Nonlinear System Identification by the Haar Multiresolution Analysis

Miroslaw Pawlak, *Member, IEEE*, and Zygmunt Hasiewicz

***Abstract*— The paper deals with the problem of reconstruction of nonlinearities in a certain class of nonlinear systems of composite structure from their input–output observations when prior information about the system is poor, thus excluding the standard parametric approach to the problem. The multiresolution idea, being the fundamental concept of modern wavelet theory, is adopted, and the Haar multiresolution analysis in particular is applied to construct nonparametric identification techniques of nonlinear characteristics. The pointwise convergence properties of the proposed identification algorithms are established. Conditions for the convergence are given; and for nonlinearities satisfying a local Lipschitz condition, the rate of convergence is evaluated. With applications in mind, the problem of data-driven selection of the optimum resolution degree in the identification procedure, essential for the multiresolution analysis, is considered as well. The theory is verified in the computer simulations.**

***Index Terms*—Composite nonlinear systems, convergence analysis, Haar multiresolution, multiresolution analysis, nonlinear system identification, nonparametric regression, wavelets.**

I. Introduction

A LARGE class of physical systems in practice are nonlinear or reveal nonlinear behavior if they are considered over a broad operating range. Hence the commonly used linearity assumption can be regarded as a first-order approximation to the observed process. System identification is the problem of complete determination of a system description (mathematical model) from an analysis of its input and output data. A large class of techniques exists for identification of linear models; see, e.g., [32], [34], [44] for an extensive discussion of this subject. Much less attention has been paid to nonlinear system identification, mostly because their analysis is generally harder and because the range of nonlinear model structures and behaviors is much broader than the range of linear model structures and behaviors. There is no universal approach to identification of nonlinear systems, and existing solutions depend strongly on a prior knowledge of the system structure; see [2]–[5], [21], [31], [33] for some techniques for nonlinear system identification. In general, the causal nonlinear (discrete time) system transforms the input data $\{X_t, t \leq n\}$ into the output signal Y_n at the time n. This transformation can be approximated in various ways, and an early approach relies on Volterra and Wiener expansions, see [5], [33], [42] and the references cited therein. These representations lead, however, to very complicated identification algorithms since multidimensional Volterra/Wiener kernels must be evaluated, often requiring an extremely large input–output data set. An alternative strategy is based on the assumption that the system structure is to some extent known. This yields the concept of block-oriented models, i.e., models consisting of linear dynamic subsystems and static nonlinear elements connected together in a certain composite structure. Signals interconnecting the subsystems are not accessible for measurement, making the identification problem not reducible to the standard situations, i.e., identification of linear dynamic systems and recovering memoryless nonlinearities. Such composite models have found numerous applications in such different and distant areas as biology, communication systems, chemical engineering, psychology and sociology; see [2], [3], [5], [10], [25], [33] and the references cited therein for some specific case studies. A class of cascade/parallel models is a popular type of block-oriented structures, i.e., when linear dynamic subsystems are in a tandem/parallel connection with a static element. Examples of such models include cascade Hammerstein, Wiener and sandwich structures and their parallel counterparts [2]–[5], [12], [14], [21], [25], [31], [33]. The popularity of these connections stems not only from their relative simplicity (allowing us to design a constructive identification algorithms) but surprisingly from their ability to approximate closely more general systems which are not necessarily of this form. This is particularly the case if one allows in the cascade/parallel models a general class of nonlinear characteristics not being able to be parameterized and smooth, e.g., not being just a polynomial of a finite order. We refer to [2], [3], [5], [10], [21], [25], and [31] for parametric identification techniques of the cascade/parallel block-oriented models with polynomial nonlinearities. The parametric restriction is often too rigid, i.e., if one chooses a parametric family of models that is not of appropriate form, then there is a danger of reaching incorrect conclusions in the system identification. In [14], and then [12], [15]–[20], [28], [29], [36], and [37], the nonparametric approach to identification of the cascade/parallel block-oriented models has been proposed. The aim of the nonparametric method is to relax assumptions on the form of an underlying nonlinear characteristic, and to let the training data decide which characteristic fits them best. These approaches are powerful in exploring fine details in nonlinear characteristics.

Manuscript received December 8, 1995; revised November 17, 1997. The work of M. Pawlak was supported by Natural Sciences and Engineering Research Council of Canada under Grant A8131 and by the Humboldt Foundation. This paper was recommended by Associate Editor J. Goetze.

M. Pawlak is with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Man., R3T 5V6, Canada.

Z. Hasiewicz is with the Institute of Engineering Cybernetics, Technical University of Wroclaw, 50-372 Wroclaw, Poland.

Publisher Item Identifier S 1057-7122(98)06496-4.

In this paper we consider the nonparametric approach to the identification of a broad class of nonlinear composite models which includes most previously defined connections. We are mostly interested in recovering a nonlinearity which is embedded in a block oriented structure containing dynamic linear subsystems and other "nuisance" nonlinearities. We illustrate our class by giving specific examples including the aforementioned popular cascade/parallel models. Our approach is based on regression analysis and we propose the identification algorithms originating from the area of nonparametric regression techniques; we refer to [7], [9], [11], [13], [23], [24], [35], [38], [40], [41], [43], [45], and [48] for the theory and applications of nonparametric curve estimation. The proposed identification algorithm is convergent for a large class of nonlinear characteristics and under very mild conditions on the system dynamics. The algorithm is based on the theory of orthogonal bases originating from multiresolution and wavelet approximations of square integrable functions. This theory provides elegant techniques for representing the levels of details of the approximated function. Multiresolution and wavelet theory has recently found applications in a remarkable diversity of disciplines such as, e.g., data compression, image analysis, signal processing, numerical analysis and statistics, see [1], [6], [8], [30], [39], [46], and [47] for a full account of the theory and applications of this subject. Little attention, however, has been paid to the application of the multiresolution and wavelet methodology to system theory and to system identification in particular; see [26] for some preliminary studies into this direction.

In this paper we apply the Haar multiresolution analysis to the identification of the proposed nonlinear composite systems. We give conditions for the identification algorithm to be pointwise convergent and find its optimal rate of convergence. As a result of these studies, the optimal local choice of the resolution level is calculated. This optimal value depends on some unknown characteristics of the system, and therefore the problem of estimating the resolution level from data is also addressed. We use the Haar multiresolution basis due to its simple structure (the scaling and wavelet functions are given explicitly) and good localization properties. Furthermore, the basis has a simple discrete structure (its values are quantized to two levels) and therefore it lends itself to a number of applications in digital circuits and systems where the discontinuous characteristics often occur. Nevertheless, it is worth noting that our considerations can be generalized to other multiresolution bases.

II. Multiresolution Analysis and the Haar System

In this section we give a brief overview of some concepts of the multiresolution and wavelet theory which are essential for our paper; see [6], [30], [46], [47] for detailed treatments of this subject. Let Z denote the set of all integer numbers. The essential idea of multiresolution analysis is to decompose the function space $L_2(R) = \{f : \int f^2(t)\,dt < \infty\}$ in an increasing sequence $\{V_m\}_{m=-\infty}^{\infty}$ of closed approximating subspaces of $L_2(R)$, i.e.,

$$\{0\} \rightarrow \cdots V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \cdots \rightarrow L_2(R) \quad (2.1)$$

with a property that the union of V_m's is dense in $L_2(R)$ and their intersection is $\{0\}$.

The subspace V_m is identified by the following two properties:

$$f(\bullet) \in V_m \Leftrightarrow f(2\bullet) \in V_{m+1},\, f(\bullet) \in V_m \\ \Leftrightarrow f(\bullet - 2^{-m}k) \in V_m, \qquad \text{for all } k \in Z. \quad (2.2)$$

There exists a function $\varphi \in V_0$, called a scaling function, such that the set of functions

$$\{\varphi_{mk}(x)\}_{k\in Z} = \{2^{m/2}\varphi(2^m x - k)\}_{k \in Z} \quad (2.3)$$

is an orthonormal basis in V_m.

Hence, any function f from $L_2(R)$ can be approximated (at resolution 2^{-m}) by its orthogonal projection f_m on V_m

$$f_m(x) = \sum_{k\in Z} a_{mk}\varphi_{mk}(x) \quad (2.4)$$

where $a_{mk} = \int_{-\infty}^{\infty} f(x)\varphi_{mk}(x)\,dx$ is the kth Fourier coefficient at resolution 2^{-m}.

Plugging the definition of a_{mk} into (2.4), we can rewrite the formula for $f_m(x)$ in the following integral form:

$$f_m(x) = \int_{-\infty}^{\infty} q_m(x, v) f(v)\,dv \quad (2.5)$$

where the kernel $q_m(x, v)$ is given by

$$q_m(x, v) = \sum_{k\in Z} \varphi_{mk}(x)\varphi_{mk}(v) = 2^m q(2^m x, 2^m v) \quad (2.6)$$

with

$$q(x, v) = q_0(x, v) = \sum_{k \in Z} \varphi(x - k)\varphi(v - k). \quad (2.7)$$

It is clear that due to the multiresolution property (2.1), the following convergence holds:

$$\lim_{m\rightarrow\infty} \int_{-\infty}^{\infty} |f_m(x) - f(x)|^2\,dx = 0. \quad (2.8)$$

The pointwise convergence of $f_m(x)$ to $f(x)$ is less trivial, and we refer to [27] and [47] for some general results into this direction; see also Lemma 1 in Section VI for the pointwise convergence of the Haar multiresolution basis.

A number of scaling functions φ with various properties has been proposed in the literature, culminating in the seminal work of Daubechies [6] on compactly supported scaling functions. A quick inspection of the conditions (2.1), (2.2), (2.3) shows that a scaling function φ determines the multiresolution analysis completely. Hence the construction of the scaling function with some desired properties like smoothness and compact support is an essential problem in the multiresolution analysis. By virtue of (2.1), one can observe that $\varphi \in V_1$ and therefore φ can be represented at the resolution 2^{-1} as follows:

$$\varphi(x) = \sqrt{2}\sum_{k\in Z} h_k \varphi(2x - k) \quad (2.9)$$

where h_k is the kth Fourier coefficient of $\varphi(x)$ in the basis $\{\varphi_{1k}(x)\}_{k\in Z}$. The formula in (2.9) forms a basis for finding φ and it is often referred to as a scaling equation.

The wavelet analysis characterizes the detail information hidden between two consecutive resolution levels. The latter is quantitatively described by the property that $\{\psi_{mk}(x)\}_{k\in Z} = \{2^{m/2}\psi(2^m x - k)\}_{k\in Z}$ forms an orthonormal basis of the detail subspace W_m, being the orthogonal complement of V_m in V_{m+1}, i.e., $V_{m+1} = V_m \oplus W_m$. Consequently we can decompose V_m as follows $V_m = \cdots W_{-1} \oplus W_0 \oplus \cdots \oplus W_{m-1}$. The wavelet function ψ (often called the mother wavelet) has a property that $\{\psi(x-k)\}_{k\in Z}$ is an orthonormal basis in W_0. Since moreover $\psi \in V_1$, then it can be expressed in terms of the scaling function as follows:

$$\psi(x) = \sqrt{2}\sum_{k\in Z} g_k \varphi(2x-k) \tag{2.10}$$

where it can be shown [6] that the Fourier coefficients $\{g_k\}$ can be determined from the formula $g_k = (-1)^k h_{-k+1}$, where $\{h_k\}$ is defined in (2.9). The latter relationship yields the concept of the so-called mirror filter.

As a consequence of the aforementioned properties, a function $f \in L_2(R)$ may be expanded in terms of the wavelet basis as follows:

$$f(x) = \sum_{m\in Z}\sum_{k\in Z} c_{mk}\psi_{mk}(x). \tag{2.11}$$

This in turn implies that the orthogonal projection $f_m(x)$ [see (2.4)] of f onto V_m has the following alternative representation in terms of $\{\psi_{mk}(x)\}_{m,k\in Z}$:

$$f_m(x) = \sum_{s=-\infty}^{m-1}\sum_{k\in Z} c_{sk}\psi_{sk}(x). \tag{2.12}$$

In this paper we utilize the multiresolution analysis based on Haar basis. This is one of the simplest examples of multiresolution systems and wavelet basis where both the scaling function φ and the wavelet function ψ are given explicitly and they are of compact support. It has been discovered recently [6] that there are other than Haar basis compactly supported wavelets which moreover can be chosen arbitrary smooth. As has already been mentioned, and will be apparent from the results of our paper, the use of the Haar basis leads to very intuitive identification algorithms with desired convergence properties and highly efficient computational features.

The scaling function of the Haar system can be taken as follows:

$$\varphi(x) = \mathbf{1}_{[0,1)}(x) \tag{2.13}$$

where $\mathbf{1}_A(x)$ denotes the indicator function of A.

Consequently the multiresolution basis function $\varphi_{mk}(x)$ is given by

$$\varphi_{mk}(x) = 2^{m/2}\mathbf{1}_{[k/2^m,\,(k+1)/2^m)}(x) \tag{2.14}$$

and the resolution space V_m is defined as follows

$$V_m = \{\text{all functions in } L_2(R) \text{ constant on all intervals } [k2^{-m}, (k+1)2^{-m}), \quad \text{for } k \in Z\}.$$

Thus the set where $\varphi_{mk}(x)$ is equal $2^{m/2}$ is a small interval of length 2^{-m}.

