Non-parametric estimation of non-linearity in a cascade time-series system by multiscale approximation

(revised version)

Zygmunt Hasiewicz

Institute of Engineering Cybernetics
Wrocław University of Technology
Wrocław, Poland

Mailing address:

Institute of Engineering Cybernetics,
Wrocław University of Technology,
Janiszewskiego 11/17,
50-372 Wrocław, Poland

phone: (48 71) 320 32 77;
fax: (48 71) 321 26 77;
e-mail: zhas@ict.pwr.wroc.pl

List of symbols:

$\phi(x)$ - scaling function
m - scale factor (resolution level); $m \in \mathrm{Z}$ - the set of integers
n - translation factor, $n \in \mathrm{Z}$
V_m - resolution space associated with the scale m
$\{V_m, m \in \mathrm{Z}\}$ - multiscale analysis of $L^2(\mathbb{R})$
$\{\phi_{mn}(x), n \in \mathrm{Z}\}$ - orthonormal basis of V_m
$F(x;m)$ - approximation of $F(x)$ in V_m
N - number of measurement data
$f(x)$ - probability density of x
$R(x)$ - non-linearity to be recovered
$R_N(x;m)$ - estimate of $cR(x)$ at the resolution level m
$a_{mn,N}$, $b_{mn,N}$ - estimates of coefficients a_{mn}, b_{mn}
$\{\lambda_p; p = 0, 1, \ldots\}$ - impulse response of linear subsystem dynamics
$\{\omega_p; p = 0, 1, \ldots\}$ - impulse response of noise filter
d - delay in the system

Number of pages: 19

Number of figures: 2

Running head: Non-parametric estimation of non-linearity by multiscale approximation

Keywords: Hammerstein system; Non-linearity recovering; Non-parametric approach; Multiscale approximation; Multiscale orthogonal series; Convergence analysis

Abstract

The paper addresses the problem of using multiscale approximation for the identification of non-linearities in Hammerstein systems. The exciting signals are random, stationary and white, with a bounded (unknown) probability density function, and system outputs are corrupted by a zero-mean stationary random noise - white or coloured. The *a priori* information is poor. In particular no parametric form of the non-linear characteristics is known in advance. To recover non-linearities, a class of non-parametric identification algorithms is proposed and investigated. The algorithms use only input-output measurements and are based on multiscale orthogonal approximations associated with scaling functions of compact support. We establish the pointwise weak consistency of such routines along with asymptotic rates of convergence. In particular, local ability of the algorithms to discover non-linear characteristics in dependence on local smoothness of the identified non-linearity, input density and the scaling function is examined. It is shown that under mild requirements the routines attain optimal rate of convergence. The form and convergence of the algorithms are insensitive to correlation of the noise.

1. Introduction

We consider the problem of recovering non-linear characteristics in complex time-series systems composed of a memoryless non-linear element and a discrete-time linear dynamic part connected in a cascade. The systems, called the Hammerstein systems, are driven by stationary white random input signals with a bounded probability density function and are disturbed by additive, white or coloured, zero-mean stationary random noise. The approach presented in the paper is suited to the case when *a priori* knowledge about the system non-linearity is small and only of qualitative nature. We merely assume that non-linearity is bounded. In such a case no finite-dimensional parametric representation of a possible non-linear characteristic can be reasonably motivated and standard parametric methods are not applicable.

To recover non-linear characteristics we propose a class of non-parametric identification algorithms which exploit only input-output observations collected from the systems and are based, on the one hand, on the idea of non-parametric regression function estimation (e.g. [11, 23]) and, on the other, on the theory of multiscale orthogonal approximations of square integrable functions, being the leading concept of wavelet theory ([8, 10, 35, 47, 50] for instance). In particular, a class of multiscale orthogonal expansions associated with scaling functions of compact support is applied to construct the algorithms.

The obtained identification routines possess good localization and parsimony properties, i.e. for reconstruction of non-linearities use only 'local' data lying in a small neighbourhood of a point at which the estimation is carried out and for practical implementation need only a number of coefficients, being merely a small fraction of the whole number of measurements employed for identification. In this sense they combine the advantages of kernel and orthogonal series algorithms worked out earlier (see the papers cited hereafter). Moreover, due to the multiple resolution ability, the algorithms allow in an easy way identification of the underlying non-linear characteristic with various precision. Hence they can be useful in exploring local details in the non-linear characteristics. Another motivation for the identification algorithms presented in the paper is that they can offer faster convergence than traditional orthogonal series algorithms being simultaneously computationally simpler, as it is for instance in the case of the Haar multiscale algorithm [39, 25].

The time-series Hammerstein systems, located in the class of block-oriented non-linear dynamical complexes (see [2, 5-7, 22] for this and other block-oriented structures), are met in various fields such as signal processing, image analysis, industrial engineering, biocybernetics and under parametric assumptions concerning non-linearity were extensively studied in the system identification literature ([12, 27, 42] and the papers cited above). A non-parametric identification algorithm for reconstruction of non-linearities in cascade Hammerstein systems has been first proposed in [16] and the approach has been next developed in [13, 14, 17-20, 31, 32, 38]. In these papers the authors applied conventional kernel or orthogonal series estimates, employing classical orthogonal polynomials, and commonly assumed that the noise disturbing the system is white. In this work, we propose algorithms originated from multiscale orthogonal expansions of functions which may be considered as an extension of the existing orthogonal series expansion methods. We admit in our considerations that the system noise can be correlated in a rather general way and we show that the approach is insensitive to a possible delay in the system dynamics.

As regards the multiresolution and wavelet theory applied in the paper, it has found recently applications in diverse areas as, e.g., signal processing [9, 34, 35, 46], data analysis [37], neural networks [51], approximation theory and statistics [1, 24, 33]. Less attention has been paid to the implementation of this methodology to system identification. We refer the reader to the surveys [28, 41] and the recent papers [39, 25, 45]. The present contribution is an extension of [39, 25]

dealing with the employment of the Haar multiresolution approximation to recovering non-linear characteristics from noisy data towards application of general multiscale approximations with scaling functions of compact support.

The draft of the