It is clear that the kernel function $q(x,v)$ defined in (2.7) is now given by

$$q(x,v) = \sum_{k\in Z}\mathbf{1}_{[k,k+1)}(x)\mathbf{1}_{[k,k+1)}(v) = \mathbf{1}_{[0,1)}(|x-v|). \tag{2.15}$$

Furthermore, the scaling equation in (2.9) takes the form

$$\varphi(x) = \varphi(2x) + \varphi(2x+1) \tag{2.16}$$

while the formula for the wavelet function in (2.10) is given by

$$\psi(x) = \varphi(2x) - \varphi(2x-1). \tag{2.17}$$

Hence the wavelet function ψ is given by

$$\psi(x) = \mathbf{1}_{[0,1/2)}(x) - \mathbf{1}_{[1/2,1)}(x) \tag{2.18}$$

and the wavelet system $\{\psi_{mk}(x)\}_{m,k\in Z}$ consists of functions which are nonzero in a small interval of length 2^{-m} and as m increases the support of $\psi_{mk}(x)$ shrinks, i.e., $\psi_{mk}(x)$ becomes taller and thinner. Let us also note that in the orthogonal series literature $\{\psi_{mk}(x)\}_{m,k\in Z}$ is usually referred to as the Haar orthonormal basis in $L_2(R)$ [47].

III. The Regression Function and Nonlinear Composite Systems

Regression analysis is a standard tool used for recovering some nonlinear relationships of two random processes. Applied to nonlinear system identification, the analysis makes it possible to recover the nonlinearities existing in a system from regression functions of the input and output processes. Hence let $\{(X_n, Y_n)\}$ be a sequence of random pairs (X_n, Y_n) representing the input and output signals of a certain dynamical system. The standard regression function of the process $\{Y_n\}$ on $\{X_n\}$ is defined as follows:

$$r(x) = E\{Y_n \mid X_n = x\}. \tag{3.1}$$

It is clear that the calculation of the regression function requires the knowledge of the probability distribution function of the processes $\{(X_n, Y_n)\}$. This is, however, rarely known in practice and one has to estimate $r(x)$ from the input-output training data $\{(X_t, Y_t), 1 \le t \le n\}$. The problem of estimation of $r(x)$ when $\{(X_t, Y_t)\}$ is a sequence of independent and identically distributed (iid) random variables has been extensively studied in the statistical literature [7], [11], [13], [23], [24], [38], [43], [48]. In this paper it is assumed that the system is excited by the iid signal $\{X_t\}$, whereas $\{Y_t\}$ being an output of a nonlinear time-invariant dynamic system is a dependent stationary stochastic process, which is in contrast to the papers cited above.

Let us now introduce a class of nonlinear composite systems examined in this paper. The class is characterized by the general property that the nonlinear characteristic of our interest can be extracted from the rest of the system.

Hence our general nonlinear model is of the following form (depicted in Fig. 1):

$$\begin{cases} O_n = \mu(X_n) + \xi_n \\ Y_n = O_n + \varepsilon_n \end{cases} \tag{3.2}$$

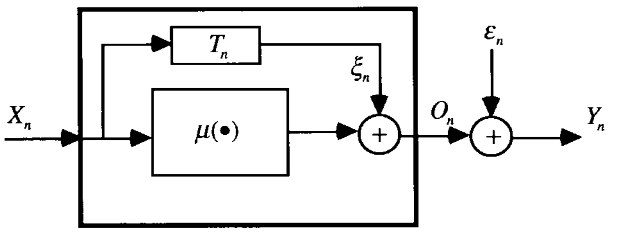


Fig. 1. Nonlinear dynamic system.

where (X_n, Y_n) is the (input, output) pair, $\mu(\bullet)$ represents the unknown system nonlinearity, $\{\xi_n\}$ is the correlated system "noise" process characterizing the history of the system, and $\{\varepsilon_n\}$ is the measurement noise. The block T_n in Fig. 1 represents an element producing the system noise process ξ_n and it is a measurable transformation (see Fig. 2) of $\{X_{n-1}, X_{n-2}, \cdots\}$ defined as follows:

$$\xi_n = T_n(X_{n-1}, X_{n-2}, \cdots) = \sum_{j=-\infty}^{n-1} s_{n-j}\lambda_{n-j}(X_j). \tag{3.3}$$

The following assumptions concerning the model in (3.2) are used in the paper.

Assumption 1: The inputs $\{X_1, X_2, \cdots\}$ form a sequence of independent and identically distributed random variables which are independent of $\{\varepsilon_n\}$. The probability density function f of $\{X_1, X_2, \cdots\}$ is unknown and satisfies the following restrictions:

$$\int_{-\infty}^{\infty} f^2(x)\,dx < \infty \tag{A1.1}$$

$$0 < \eta \le f(x) \tag{A1.2}$$

for all $x \in R$ and some unknown η.

Assumption 2: The system noise process $\{\xi_n\}$ is defined by (3.3) and depicted in Fig. 2, where $\{\lambda_j(\bullet)\}$ is a sequence of measurable functions and $\{s_j\}$ is a sequence of numbers. Furthermore we assume that

$$E\lambda_j(X) = 0, \qquad j = 1, 2, \cdots \tag{A2.1}$$

$$\sum_{j=1}^{\infty} s_j^2 E\lambda_j^2(X) < \infty \tag{A2.2}$$

$$\sum_{j=1}^{\infty} |s_j||\lambda_j(x)| < \infty, \qquad \text{for almost all } x \in R \tag{A2.3}$$

$$\sum_{t=1}^{\infty}\sum_{j=1}^{\infty} |s_j s_{t+j}| E\{|\lambda_j(X)\lambda_{t+j}(X)|\} < \infty. \tag{A2.4}$$

Assumption 3: The nonlinear characteristic $\mu(\bullet)$ is a measurable function satisfying the following condition:

$$E\mu^2(X) < \infty \tag{A3.1}$$

$$\int_{-\infty}^{\infty} (\mu(x)f(x))^2\,dx < \infty. \tag{A3.2}$$

Assumption 4: The measurement noise $\{\varepsilon_n\}$ is uncorrelated and such that

$$E\varepsilon_n = 0, \qquad \operatorname{var}\varepsilon_n = \sigma^2 < \infty.$$

Let us elaborate on the role of the above conditions. The process $\{\xi_n\}$ has an infinite nonlinear moving average representation and its realization for the case of the pth-order moving average (FIR system) is depicted in Fig. 2 (D is the delay operator). The restriction (A1.1) is required since we use the $L_2(R)$ multiresolution decomposition of $f(x)$. The condition (A1.2) says that we consider the estimation problem in such points on R where the input density is high, i.e., where $f(x)$ is strictly bounded away from zero. The Assumptions (A2.1) and (A2.2) are necessary for $\{\xi_n\}$ to be the second order covariance stationary stochastic process with $E\xi_n = 0$, $\operatorname{var}\xi_n < \infty$ and $\operatorname{cov}(\xi_n, \xi_{n+t}) = \sum_{j=1}^{\infty} s_j s_{j+1} E\{\lambda_j(X)\lambda_{j+t}(X)\}$, $|t| > 1$. This along with Assumptions (A3.1) and 4 makes the output process $\{Y_n\}$ well defined, i.e., it is also a second order covariance stationary stochastic process. It is worth noting that $\{Y_n\}$ is not strictly stationary process.

The conditions (A2.3), (A2.4), (A3.2) are related to our identification procedure for recovering $\mu(\bullet)$ and they will be discussed later. Let us note only that (A2.3) is meant in the Lebesque measure sense, i.e., it holds at all points $x \in R$, except sets with zero Lebesque measure. In particular, (A2.3) is true at all points where $\{\lambda_j(x), j = 1, 2, \cdots\}$ are continuous.

It is a fundamental fact for our paper to observe that

$$E\{Y_n \mid X_n = x\} = \mu(x) \tag{3.4}$$

i.e., the system nonlinearity is just equal to the standard regression function $r(x)$ defined in (3.1). Thus by estimating the regression in (3.4) we can recover the nonlinearity $\mu(x)$.

Surprisingly there is a large class of block-oriented nonlinear models which fall into the description given in (3.2), (3.3). In the next section we give a number of specific examples which include both well-known structures as well as some new models. A detailed discussion of Assumption 2 in all examples is given.

IV. Examples of Block-Oriented Models

Example 1 (Memoryless System): The simplest situation represented by (3.2) is the memoryless system, i.e.,

$$Y_n = \theta(X_n) + \varepsilon_n$$

shown in Fig. 3.

It is clear that this is a special case of (3.2) with $\xi_n = 0$ and $\mu(x) = \theta(x)$. We refer to [48] for a recent overview of nonparametric techniques for estimation of memoryless systems. Wavelet-based techniques for this model are examined in [1].

Example 2 (Cascade System): The second system is dynamic and has a cascade structure. It consists of a nonlinear static element $\theta(\bullet)$ followed by a linear dynamic system with the impulse response function $\{h_i\}$, see Fig. 4. Such a system is often referred to as the Hammerstein system.

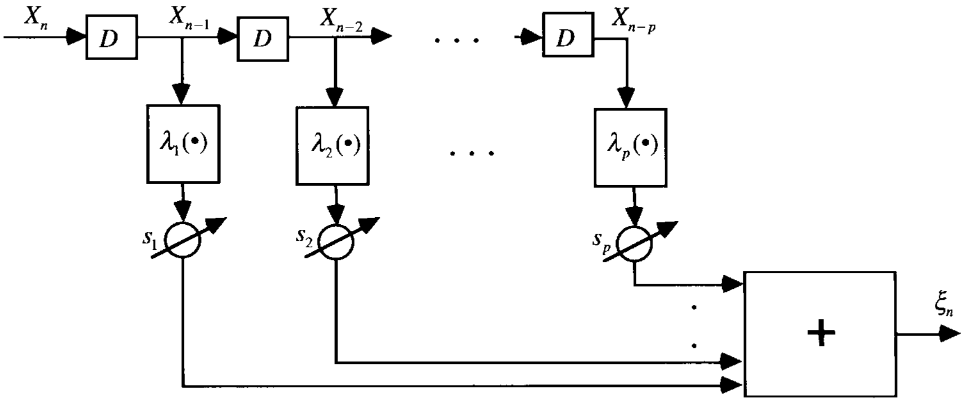


Fig. 2. System noise $\{\xi_n\}$ structure.

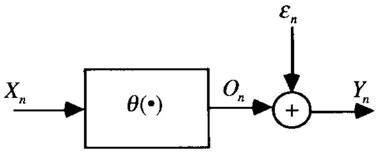


Fig. 3. Nonlinear memoryless system.

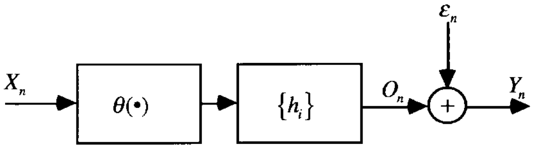


Fig. 4. Cascade nonlinear system.

The system input–output relationship is given by

$$Y_n = \sum_{j=-\infty}^{n} h_{n-j}\theta(X_j) + \varepsilon_n$$

which clearly can be written in the form of (3.2) with

$$\mu(x) = \theta(x) + E\theta(X)\sum_{j=1}^{\infty} h_j \tag{4.1}$$

and

$$\xi_n = \sum_{j=-\infty}^{n-1} h_{n-j}(\theta(X_j) - E\theta(X_j))$$

where it has been assumed, without loss of generality, that $h_0 = 1$.

Hence here $\lambda_j(X) = \lambda(X) = \theta(X) - E\theta(X)$ all j. Furthermore, one can easily observe that Assumptions 2 and 3 are satisfied if

$$E\theta^2(X) < \infty$$
$$\sum_{j=1}^{\infty}|h_j| < \infty, \qquad \sum_{t=1}^{\infty}\sum_{j=1}^{\infty}|h_j||h_{j+t}| < \infty. \tag{4.2}$$

The latter condition holds for any BIBO stable system with the impulse response $\{h_i\}$ being square summable.

It is also clear from (4.1) that one can only estimate $\theta(\bullet)$ up to an additive constant, the property which is independent of any identification procedure. In order to eliminate the constant, some prior information on $\theta(\bullet)$ must be incorporated. In particular, if $E\theta(X) = 0$ (which takes place if, e.g., $f(\bullet)$ is symmetric and $\theta(\bullet)$ is odd), then $\mu(x) = \theta(x)$. Also, if we know that $\theta(0) = 0$ (which is often the case), then $\theta(x) = \mu(x) - \mu(0)$. In the next section we introduce a consistent estimate of $\mu(x)$, and in the light of the aforementioned relationships between $\mu(x)$ and $\theta(x)$ this also yields a consistent estimate of the nonlinearity $\theta(x)$. We refer to [14]–[20], [28], [29], [36] for various nonparametric identification algorithms of the Hammerstein system.

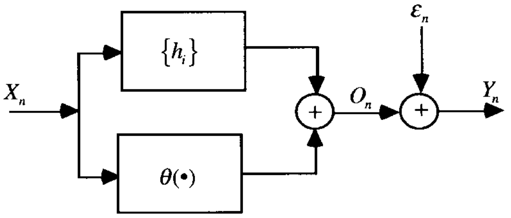


Fig. 5. Parallel nonlinear system.

Example 3 (Parallel System): As a complement to the previous example, a system of the parallel structure (depicted in Fig. 5) is considered here.

The system input–output equation is given by following formula:

$$Y_n = \theta(X_n) + \sum_{j=-\infty}^{n} h_{n-j}X_j + \varepsilon_n$$

which can be represented in the form (3.2) with (putting $h_0 = 1$)

$$\mu(x) = \theta(x) + x + EX\sum_{j=1}^{\infty} h_j$$
$$\xi_n = \sum_{j=-\infty}^{n-1} h_{n-j}(X_j - EX_j).$$

Hence $\lambda_j(X) = \lambda(X) = X - EX$ and it is clear that Assumptions 2 and 3 are met if $\{h_j\}$ satisfies the conditions as in Example 2, see (4.2), and

$$E\theta^2(X) < \infty, \quad EX^2 < \infty.$$

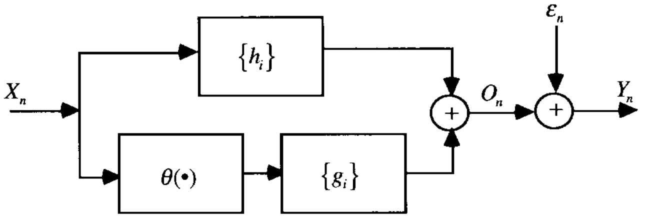


Fig. 6. Parallel–series nonlinear system.

The above formulas reveal that if $EX = 0$, then $\theta(x) = \mu(x) - x$, and again a consistent estimate of $\mu(x)$ yields a convergent identification algorithm for $\theta(x)$. Nonparametric identification algorithms for the parallel structures are studied in [19] and [20].

Example 4 (Parallel-Series System): A combination of the structures in Examples 2 and 3 leads to another nonlinear block-oriented system depicted in Fig. 6.

This is an example of the system containing two dynamical elements $\{g_j\}$, $\{h_j\}$, and having the following input–output description:

$$Y_n = \sum_{i=-\infty}^{n} g_{n-i}\theta(X_i) + \sum_{j=-\infty}^{n} h_{n-j}X_j + \varepsilon_n$$

being transformable to the representation in (3.2) with

$$\begin{aligned}\mu(x) =& \theta(x) + x + E\theta(X)\sum_{j=1}^{\infty} g_j + EX\sum_{i=1}^{\infty} h_i\\ \xi_n =& \sum_{j=-\infty}^{n-1}\Big\{g_{n-j}(\theta(X_j) - E\theta(X_j))\\ &+ h_{n-j}(X_j - EX_j)\Big\}. \qquad (4.3)\end{aligned}$$

It is worth noting that in this case the process $\{\xi_n\}$ is not exactly in the form as in Assumption 2, i.e., the convolution between $\{s_j\}$ and $\{\lambda_j(X_{n-j})\}$. Such a representation is, however, possible by augmenting the convolution formula to vector sequences, i.e., by defining $\boldsymbol{s}_j = (s_j^1, s_j^2)$, where $s_j^1 = g_j$, $s_j^2 = h_j$ and $\Lambda(X) = (\lambda^1(X), \lambda^2(X))$ with $\lambda^1(X) = \theta(X) - E\theta(X)$, $\lambda^2(X) = X - EX$ we can rewrite (4.3) as follows:

$$\xi_n = \sum_{j=-\infty}^{n-1} \mathbf{s}_{n-j} \bullet \Lambda(X_j)$$

where $\boldsymbol{a} \bullet \boldsymbol{b}$ is the inner product of vectors $\boldsymbol{a}$ and $\boldsymbol{b}$.

It is also clear that Assumptions 2 and 3 are met when both sequences $\{g_j\}$, $\{h_j\}$ satisfy the condition in (4.2), and furthermore $E\theta^2(X) < \infty$, $EX^2 < \infty$ must hold.

Example 5: Our final example concerns a system with two nonlinearities (Fig. 7), where $\theta(\bullet)$ is the one to be estimated and the other $\theta_0(\bullet)$ is a "nuisance" nonlinearity (known or not). Note that if $\theta_0(\bullet) \equiv 0$, then the system in Example 2 is recovered.

Fig. 7 reveals that

$$Y_n = \sum_{j=-\infty}^{n} h_{n-j}(\theta(X_j) + \theta_0(X_{j-1})) + \varepsilon_n$$

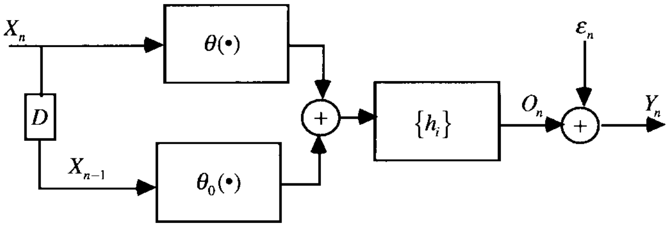


Fig. 7. Nonlinear system with two nonlinearities.

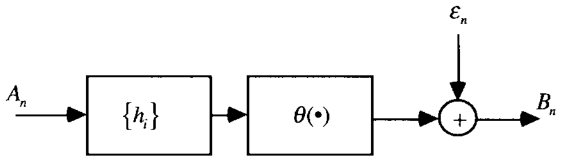


Fig. 8. Wiener nonlinear system.

which can be rewritten in the form of (3.2) with

$$\mu(x) = \theta(x) + E\theta_0(X) + E(\theta(X) + \theta_0(X))\sum_{j=1}^{\infty} h_j$$

$$\xi_n = (\theta_0(X_{n-1}) - E\theta_0(X_{n-1})) + \sum_{j=-\infty}^{n-1} h_{n-j}\gamma_j$$

where $\gamma_j = \theta(X_{j-1}) + \theta_0(X_{j-1}) - E(\theta(X_{j-1} + \theta_0(X_{j-1}))$.

The process ξ_n is in the form as in Assumption 2 with

$$s_i = \begin{cases}1, & \text{if } i = 1\\ h_{i-1}, & \text{if } i \geq 2\end{cases}$$

$$\lambda_i(X) = \begin{cases}\theta_0(X) - E\theta_0(X), & \text{if } i = 1\\ \theta(X) + \theta_0(X) - E\{\theta(X) + \theta_0(X)\}, & \text{if } i \geq 2.\end{cases}$$

Assumptions 2 and 3 hold if $\{h_j\}$ satisfies (4.2), and moreover $E\theta^2(X) < \infty$, $E\theta_0^2(X) < \infty$ must hold. Once again, if $\theta(0) = 0$, then $\theta(x) = \mu(x) - \mu(0)$ for all x.

Some other examples of systems being representable in the form (3.2) can be easily derived. Nevertheless, there are cases of block-oriented structures which are not straightforwardly expressed in that form. This includes, e.g., Wiener and sandwich systems. The Wiener structure is a tandem composition (Fig. 8) of linear dynamic system and nonlinear static element, i.e.,

$$B_n = \theta\left(\sum_{j=-\infty}^{n} h_{n-j}A_j\right) + \varepsilon_n.$$

If the measurement noise ε_n is equal to zero and $\theta(\bullet)$ is invertible, then one can write

$$\begin{cases}O_n &= \mu(X_n) + \xi_n\\ Y_n &= O_n\end{cases} \qquad (4.4)$$

where $X_n = B_n$, $Y_n = A_n$, $\mu(y) = \theta^{-1}(y)$—the inverse of $\theta(\bullet)$ and $\xi_n = -\sum_{j=-\infty}^{n-1} h_{n-j}A_j$. Clearly, (4.4) is in the form of (3.2) with the role of the input and output signals exchanged. Such an approach has been used in [12], where the detail analysis of identification algorithms of $\theta^{-1}(\bullet)$ based on classical orthogonal series expansions is given, see also [37] for an alternative method.

V. Identification Algorithms

Due to the fundamental property in (3.4), we can treat $\mu(x)$ as a standard regression function of Y_n on $X_n = x$. In order to construct an estimate of the regression function, let us first observe that

$$\mu(x) = \frac{g(x)}{f(x)}$$

where $g(x) = \mu(x)f(x)$ for every x where the Assumption (A1.2) holds. Owing to Assumption (A1.1), (A3.2) and using the results of Section II, we can approximate $g(x)$ and $f(x)$ at the resolution 2^{-m} as follows:

$$g_m(x) = \sum_{k\in Z} a_{mk}\varphi_{mk}(x), \quad f_m(x) = \sum_{k\in Z} b_{mk}\varphi_{mk}(x) \tag{5.1}$$

where one can easily observe that

$$a_{mk} = \int_{-\infty}^{\infty} \mu(x)\varphi_{mk}(x)f(x)\,dx = E\{Y_n\varphi_{mk}(X_n)\}$$

and

$$b_{mk} = \int_{-\infty}^{\infty} \varphi_{mk}(x)f(x)\,dx = E\{\varphi_{mk}(X_n)\}.$$

Empirical counterparts of $g_m(x)$ and $f_m(x)$ in (5.1) can be easily constructed first by replacing the expected values in the formulas for a_{mk} and b_{mk} by their natural estimates

$$\hat{a}_{mk} = n^{-1}\sum_{i=1}^{n} Y_i\varphi_{mk}(X_i), \quad \hat{b}_{mk} = n^{-1}\sum_{i=1}^{n} \varphi_{mk}(X_i) \tag{5.2}$$

and next by cutting off the number of terms in (5.1) to some finite value referred to in this paper as a truncation value N. All these things yield the following estimator of $\mu(x)$ at the resolution 2^{-m} based on $2N+1$ terms in the expansions (5.1):

$$\hat{\mu}(x) = \frac{\sum\limits_{|k|\le N} \hat{a}_{mk}\varphi_{mk}(x)}{\sum\limits_{|k|\le N} \hat{b}_{mk}\varphi_{mk}(x)} \tag{5.3}$$

where without loss of generality we use the same truncation value for estimating $g_m(x)$ and $f_m(x)$.

It is worth noting that $\hat{a}_{mk}$, $\hat{b}_{mk}$ are unbiased estimators of a_{mk}, b_{mk}, i.e., $E\hat{a}_{mk} = a_{mk}$, $E\hat{b}_{mk} = b_{mk}$. Recalling the definition of the kernel function [see (2.6)] associated with the basis $\{\varphi_{mk}(x)\}$, we can rewrite the estimator $\hat{\mu}(x)$ in the following form:

$$\hat{\mu}(x) = \frac{n^{-1}\sum\limits_{i=1}^{n} Y_i q_{mN}(x, X_i)}{n^{-1}\sum\limits_{i=1}^{n} q_{mN}(x, X_i)} \tag{5.4}$$

where

$$\begin{aligned} q_{mN}(x,v) &= \sum_{|k|\le N} \varphi_{mk}(x)\varphi_{mk}(v) \\ &= 2^m \sum_{|k|\le N} \varphi(2^m x - k)\varphi(2^m v - k) \end{aligned} \tag{5.5}$$

is the kernel of order N at the resolution 2^{-m}. Hence $q_{mN}(x,v)$ is the truncated version of the kernel $q_m(x,v)$ defined in (2.6), i.e., $q_m(x,v) = \lim_{N\to\infty} q_{mN}(x,v)$.

Let us also note that the expression in the numerator in (5.3) [or (5.4)] is an estimator of $g(x)$, whereas the denominator in these formulas is an estimator of $f(x)$. Hence we shall denote

$$\hat{g}(x) = \sum_{|k|\le N} \hat{a}_{mk}\varphi_{mk}(x) = n^{-1}\sum_{i=1}^{n} Y_i q_{mN}(x, X_i) \tag{5.6}$$

$$\hat{f}(x) = \sum_{|k|\le N} \hat{b}_{mk}\varphi_{mk}(x) = n^{-1}\sum_{i=1}^{n} q_{mN}(x, X_i). \tag{5.7}$$

For the particular case of the Haar multiresolution basis [see (2.13)–(2.15)], the above equations can be further simplified. In fact, the kernel $q_{mN}(x,v)$ has the form

$$\begin{aligned} &q_{mN}(x,v) \\ &\quad = 2^m \sum_{|k|\le N} \mathbf{1}_{[k/2^m,\,(k+1)/2^m)}(x)\mathbf{1}_{[k/2^m,\,(k+1)/2^m)}(v). \end{aligned} \tag{5.8}$$

Applying this in (5.4) yields the following simple histogram-like form for $\hat{\mu}(x)$:

$$\hat{\mu}(x) = \frac{\sum_{\{i:X_i\in A_m(x)\}} Y_i}{\#\{i : X_i \in A_m(x)\}} \tag{5.9}$$

where $A_m(x)$ is one of the $\{[k/2^m,\, k+1/2^m), |k| \le N\}$ intervals where x falls in.

It is an important fact to observe that (5.9) is well defined as long as

$$N \ge 2^m|x|. \tag{5.10}$$

Hence the truncation value N must be sufficiently large and (5.10) gives the lower bound for N. It is worth noting that one can use $N = \infty$ yielding, due to (2.6) and (2.7), the following counterpart of (5.4):

$$\tilde{\mu}(x) = \frac{n^{-1}\sum\limits_{i=1}^{n} Y_i q_m(x, X_i)}{n^{-1}\sum\limits_{i=1}^{n} q_m(x, X_i)} \tag{5.11}$$

where for the Haar multiresolution basis [see (2.15)] we have

$$q_m(x,v) = 2^m\varphi(2^m|x-v|).$$

It is clear that if for a given x, N satisfies (5.10) the estimators $\hat{\mu}$ and $\tilde{\mu}$ are equivalent. Hence throughout the paper, without loss of generality, we will examine the estimator $\hat{\mu}$.

Due to (5.10) it is plain that the truncation point N has to merely grow sufficiently fast with m in order to assure consistency. On the other hand, the resolution level m plays a much more important role in both asymptotic and finite sample size performance of the estimators. For the convergence property, i.e., that

$$\hat{\mu}(x) \to \mu(x) \text{ as } n \to \infty$$
$$\text{in probability for almost all } x \in R$$

it is shown that the resolution level m must be chosen as a function of the sample size n, i.e., $m = m(n)$ in such a way that

$$m(n) \to \infty \tag{5.12}$$

and

$$\frac{2^{m(n)}}{n} \to 0 \tag{5.13}$$

as $n \to \infty$.

Roughly speaking, the condition in (5.12) controls the bias of $\hat{\mu}(x)$ whereas (5.13) appears as a leading term in the variance of $\hat{\mu}(x)$.

Furthermore, under some mild smoothness conditions on $\mu(\bullet)$ and $f(\bullet)$, we will demonstrate that an optimal value of $m(n)$ exists (realizing the bias-variance tradeoff), and it is of order $\frac{1}{3}\log_2(n)$. Note that this and (5.10) lead to the following bound on the truncation value N:

$$N \geq n^{1/3}|x|. \tag{5.14}$$

The identification algorithm in (5.3) is in the form of the orthogonal series method for estimating the regression function $\hat{\mu}(x)$. In [13] an estimate of such a form based on classical orthogonal polynomials has been studied for independent pairs $\{(X_i, Y_i)\}$. This technique has been extended to some block-oriented models in [17], [20], [28], [36]. In particular, it has been proved that the convergence holds if $\mu(x)$ is differentiable at x which is consistent with the well-known [47] fact that there are examples of continuous functions whose orthogonal series diverge. On the contrary, the wavelet expansions converge for all continuous functions, and consequently they can be applied to a broader class of nonlinear characteristics.

VI. Convergence Analysis

In this section we give a detailed analysis of the convergence properties of our identification algorithms. In particular, the sufficient conditions for the pointwise convergence of the estimators to the unknown nonlinearity are given. The convergence properties hold for all input densities $f(\bullet)$ and all measurable nonlinearities $\mu(\bullet)$ which satisfy Assumptions 1 and (A3.2). No continuity conditions for the characteristic $\mu(\bullet)$ are required. We should stress again that the latter property is not shared by estimates employing the usual classical orthogonal systems [13], [17], [20], [28], [36].

In order to establish the convergence results, we need the following preliminary results concerning the kernel function in (5.8).

Lemma 1: Let $w(\bullet)$ be a measurable function for which $\int_{-\infty}^{\infty} |w(x)|\,dx < \infty$. If the truncation value N satisfies (5.10), then

$$\int_{-\infty}^{\infty} q_{mN}(x,z)w(z)\,dz \to w(x)$$
$$\text{as } m \to \infty \qquad \text{for almost every } x \in R.$$

The proof of this result is given in the Appendix. The convergence in Lemma 1 holds in particular in each point x where $w(x)$ is continuous. It is also clear that an analogous result is true for the nontruncated kernel $q_m(x,v) = \lim_{N\to\infty} q_{mN}(x,v)$.

Lemma 2: Under the conditions of Lemma 1 we have

$$2^{-m}\int_{-\infty}^{\infty} q_{mN}^2(x,z)w(z)\,dz \to w(x)$$
$$\text{as } m \to \infty \qquad \text{for almost every } x \in R.$$

This result is proved in the Appendix. Lemma 2 holds also for the kernel $q_m(x,v)$.

We are now in a position to prove the convergence result for the estimator $\hat{\mu}(x)$ defined in (5.2)–(5.4), or (5.9). Let $a_n \approx b_n$ for the number sequences $\{a_n\}$, $\{b_n\}$ denote the fact that $\lim_{n\to\infty} a_n/b_n = 1$.

Theorem 1: Let Assumptions 1–4 of Section III be satisfied. If (5.10), (5.12), and (5.13) hold then

$$\hat{\mu}(x) \to \mu(x) \qquad \text{as } n \to \infty \text{ in probability}$$

for almost every $x \in R$ and in particular at every x where $\mu(x)$ and $f(x)$ are continuous.

Proof of Theorem 1: Let us recall that $\hat{\mu}(x) = \hat{g}(x)/\hat{f}(x)$, where $\hat{g}(x)$, $\hat{f}(x)$ are defined in (5.6) and (5.7). It suffices to show that $\hat{g}(x)$ and $\hat{f}(x)$ converge (in probability) to $g(x)$ and $f(x)$, respectively.

Let us first consider $\hat{g}(x)$ by noting that [see (5.6)] $\hat{g}(x) = n^{-1}\sum_{i=1}^{n} Y_i q_{mN}(x, X_i)$ and [see (3.2)]

$$Y_i = \mu(X_i) + \xi_i + \varepsilon_i \tag{6.1}$$

where ξ_i is the system noise and ε_i is the measurement noise.

By this we have

$$\begin{aligned}\operatorname{var}\hat{g}(x) &= n^{-2}\sum_{i=1}^{n}\operatorname{var}(Y_i q_{mN}(x,X_i)) \\ &\quad + 2n^{-2}\sum_{r=2}^{n}\sum_{p=1}^{r-1}\operatorname{cov}[Y_p q_{mN}(x,X_p), Y_r q_{mN}(x,X_r)] \\ &= L_1(x) + L_2(x).\end{aligned} \tag{6.2}$$

Assumption 4 and (6.1) imply that

$$\begin{aligned}nL_1(x) &= \operatorname{var}(\mu(X_n)q_{mN}(x,X_n)) + \operatorname{var}(\xi_n)Eq_{mN}^2(x,X_n) \\ &\quad + \sigma^2 Eq_{mN}^2(x,X_n)\end{aligned} \tag{6.3}$$

where $\sigma^2 = \operatorname{var}\varepsilon_n$.

Owing to Lemmas 1 and 2 and Assumption (A3.1) we have

$$\operatorname{var}(\mu(X_n)q_{mN}(x,X_n)) \approx 2^m\mu^2(x)f(x) - (\mu(x)f(x))^2$$
$$\text{as } m \to \infty$$

for almost every $x \in R$.

Due to the same reasons

$$Eq_{mN}^2(x,X_n) \approx 2^m f(x) \qquad \text{as } m \to \infty$$

for almost every $x \in R$.

Note also that Assumptions (A2.1) and (A2.2) imply

$$\operatorname{var}(\xi_n) = \sum_{j=1}^{\infty} s_j^2 E\lambda_j^2(X) < \infty.$$

Hence the term $L_1(x)$ in (6.2) is of order

$$L_1(x) \approx a(x)\frac{2^m}{n} \qquad \text{as } m \to \infty \tag{6.4}$$

for almost every $x \in R$, where

$$a(x) = f(x)\{\mu^2(x) + \text{var}(\xi_n) + \sigma^2\}.$$

Let us now turn our attention to the term $L_2(x)$ in (6.2). By virtue of (6.2) and Assumption 4 we can rewrite the covariance in $L_2(x)$ as follows:

$$\begin{aligned}\text{cov}&[Y_p q_{mN}(x, X_p), Y_r q_{mN}(x, X_r)]\\ &= \text{cov}[\mu(X_p)q_{mN}(x, X_p), \mu(X_r)q_{mN}(x, X_r)]\\ &\quad + \text{cov}[\mu(X_p)q_{mN}(x, X_p), \xi_r q_{mN}(x, X_r)]\\ &\quad + \text{cov}[\xi_p q_{mN}(x, X_p), \mu(X_r)q_{mN}(x, X_r)]\\ &\quad + \text{cov}[\xi_p q_{mN}(x, X_p), \xi_r q_{mN}(x, X_r)]\\ &= C_1(x) + C_2(x) + C_3(x) + C_4(x).\end{aligned} \tag{6.5}$$

From Assumption 1 and the fact that for $p < r$ the random variable ξ_p is independent of X_r, we can easily conclude that $C_1(x) = C_3(x) = 0$.

Using the definition of ξ_r (see Assumption 2) and the fact that $E\lambda_j(X) = 0$ we can obtain

$$\begin{aligned}C_2&(x)\\ &= \sum_{j=-\infty}^{r-1} s_{r-j}\,\text{cov}[\mu(X_p)q_{mN}(x, X_p), \lambda_{r-j}(X_j)q_{mN}(x, X_r)]\\ &= s_{r-p}\,\text{cov}[\mu(X_p)q_{mN}(x, X_p), \lambda_{r-p}(X_p)q_{mN}(x, X_r)]\\ &= s_{r-p}E\{\lambda_{r-p}(X_p)\mu(X_p)q_{mN}(x, X_p)\}E\{q_{mN}(x, X_r)\}.\end{aligned}$$

By virtue of Lemma 1 we have

$$C_2(x) \approx s_{r-p}\lambda_{r-p}(x)\mu(x)f^2(x)$$

for almost every $x \in R$.

Regarding the term $C_4(x)$ in (6.5) we can obtain for $r > p$

$$\begin{aligned}C_4&(x)\\ &= \sum_{i=-\infty}^{p-1}\sum_{j=-\infty}^{r-1} s_{p-i}s_{r-j}\,\text{cov}[\lambda_{p-i}(X_i)q_{mN}(x, X_p),\\ &\qquad\qquad \lambda_{r-j}(X_j)q_{mN}(x, X_r)]\\ &= \sum_{i=-\infty}^{p-1}\sum_{j=-\infty}^{p-1} s_{p-i}s_{r-j}\,\text{cov}[\lambda_{p-i}(X_i)q_{mN}(x, X_p),\\ &\qquad\qquad \lambda_{r-j}(X_j)q_{mN}(x, X_r)]\\ &\quad + s_{r-p}\sum_{i=-\infty}^{p-1} s_{p-i}\,\text{cov}[\lambda_{p-i}(X_i)q_{mN}(x, X_p),\\ &\qquad\qquad \lambda_{r-p}(X_p)q_{mN}(x, X_r)]\\ &\quad + \sum_{i=-\infty}^{p-1}\sum_{j=p+1}^{r-1} s_{p-i}s_{r-j}\,\text{cov}[\lambda_{p-i}(X_i)q_{mN}(x, X_p),\\ &\qquad\qquad \lambda_{r-j}(X_j)q_{mN}(x, X_r)]\\ &= C_{41}(x) + C_{42}(x) + C_{43}(x).\end{aligned}$$

It is clear that $C_{43}(x) = 0$. Concerning the term $C_{42}(x)$ let us observe that since $i \le p - 1$ and $p < r$, therefore $\text{cov}[\lambda_{p-i}(X_i)q_{mN}(x, X_p), \lambda_{r-p}(X_p)q_{mN}(x, X_r)] = 0$, and consequently $C_{42}(x) = 0$.

Let us finally consider the term $C_{41}(x)$.

Clearly,

$$\begin{aligned}C_{41}&(x)\\ &= \sum_{i=-\infty}^{p-1} s_{p-i}s_{r-i}\,\text{cov}[\lambda_{p-i}(X_i)q_{mN}(x, X_p),\\ &\qquad\qquad \lambda_{r-i}(X_i)q_{mN}(x, X_r)]\\ &\quad + \sum_{i=-\infty}^{p-1}\sum_{j=-\infty, i\neq j}^{r-1} s_{p-i}s_{r-j}\,\text{cov}[\lambda_{p-i}(X_i)q_{mN}(x, X_p),\\ &\qquad\qquad \lambda_{r-j}(X_j)q_{mN}(x, X_r)].\end{aligned}$$

The covariance in the second term is equal to zero, while for the first one it is given by

$$\begin{aligned}E\{&\lambda_{p-i}(X_i)\lambda_{r-i}(X_i)q_{mN}(x, X_p)q_{mN}(x, X_r)\}\\ &= E\{\lambda_{p-i}(X_i)\lambda_{r-i}(X_i)\}E\{q_{mN}(x, X_p)\}E\{q_{mN}(x, X_r)\}\\ &\approx f^2(x)E\{\lambda_{p-i}(X)\lambda_{r-i}(X)\}\end{aligned}$$

for almost every $x \in R$.

Hence,

$$C_4(x) \approx f^2(x)\sum_{l=1}^{\infty} s_l s_{r-p+l}E\{\lambda_l(X)\lambda_{r-p+l}(X)\}$$

for almost every $x \in R$.

All these considerations yield the following asymptotic expression for $L_2(x)$ in (6.2):

$$\begin{aligned}L_2(x) &\approx 2f^2(x)n^{-2}\left\{\mu(x)\sum_{r=2}^{n}\sum_{p=1}^{r-1} s_{r-p}\lambda_{r-p}(x)\right.\\ &\qquad \left. + \sum_{r=2}^{n}\sum_{p=1}^{r-1}\sum_{l=1}^{\infty} s_l s_{r-p+l}E\{\lambda_l(X)\lambda_{r-p+l}(X)\}\right\}\\ &= 2f^2(x)n^{-1}\left\{\mu(x)\sum_{r=1}^{n-1}\left(1 - \frac{i}{n}\right)s_i\lambda_i(x)\right.\\ &\qquad \left. + \sum_{l=1}^{n-1}\left(1 - \frac{l}{n}\right)w_l\right\}\end{aligned} \tag{6.6}$$

for almost every $x \in R$, where

$$w_l = \sum_{i=1}^{\infty} s_i s_{i+l}E\{\lambda_i(X)\lambda_{i+l}(X)\}.$$

Recalling the fact (Cesaro summation formula) that if $\sum_{i=1}^{\infty}|g_i| < \infty$ then, $\sum_{i=1}^{n-1}(1 - (i/n))g_i$ has a finite limit as $n \to \infty$, and applying Assumptions (A2.3) and (A2.4) we find that (6.6) is of order $O(n^{-1})$. Hence, we have established that

$$L_2(x) = O\left(\frac{1}{n}\right)$$

for almost every $x \in R$.

All these considerations show that

$$\operatorname{var}(\hat{g}(x)) \approx a(x)\frac{2^m}{n} \qquad \text{as } m \to \infty \tag{6.7}$$

for almost every $x \in R$, where $a(x)$ is defined in (6.4).

Concerning the bias term for $\hat{g}(x)$ let us observe that

$$E\{\hat{g}(x)\} = E\{\mu(X_0)q_{mN}(x, X_0)\}$$

which due to Lemma 1 converges to $g(x) = \mu(x)f(x)$, for almost every $x \in R$. Hence we have proved that

$$\hat{g}(x) \to g(x) \quad \text{as } n \to \infty, \qquad \text{in probability}$$

for almost every $x \in R$ provided that (5.12) and (5.13) hold.

Regarding

$$\hat{f}(x) = n^{-1}\sum_{i=1}^{n} q_{mN}(x, X_i)$$

let us observe that

$$E\hat{f}(x) = E\{q_{mN}(x, X_0)\} \to f(x) \qquad \text{as } m \to \infty$$

for almost every $x \in R$.

Furthermore,

$$\operatorname{var}(\hat{f}(x)) = n^{-1}\operatorname{var}(q_{mN}(x, X_0)) \approx f(x)\frac{2^m}{n} - \frac{f^2(x)}{n} \tag{6.8}$$

for almost every $x \in R$.

Thus,

$$\hat{f}(x) \to f(x) \quad \text{as } n \to \infty, \qquad \text{in probability}$$

for almost every $x \in R$.

The proof of Theorem 1 is thus complete. □

VII. Convergence Rates

Theorem 1 establishes the pointwise convergence of our identification algorithms under very mild conditions on the input density $f(\bullet)$ and the unknown nonlinearity $\mu(\bullet)$. For instance, if $f(\bullet)$ is the standard $N(0,1)$ normal density, then (A3.2) is satisfied for any nonlinearity which does not grow faster than $e^{\alpha x^2}$, $\alpha < 1$. These conditions are also met if $|\mu(x) \le c_1|x| + c_2$ and $\int_{-\infty}^{\infty} x^2 f^2(x)\,dx < \infty$. Nevertheless, in order to get further insight into the behavior of our algorithms let us consider the question of the convergence rate. This, in particular, will allow us to select the locally optimal resolution level $m(n)$ yielding an asymptotically best rate of convergence. To this end we need some further local regularity conditions on $\mu(\bullet)$ and $f(\bullet)$. Hence, suppose that $\mu(\bullet)$, $f(\bullet)$ are bounded and satisfy the local Lipschitz condition at the point x_0 with exponents $\alpha \in (0,1]$ and $\beta \in (0,1]$ correspondingly, i.e.,

$$|\mu(x_0 + h) - \mu(x_0)| \le L_\mu |h|^\alpha \tag{7.1}$$

$$|f(x_0 + h) - f(x_0)| \le L_f |h|^\beta \tag{7.2}$$

where L_μ, L_f are some positive constants and h, $|h| < 1$ determines a small neighborhood around x_0.

The assumption in (7.1) says that $\mu(\bullet)$ has a fractional derivative of order α, $0 < \alpha \le 1$ at the point x_0. In particular, if $\mu(\bullet)$ has a bounded ordinary derivative at x_0, then it satisfies (7.1) with $\alpha = 1$ and $L_\mu = |\mu^{(1)}(x_0)|$. Note also that $\mu(\bullet)$ need not be continuous on R. The interpretation of (7.2) is analogous.

In order to establish the convergence rate, we need first the following fact (see the Appendix for the proof).

Lemma 3: Let $\hat{\mu}(x) = \hat{g}(x)/\hat{f}(x)$, $\hat{f}(x) \ne 0$ be a certain estimate of $\mu(x) = g(x)/f(x)$, $f(x) \ne 0$. Then, for $t > 0$,

$$|\hat{g}(x) - g(x)| \le tf(x)/(t + |\mu(x)| + 1)$$

and

$$|\hat{f}(x) - f(x)| \le tf(x)/(t + |\mu(x)| + 1)$$

imply

$$|\hat{\mu}(x) - \mu(x)| \le t.$$

By Lemma 3, the identity (11.1) used in the Proof of Lemma 3 (see the Appendix) and Chebyshev's inequality (see [13] for similar facts) we can easily establish the following result. In what follows, we say that $Z_n = O(a_n)$ in probability for a sequence of random variables $\{Z_n\}$ if $\alpha_n Z_n / a_n \to 0$ in probability as $n \to \infty$, for all sequences $\{\alpha_n\}$ convergent to zero.

Lemma 4: Let $\hat{\mu}(x) = \hat{g}(x)/\hat{f}(x)$ be an estimate of $\mu(x) = g(x)/f(x)$. Suppose that for some positive γ_f, γ_g and some point $x \in R$ we have

$$E\{\hat{f}(x) - f(x)\}^2 = O(n^{-\gamma_f})$$

and

$$E\{\hat{g}(x) - g(x)\}^2 = O(n^{-\gamma_g})$$

then

$$\hat{\mu}(x) = \mu(x) + O(n^{-\min(\gamma_f, \gamma_g)/2}) \qquad \text{in probability}$$

and

$$E\{\hat{\mu}(x) - \mu(x)\}^2 = O(n^{-\min(\gamma_f, \gamma_g)}).$$

We are now in a position to give a result concerning the local rate of convergence of $\hat{\mu}(\bullet)$. Let $[a]$ denote the integer part of a and let $\delta = \min(\alpha, \beta)$, where α and β are Lipschitz coefficients defined in (7.1) and (7.2), respectively.

Theorem 2: Let all the conditions of Theorem 1 be satisfied. Let at the point x_0 $\mu(\bullet)$ and $f(\bullet)$ meet Assumptions (7.1) and (7.2).

If the resolution level $m(n)$ is selected as

$$m(n) = \left[\frac{1}{2\delta + 1}\log_2(n)\right]$$

and if the truncation point N [number of summands in (5.3)] satisfies

$$N \ge n^{1/(2\delta+1)}|x_0|$$

then

$$\hat{\mu}(x_0) = \mu(x_0) + O(n^{-\delta/(2\delta+1)}) \qquad \text{in probability} \tag{7.3}$$

and

$$E\{\hat{\mu}(x_0)-\mu(x_0)\}^2 = O(n^{-2\delta/(2\delta+1)}). \tag{7.4}$$

Remark 7.1: Thus if the nonlinear characteristic $\mu(\bullet)$ is more rough than the input density $f(\bullet)$, i.e., $\alpha < \beta$, then the rate is determined by the smoothness of $\mu(\bullet)$. In particular, if $\alpha = \beta = 1$, then the rate is of order $O(n^{-1/3})$ in probability, or $O(n^{-2/3})$ in the mean squared-error sense, where the resolution level m can be selected as

$$m(n) = \left[\tfrac{1}{3}\log_2(n)\right]. \tag{7.5}$$

Let us observe that the rate obtained in Theorem 2 is optimal since it agrees with the best possible rate for nonparametric regression estimation established in [45].

It is also worth noting that the smoother the functions $\mu(\bullet)$ and $f(\bullet)$ are, the slower the parameters m and N grow.

Proof of Theorem 2: By recalling that $\hat{\mu}(x) = \hat{g}(x)/\hat{f}(x)$, where $\hat{g}(x)$, $\hat{f}(x)$ are defined in (5.6) and (5.7), respectively, and by using Lemma 3 we have for any $t > 0$ and x_0 such that Assumption (A1.2) is satisfied

$$\begin{aligned}&\boldsymbol{P}\{|\hat{\mu}(x_0)-\mu(x_0)| > t\}\\ &\quad\le \boldsymbol{P}\{|\hat{g}(x_0)-g(x_0)| > \bar{t}\} + \boldsymbol{P}\{|\hat{f}(x_0)-f(x_0)| > \bar{t}\}\end{aligned}$$

where $\bar{t} = tf(x_0)/(t+|\mu(x_0)|+1)$.

By this and Chebyshev's inequality, it suffices to examine the mean squared errors of $\hat{g}(x_0)$ and $\hat{f}(x_0)$. We have already shown [see (6.7) and (6.8)] that

$$\operatorname{var}\hat{g}(x_0) \approx a(x_0)\frac{2^m}{n} \tag{7.6}$$

where $a(x_0) = f(x_0)(\mu^2(x_0) + \operatorname{var}\xi_n + \sigma^2)$ and

$$\operatorname{var}\hat{f}(x_0) \approx f(x_0)\frac{2^m}{n}. \tag{7.7}$$

Hence, only the bias terms of $\hat{g}(x_0)$ and $\hat{f}(x_0)$ must be considered.

Let us first observe that if $\mu(\bullet)$ and $\hat{f}(\bullet)$ satisfy the conditions (7.1) and (7.2), then for the function $g(x) = \mu(x)f(x)$ we have

$$|g(x_0+h)-g(x_0)| \le L_g|h|^\delta$$

with $\delta = \min(\alpha,\beta)$ and $L_g = M_fL_\mu + M_\mu L_f$, where $M_f = \sup_x f(x)$, $M_\mu = \sup_x|\mu(x)|$, i.e., $g(\bullet)$ is also Lipschitz with the exponent δ.

In the Proof of Lemma 1 (see the Appendix) we have already observed that for $N \ge |x_0|2^m$ we have

$$\begin{aligned}&E\hat{g}(x_0)\\ &\quad= \int_{-\infty}^{\infty} q_{mN}(x_0,z)g(z)\,dz = \frac{1}{|A_m(x_0)|}\int_{A_m(x_0)} g(z)\,dz\end{aligned}$$

for $A_m(x_0)$ being one of the intervals $\{[k/2^m, (k+1)/2^m), |k| \le N\}$ where x_0 belongs to and $|A_m(x_0)| = 2^{-m}$. By this we have

$$\begin{aligned}|E\hat{g}(x_0)-g(x_0)| &\le \frac{1}{|A_m(x_0)|}\int_{A_m(x_0)}|g(z)-g(x_0)|\,dz\\ &\le L_g2^{-m\delta}.\end{aligned} \tag{7.8}$$

By the analogous considerations we can infer that

$$|E\hat{f}(x_0)-f(x_0)| \le L_f2^{-m\beta}. \tag{7.9}$$

Hence (7.6)–(7.9) yield the following:

$$E\{\hat{g}(x_0)-g(x_0)\}^2 = O\left(\frac{2^m}{n}\right) + O(2^{-2m\delta})$$

and

$$E\{\hat{f}(x_0)-f(x_0)\}^2 = O\left(\frac{2^m}{n}\right) + O(2^{-2m\beta}).$$

Direct minimization of these expressions with respect to m and Lemma 4 conclude the Proof of Theorem 2. □

Remark 7.2: Lemma 4 [see the identity in (11.1)] allows us to examine the exact local (at given x_0) asymptotic rate of the mean-squared error for $\hat{\mu}(x_0)$ provided that $\mu(\bullet)$ and $f(\bullet)$ satisfy some stronger smoothing conditions than in (7.1) and (7.2). In fact, let $\mu(\bullet)$ and $f(\bullet)$ possess two derivatives, with the properties that $\mu^{(1)}(\bullet)$ and $f^{(1)}(\bullet)$ are continuous at the point x_0, and $\mu^{(2)}(\bullet)$ and $f^{(2)}(\bullet)$ are bounded on R. Let the point x_0 belong to the interval $[k/2^m, (k+1)/2^m)$, i.e., it can be represented as $x_0 = (k+\varrho)/2^m$ for some $0 \le \varrho < 1$. Then using the aforementioned results and borrowing some rather complicated techniques from [35], we can show after some algebra that

$$E\hat{\mu}(x_0) = \mu(x_0) - \frac{(\varrho-1/2)}{2^m}\mu^{(1)}(x_0) + O(2^{-2m}) \tag{7.10}$$

and

$$\operatorname{var}\hat{\mu}(x_0) \approx \frac{2^m}{n}\frac{\operatorname{var}(\xi_n)+\sigma^2}{f(x_0)}.$$

This yields the following exact asymptotic formula for the mean-squared error of $\hat{\mu}(x_0)$:

$$\begin{aligned}&E(\hat{\mu}(x_0)-\mu(x_0))^2\\ &\quad\approx \frac{2^m}{n}\frac{\operatorname{var}(\xi_n)+\sigma^2}{f(x_0)} + \frac{((\varrho-1/2)\mu^{(1)}(x_0))^2}{2^{2m}}.\end{aligned} \tag{7.11}$$

The direct minimization of this expression yields the following formula for the optimal resolution level:

$$m(n) = \left[\tfrac{1}{3}\log_2(n) + \tfrac{1}{3}\log_2(\gamma(x_0))\right] \tag{7.12}$$

where

$$\rho(x_0) = \frac{2((\varrho-1/2)\mu^{(1)}(x_0))^2f(x_0)}{\operatorname{var}(\xi_n)+\sigma^2}. \tag{7.13}$$

Hence (7.12), contrary to (7.5), gives the optimal local [depending on the pointwise properties of $\mu(x)$ and $f(x)$] value of the resolution degree. See also Section VIII for further discussion concerning the choice of the resolution level. Plugging (7.12) into (7.11) gives the optimal value of the resulting local error

$$\begin{aligned}&E\{\hat{\mu}(x)-\mu(x)\}^2\\ &\quad\approx 32^{-2/3}\left(\frac{(\varrho-1/2)\mu^{(1)}(x_0)(\operatorname{var}(\xi_n)+\sigma^2)}{f(x_0)}\right)^{2/3}n^{-2/3}.\end{aligned} \tag{7.14}$$

It is worth noting that if x_0 is the middle point of the interval $[k/2^m, (k+1)/2^m)$, i.e., if $\varrho = 1/2$, then the bias in (7.10) is of order $O(2^{-2m})$ and we obtain the faster rate $O(n^{-4/5})$, where $m(n)$ is selected as $m(n) = [\frac{1}{5}\log_2(n)]$. Note that the middle point of $[k/2^m, (k+1)/2^m)$ corresponds to the discontinuity point of the wavelet orthonormal function $\psi_{mk}(x)$ defined in Section II [see (2.18)].

Remark 7.3: Theorem 2 and Remark 7.2 establish the pointwise rate of convergence of our estimate $\hat{\mu}(x)$ for a large class of nonlinear characteristics $\mu(\bullet)$ and input densities $f(\bullet)$.

Let us also consider the most comfortable situation for our estimation techniques, i.e., when both $\mu(\bullet)$ and $f(\bullet)$ belong to the Haar multiresolution class V_m, i.e., the class (see Section II) of all piecewise constant functions with possible jumps at the integer multiple of 2^{-m}, where now m is a fixed integer. It is important to observe that this also implies $g = \mu f \in V_m$. It is then clear that both $\hat{g}(x)$ and $\hat{f}(x)$ are unbiased estimators, i.e., $E\hat{g}(x) = g(x)$, $E\hat{f}(x) = f(x)$, provided that the truncation value N satisfies (5.10). This, however, does not imply that $E\hat{\mu}(x) = \mu(x)$. The latter is due to the fact that $\hat{\mu}(x)$ has the ratio form, i.e., $\hat{\mu}(x) = \hat{g}(x)/\hat{f}(x)$. Nevertheless, arguing, as in Remark 7.2, we can conclude that if μ, $f \in V_m$ then

$$E\hat{\mu}(x) = \mu(x) + O(n^{-4/3}).$$

Hence the estimator bias is greatly reduced, i.e., we have the error $O(n^{-4/3})$ instead of $O(n^{-1/3})$ as in (7.10).

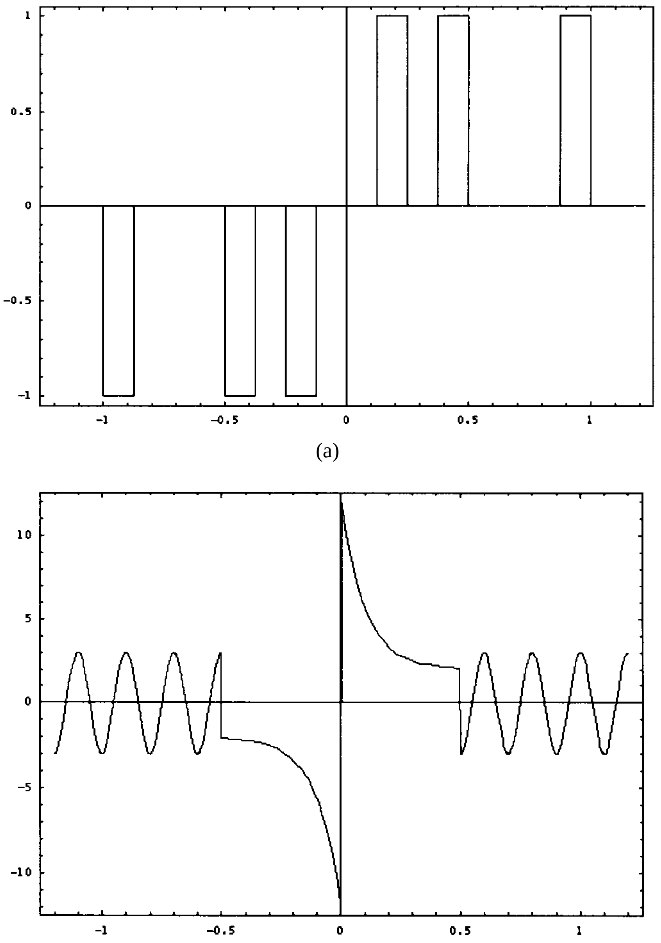
VIII. Selecting Resolution Level

The discussion in Section VII reveals (see in particular Remark 7.2) the importance of proper selection of the resolution level $m(n)$. The formula given in (7.12) gives an asymptotically optimal value of $m(n)$. The function $\gamma(x)$ defined in (7.13) depends, however, on some unknown characteristics of the system, i.e., on $\mu^{(1)}(x)$, $f(x)$, $\mathrm{var}(\xi_n)$, and σ^2. Some pilot estimates of these quantities would lead to a plug-in formula for the resolution level. The input density $f(x)$ can be estimated by the estimator $\hat{f}(x)$ given in (5.7). This estimator depends, however, also on $m(n)$, and we would recommend to use the value suggested in Remark 7.1, i.e., $m(n) = [\frac{1}{3}\log_2(n)]$. With such a value of $m(n)$, $\hat{f}(x)$ is a consistent estimate of $f(x)$. The nonlinear characteristic derivative $\mu^{(1)}(x)$ could be estimated by

$$\frac{\hat{\mu}(x+\Delta) - \hat{\mu}(x-\Delta)}{2\Delta}$$

with some appropriately defined Δ, e.g., $\Delta = n^{-1/3}$ based on the discussion in Remark 7.1. Estimation of the variance values $\mathrm{var}(\xi_n)$ and σ^2 can be specified experimentally. Although this procedure could, eventually, be implemented, it does not seem to be practical and requires too many arbitrary parameters to choose.

Let us propose an alternative approach based again on the formulas in (7.12) and (7.13) and some prior knowledge about the system characteristics.



(a)

(b)

Fig. 9. Nonlinear characteristics: (a) the nonlinearity in (9.1) and (b) the nonlinearity in (9.4).

Let us assume that the nonlinear characteristic $\mu(\bullet)$ satisfies for every $x \in R$ the following condition:

$$|\mu^{(1)}(x)| \le \overline{\mu}$$

for some positive $\overline{\mu}$, i.e., $\mu(\bullet)$ has a bounded variability. Let also $\sup_x f(x) = M_f$. It is then clear that

$$\gamma(x) = \frac{2((\varrho - 1/2)\mu^{(1)}(x))^2 f(x)}{\mathrm{var}(\xi_n) + \sigma^2} \le \frac{\overline{\mu}^2 M_f}{2\sigma^2}.$$

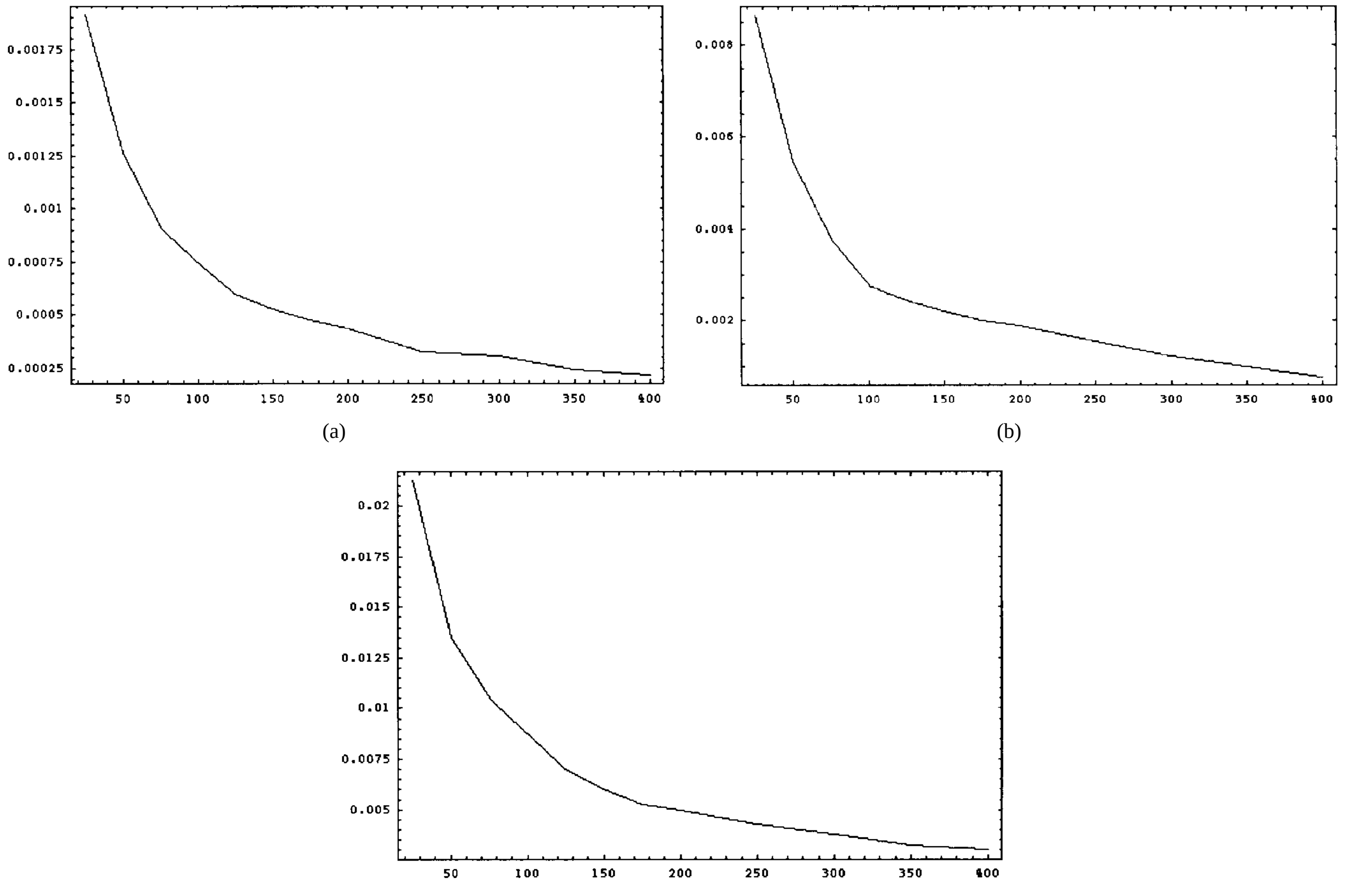
Hence, provided that values $\overline{\mu}$, M_f, σ^2 are known (which can be the case in a number of practical situations), we can choose $m(n)$ as

$$m(n) = \left[\frac{1}{3}\log_2(n) + \frac{1}{3}\log_2\left(\frac{\overline{\mu}^2 M_f}{2\sigma^2}\right)\right]. \tag{8.1}$$

Note that this is a rather pessimistic choice of $m(n)$, i.e., it is larger than the optimal value minimizing the mean squared error $E\{\hat{\mu}(x) - \mu(x)\}^2$.

A fully automatic choice of $m(n)$ can rely on the cross validation methodology where $m(n)$ is selected as a minimum of the so-called prediction error

$$CV(m) = n^{-1}\sum_{j=1}^{n}(Y_j - \hat{\mu}_{-j}(X_j))^2$$



(a)

(b)

(c)

Fig. 10. *Error* versus n for the nonlinearity in (9.1): (a) memoryless system, (b) cascade system, and (c) parallel system.

where $\hat{\mu}_{-j}(x)$ is the version of $\hat{\mu}(x)$ calculated from all the data points except the jth; see [24] for some alternative cross-validation methods of choosing smoothing parameters. Hence the data dependent $m(n)$ is selected as

$$\hat{m}(n) = \arg\min_{m} CV(m). \tag{8.2}$$

This technique requires a considerable amount of computing as the estimate has to be formed n times. Furthermore, the minimum in (8.2) is taken over all integers. In order to reduce this computational burden, we propose to minimize $CV(m)$ over a certain range of the resolution levels $m_{\min}(n) \leq m(n) \leq m_{\max}(n)$. As has already been noted (see Remark 7.2), under the most preferable circumstances $m(n)$ can be selected as low as $[\frac{1}{5}\log_2(n)]$. This can be used as a lower bound for $m(n)$, i.e.,

$$m_{\min}(n) = \left[\tfrac{1}{5}\log_2(n)\right].$$

As the upper bound for $m(n)$ we can use either $m_{\max}(n) = [\frac{1}{3}\log_2(n)]$ or the formula given in (8.1). All these considerations lead to the following choice of $m(n)$:

$$\hat{m}(n) = \arg \min_{m\in[m_{\min}(n), m_{\max}(n)]} CV(m).$$

Hence, the optimal $m(n)$ can be found by calculating $CV(m)$ from the coarsest scale $m_{\min}(n)$ to the finest one $m_{\max}(n)$. Note that the cross-validation choice produces the global value of $m(n)$, i.e., the value which is independent of x at which $\hat{\mu}(x)$ is computed.

IX. Simulation Examples

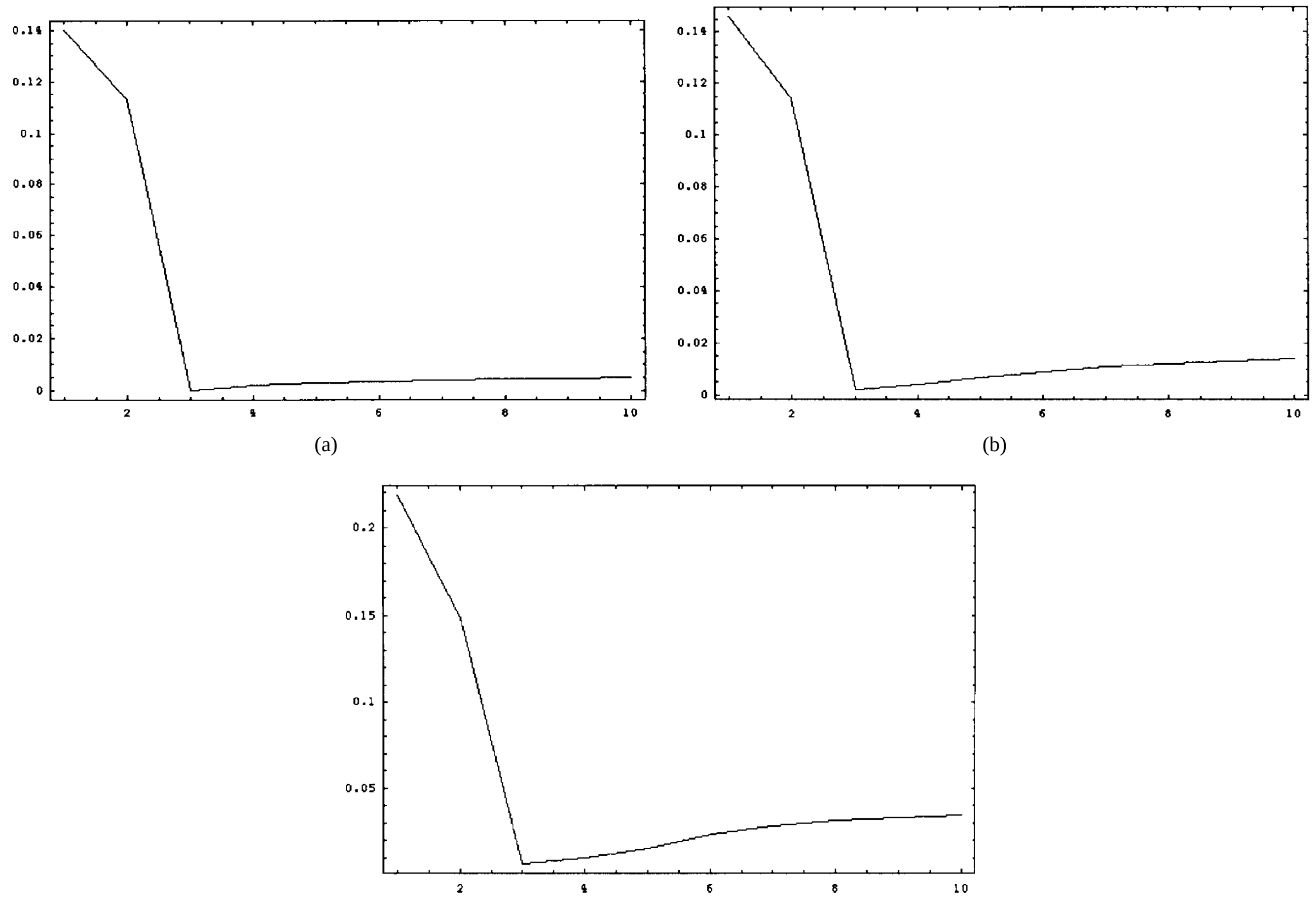
To evaluate the accuracy of our identification algorithms for small and moderate sample sizes, we perform some simulation studies. In all our experiments the input signal $\{X_n\}$ is uniformly distributed over the interval $[-3, 3]$. The measurement noise $\{\varepsilon_n\}$ is also uniformly distributed in $[-0.1, 0.1]$. The range of the input signal implies that we can specify the truncation parameter N, see (5.10), as $N = 3 \times 2^m$, where m is the resolution level. The efficacy of the identification procedure $\hat{\mu}(x)$ [see (5.3), (5.4), and (5.9) for various equivalent forms of $\hat{\mu}(x)$] is measured by the following criterion:

$$\mathit{Error} = n^{-1}E\left\{\sum_{j=1}^{n}|\hat{\mu}(X_j) - \mu(X_j)|^2\right\}$$

where the expectation sign is realized in simulations by averaging over 20 different training samples each of the size n.

In the first experiment a nonlinearity [see Fig. 9(a)]

$$\theta(x) = \begin{cases} 1, & \text{if } x \in [0.125, 0.25] \cup [0.375, 0.5] \cup [0.875, 1] \\ -1, & \text{if } x \in [-0.125, -0.25] \cup [-0.375, -0.5] \\ & \qquad \cup\ [-0.875, -1] \\ 0, & \text{otherwise} \end{cases} \tag{9.1}$$



(a)

(b)

(c)

Fig. 11. *Error* versus m for nonlinearity in (9.1): (a) memoryless system, (b) cascade system, and (c) parallel system.

is used in three different settings, i.e., for the memoryless, cascade, and parallel models, see Examples 1, 2, 3 in Section IV, respectively. Since $\theta(x)$ in (9.1) is piecewise constant, this is an example of the nonlinearity well adapted to the Haar multiresolution basis.

The cascade and parallel models are described, respectively, by the following state equations:

$$\begin{cases} O_n &= -0.2O_{n-1} + \theta(X_n) \\ Y_n &= O_n + \varepsilon_n \end{cases} \tag{9.2}$$

$$\begin{cases} O_n &= -0.2O_{n-1} + X_n \\ Y_n &= O_n + \theta(X_n) + \varepsilon_n. \end{cases} \tag{9.3}$$

It is worth nothing that we have $E\theta(X_n) = EX_n = 0$ and therefore $\hat{\mu}(x)$ and $\hat{\mu}(x) - x$ are consistent estimates of the nonlinearity $\theta(x)$ in the cascade and parallel models, respectively (see discussion in Section IV).

Fig. 10 depicts the *Error* as a function of the sample size n. It is seen that the *Error* for the memoryless model is the smallest. Surprisingly the *Error* for the cascade model is about 2–3 times smaller than that of the parallel structure.

The value of the resolution level m has been set to 3 in the all three cases. This is due to the fact that this value minimizes the *Error* for a small and moderate number of observations. In fact, Fig. 11 displays the *Error* versus m for $n = 150$ observations. A clear global minimum at $m = 3$ is seen. Hence, the optimal partition of the x-axis is $1/2^m = 1/8$. Note that this agrees with the structure of the nonlinearity in (9.1) which is constant on the intervals of the size 1/8.

In the second experiment, a nonlinearity in (9.2) and (9.3) not well adapted to the Haar basis has been selected, i.e.,

$$\theta(x) = \begin{cases} 10\exp(-10x) + 2, & \text{for } 0 \le x \le 0.5 \\ 3\cos(10\pi x), & \text{for } x > 0.5 \end{cases} \tag{9.4}$$

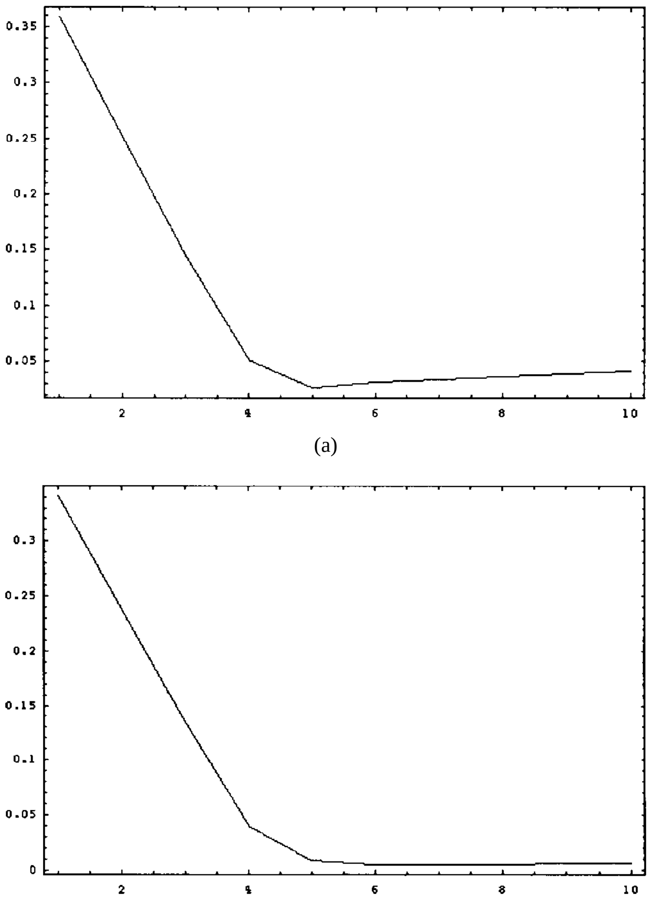
and $-\theta(-x)$ for $x < 0$. Fig. 9(b) displays this nonlinearity.

Fig. 12 depicts the *Error* versus m for the cascade and parallel structures based on $n = 150$ observations. Since the nonlinearity is not well suited for the Haar basis, the larger m is required; the optimal m equals 5 (cascade model) and 6 (parallel model). For the memoryless model, that value is even larger and equals 10. The overall performance of $\hat{\mu}(x)$ is now considerably poorer.

Finally the model discussed in Example 5 of Section IV has been taken into account, i.e., the model represented by the following equation:

$$\begin{cases} O_n = -0.2O_{n-1} + \theta(X_n) + \theta_0(X_{n-1}) \\ Y_n = O_n + \varepsilon_n \end{cases} \tag{9.5}$$

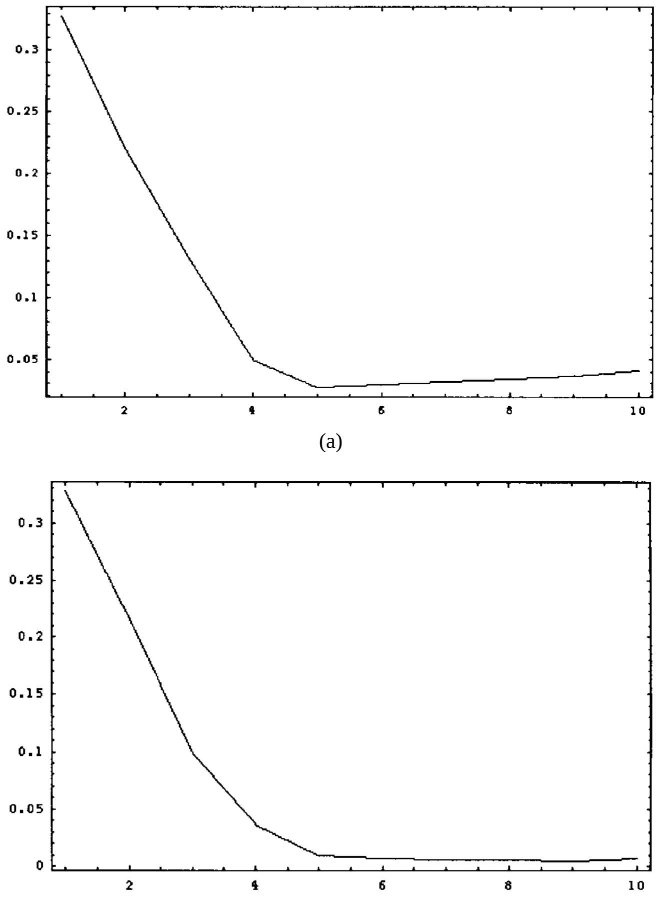
where $\theta_0(x) = 0.1x^3$ is the nuisance nonlinearity and $\theta(x)$ is defined as in (9.4). Fig. 13 displays the *Error* versus m. An optimal resolution level is equal to 5 for $n = 150$ observations.



(a)

(b)

Fig. 12. *Error* versus m for the nonlinearity in (9.4): (a) cascade system and (b) parallel system.



(a)

(b)

Fig. 13. *Error* versus m for the system with two nonlinearities.

In the same figure [Fig. 13(b)] we show the *Error* of the version of (9.5) where the dynamical subsystem is set to zero (i.e., the value -0.2 in (9.5) is replaced by 0). Let us observe that an optimal m is now considerably greater and equals 9. This reveals that the presence of dynamical subsystems in composite models to some extent helps in identification, and it greatly influences the accuracy of identification algorithms for recovering nonlinear elements.

X. Concluding Remarks

In this paper we have developed the Haar multiresolution identification algorithm for recovering nonlinearities in a broad class of block structured nonlinear systems. The introduced class of systems includes known cascade and parallel structures as well as some new connections. No *a priori* information about the nonlinear characteristics and input signal probability density function is required, making the identification problem nonparametric. Using the concept of the standard regression function, the nonparametric Haar multiresolution identification algorithm is formed and its rigorous convergence properties are proved. In particular, the best possible local rate of convergence and optimal selection of the resolution degree are established. The convergence results hold under very mild restrictions on the nonlinear characteristic and the input density function as well as on the system dynamics. Besides these theoretical properties, our algorithm is very easy and fast to compute. In fact, observe that to memorize our estimate [see (5.3)] it suffices to store (for every x) $2N+1$ numbers $\hat{a}_{mk}$ and $\hat{b}_{mk}$, where N increases much slower than n. Indeed, in Theorem 2 we have shown that N is of order $n^{1/(2\delta+1)}$, where the parameter δ controls the smoothness of the nonlinear characteristic $\mu(\bullet)$ and the input density $f(\bullet)$. The smoother $\mu(\bullet)$ and $f(\bullet)$ are, the smaller the truncation value N is. Alternatively, the version of $\hat{\mu}(x)$ in (5.9) can be used for calculations. Here only a simple binning process is required for determination of $\hat{\mu}(x)$.

Some further studies could be carried out by combining the multiresolution approach with the concept of wavelet basis. As has already been mentioned (see Section II), the resolution space V_{m+1} can be decomposed as $V_{m+1} = V_m \oplus W_m$, where W_m is the wavelet subspace equipped with the orthonormal set $\{2^{m/2}\psi(2^m x - k\}_{k\in Z}$, $\psi(\bullet)$ is the wavelet function. A simple iteration of the decomposition leads to

$$V_{m_0+r} = V_{m_0} \oplus W_{m_0} \oplus W_{m_0+1} \oplus \cdots \oplus W_{m_0+r-1}$$

for any integers m_0 and $r > 0$.

Hence a function $f \in L_2(R)$ can be approximated at the resolution $2^{-(m_0+r)}$ as follows:

$$f_{m_0+r}(x) = \sum_{k\in Z} a_{m_0k}\varphi_{m_0k}(x) + \sum_{s=m_0}^{m_0+r-1} \sum_{k\in Z} c_{sk}\psi_{sk}(x) \tag{10.1}$$

i.e., $f_{m_0+r}(x)$ is the orthogonal projection of f onto the resolution subspace V_{m_0+r}.

The first term in (10.1) represents our initial guess, whereas the second one adds further layers of information about $f(\bullet)$. An empirical version of (10.1) (with estimated coefficients $\{a_{m_0k}\}$, $\{c_{sk}\}$) would be an attractive alternative to our estimation technique. It is clear that here both m_0 (the initial resolution level) and r (the additional number of resolution layers) should be appropriately specified in order to establish consistency results. We conjecture that the choice of m_0 is less critical than our $m(n)$—the resolution degree specifying the estimators studied in this paper. We refer to [22] and the references cited therein for some results on probability density estimation techniques employing the representation in (10.1).

In this paper we have used the classical Haar multiresolution analysis. The Haar basis is very well localized and easily understood since the supports of the basis functions are dyadic intervals. Furthermore they are step functions, making them well adapted to discontinuous characteristics. For continuous characteristics, however, one should use smooth scaling functions developed by Daubechies [6]. It seems that our results can be extended to this case as well. Nevertheless, the smooth scaling functions are rarely given in an explicit form, and they have to be numerically determined from the scaling equation (2.9). An interesting situation arises when one deals with characteristics of the mixed nature being, e.g., a piecewise continuous. In such a case, a multiresolution basis which is a certain combination of the Haar system and smooth multiresolution functions could be employed.

Appendix

Proof of Lemma 1: Let us observe that for $N \geq 2^m|x|$ and the kernel $q_{mN}(x,y)$ defined in (5.8), we have

$$\begin{aligned}\int_{-\infty}^{\infty} & q_{mN}(x,z)w(z)\,dz \\ &= 2^m \sum_{|k|\leq N} \mathbf{1}_{[k/2^m,\,(k+1)/2^m]}(x) \int_{k/2^m}^{(k+1)/2^m} w(z)\,dz \\ &= \frac{1}{|A_m(x)|}\int_{A_m(x)} w(z)\,dz\end{aligned}$$

where $A_m(x)$ is one of the $\{[k/2^m,(k+1)/2^m), |k| \leq N\}$ intervals where x falls and $|A|$ denotes the measure of the set A.

A straightforward application of the Lebesque density theorem (see [49, pp. 108–109]), gives

$$\begin{aligned}\frac{1}{|A_m(x)|} & \int_{A_m(x)} w(z)\,dz \\ & \to w(x) \qquad \text{as } m \to \infty \qquad \text{for almost every } x.\end{aligned}$$

The Proof of Lemma 1 is thus complete. □

Proof of Lemma 2: Lemma 2 results straightforwardly from the observation that

$$\int_{-\infty}^{\infty} q_{mN}^2(x,z)w(z)\,dz = 2^m \int_{-\infty}^{\infty} q_{mN}(x,z)w(z)\,dz$$

and application of Lemma 1. □

Proof of Lemma 3: We begin with the following identity:

$$\frac{\hat{g}}{\hat{f}} = \mu + \mu\frac{(f-\hat{f})}{f} + \frac{(\hat{g}-g)}{f} + \left(\frac{\hat{g}}{\hat{f}} - \mu\right)\frac{(f-\hat{f})}{f} \qquad (11.1)$$

where $\mu = g/f$.

Thus for $t > 0$ and if

$$|\hat{g} - g| \leq tf/(t+|\mu|+1) \quad \text{and} \quad |\hat{f} - f| \leq tf(t+|\mu|+1)$$

we have

$$\left|\frac{\hat{g}}{\hat{f}} - \mu\right| \leq t.$$

The Proof of Lemma 3 is complete. □

Acknowledgment

The authors thank the reviewers for many helpful suggestions.

References

[1] A. Antoniadis and G. Oppenheim, Eds., *Wavelets and Statistics, Lecture Notes in Statistics*, vol. 103. New York: Springer-Verlag, 1995.

[2] J. S. Bendat, *Nonlinear System Analysis and Identification*. New York: Wiley, 1990.

[3] S. A. Billings, “Identification of nonlinear systems—A survey,” *Proc. Inst. Elect. Eng.*, vol. 127, pp. 272–285, 1980.

[4] D. R. Brillinger, “The identification of a particular nonlinear time series system,” *Biometrika*, vol. 64, pp. 509–515, 1977.

[5] H. W. Chen, “Modeling and identification of parallel nonlinear systems: Structural classification and parameter estimation methods,” *Proc. IEEE*, vol. 83, pp. 39–66, 1995.

[6] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia: SIAM, 1992.

[7] L. Devroye and L. Györfi, *Nonparametric Density Estimation: The L_1 View*. New York: Wiley, 1985.

[8] D. L. Donoho and I. M. Johnstone, “Ideal spatial adaptation by wavelet shrinkage,” *Biometrika*, vol. 81, pp. 425–455, 1994.

[9] J. Engel, “Density estimation with Haar series,” *Stat. Probab. Lett.*, vol. 9, pp. 111–117, 1990.

[10] E. Eskinat, S. H. Johnson, and W. L. Luyben, “Use Hammerstein models in identification of nonlinear systems,” *Amer. Inst. Chem. Eng.*, vol. 37, pp. 255–268, 1991.

[11] R. L. Eubank, *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker, 1988.

[12] W. Greblicki, “Nonparametric identification of Wiener systems by orthogonal series,” *IEEE Trans. Automat. Contr.*, vol. 10, pp. 2077–2086, 1994.

[13] W. Greblicki and M. Pawlak, “Fourier and Hermite series estimates of regression functions,” *Ann. Inst. Statist. Math.*, vol. 37A, pp. 443–454, 1985.

[14] ——, “Identification of discrete Hammerstein systems using kernel regression estimates,” *IEEE Trans. Automat. Contr.*, vol. 31, pp. 74–77, 1986.

[15] ——, “Cascade nonlinear system identification by a nonparametric method,” *Int. J. Syst. Sci.*, vol. 25, pp. 129–153, 1994.

[16] ——, “Nonparametric identification of Hammerstein systems,” *IEEE Trans. Inform. Theory*, vol. 35, pp. 409–418, 1989.

[17] ——, “Nonparametric identification of a cascade nonlinear time series system,” *Signal Processing*, vol. 22, pp. 61–75, 1991.

[18] , “Nonparametric identification of a particular nonlinear time series system,” *IEEE Trans. Signal Processing*, vol. 40, pp. 985–989, Apr. 1992.

[19] ——, “Dynamic system identification with order statistics,” *IEEE Trans. Inform. Theory*, vol. 40, pp. 1474–1489, 1994.

[20] ——, “Nonparametric recovering nonlinearities in block-oriented system with the help of Laguerre polynomials,” *Contr. Theory Advanced Technol.*, vol. 10, pp. 771–791, 1994.

[21] R. Haber and H. Unbenhauen, “Structure identification of nonlinear dynamic systems—A survey on input/output approaches,” *Automatica*, vol. 26, pp. 651–677, 1990.

[22] P. Hall and P. Patil, “Formulae for mean integrated squared error of nonlinear wavelet-based density estimators,” *Ann. Stat.*, vol. 23, pp. 905–928, 1995.

[23] W. Härdle, *Applied Nonparametric Regression*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[24] W. Härdle, P. Hall, and J. S. Marron, "How far are automatically chosen regression smoothing parameters from their optimum? (with discussion)," *J. Amer. Stat. Assoc.*, vol. 83, pp. 86–101, 1988.

[25] I. W. Hunter and M. J. Korenberg, "The identification of nonlinear biological systems: Wiener and Hammerstein cascade models," *Biol. Cybern.*, vol. 55, pp. 135–144, 1986.

[26] A. Juditsky, H. Hjalmaarsson, A. Benveniste, B. Delyon, L. Ljung, J. Sjoberg, and Q. Zhang, "Nonlinear black-box models in system identification: Mathematical foundations," *Automatica*, vol. 31, pp. 1725–1750, 1995.

[27] S. E. Kelly, M. A. Kon, and L. A. Raphael, "Pointwise convergence of wavelet expansions," *Bull. Amer. Math. Soc.*, vol. 30, pp. 87–94, 1994.

[28] A. Krzyzak, "Identification of discrete Hammerstein systems by the Fourier series regression estimate," *Int. J. Syst. Sci.*, vol. 20, pp. 1729–1744, 1989.

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

[30] S. Mallat, "Multiresolution approximation and wavelets," *Trans. Amer. Math. Soc.*, vol. 315, pp. 69–88, 1989.

[31] H. E. Liao and W. S. Sethares, "Suboptimal identification of nonlinear ARMA models using an orthogonality approach," *IEEE Trans. Circuits Syst. I*, vol. 42, pp. 14–22, 1995.

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

[33] P. Z. Marmarelis and V. Z. Marmarelis, *Analysis of Physiological Systems. The White Noise Approach*. New York: Plenum, 1978.

[34] J. P. Norton, *An Introduction to Identification*. London, U.K.: Academic, 1986.

[35] M. Pawlak, "On the almost everywhere properties of the kernel regression estimate," *Ann. Inst. Statist. Math.*, vol. 43, pp. 311–326, 1991.

[36] ——, "On the series expansion approach to the identification of Hammerstein systems," *IEEE Trans. Automat. Contr.*, vol. 36, pp. 763–767, 1991.

[37] M. Pawlak and W. Greblicki, "On nonparametric identification of cascade nonlinear systems," in *Proc. 33rd Conf. Decision Contr.*, 1994, pp. 2866–2867.

[38] B. L. S. Prakasa Rao, *Nonparametric Functional Estimation*. New York: Academic, 1983.

[39] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Mag.*, vol. 8, pp. 14–38, 1991.

[40] L. Rutkowski, "A general approach for nonparametric fitting of functions and their derivatives with application to linear circuits identification," *IEEE Trans. Circuits Syst.*, vol. 33, pp. 812–818, 1986.

[41] L. Rutkowski and E. Rafajlowicz, "On global rate of convergence of some nonparametric identification procedures," *IEEE Trans. Automat. Contr.*, vol. 34, pp. 1089–1091, 1989.

[42] I. W. Sanberg, "Approximation theorems for discrete-time systems," *IEEE Trans. Circuits Syst.*, vol. 38, pp. 564–566, 1991.

[43] D. W. Scott, *Multivariate Density Estimation. Theory, Practice, and Visualization*. New York: Wiley, 1992.

[44] T. Söderström and P. Stoica, *System Identification*. Englewood Cliffs, NJ: Prentice Hall, 1989.

[45] C. J. Stone, "Optimal rates of convergence for nonparametric estimators," *Ann. Stat.*, vol. 8, pp. 1348–1360, 1980.

[46] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice Hall, 1995.

[47] G. Walter, *Wavelets and Other Orthogonal Systems with Applications*. Boca Raton, FL: CRC Press, 1994.

[48] M. P. Wand and M. C. Jones, *Kernel Smoothing*. London: Chapman and Hall, 1995.

[49] R. L. Wheeden and A. Zygmund, *Measure and Integral*. New York: Marcel Dekker, 1977.

Miroslaw Pawlak (M'85) received the M.Sc. and Ph.D. degrees in computer engineering from the Technical University of Wroclaw, Poland, in 1978 and 1982, respectively.

In 1982 he joined the Department of Technical Cybernetics, Technical University of Wroclaw, as an Assistant Professor. From 1984 to 1985 he was a visiting Research Associate at the Department of Computer Science, Concordia University, Montreal, Canada. Since 1985 he has been with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Canada, where he is currently an Associate Professor. During 1992 he was at the Technical University of Wroclaw and at the University of Ulm, Germany, as an Alexander von Humboldt Foundation Fellow. He has also held visiting appointments at the University of California at Irvine, the University of Pittsburgh, the University of North Carolina at Chapel Hill, the University of Ulm, Germany, and Vadrina University, Frankfurt, Germany. His research interests are in statistical signal/image processing, pattern recognition, nonlinear system modeling and estimation problems in telecommunication systems. He is an associate editor for *Pattern Recognition* journal.

Zygmunt Hasiewicz was born in 1948 in Poland. He received the M.Sc. and Ph.D. degrees in control engineering from the Technical University of Wroclaw, Poland, in 1971 and 1974, respectively, and the D.Sc. degree from the Technical University of Warsaw, Poland, in 1993.

In 1971 he joined the Institute of Engineering Cybernetics, Technical University of Wroclaw, where he is currently an Associate Professor. In 1976 he held visiting appointments at Lille University, France, and in 1995 he was a visiting professor at the University of Manitoba, Winnipeg, Canada. His research interests include nonlinear system modeling and statistical methods in complex system identification. He is a reviewer for *Zentralblatt für Mathematik*.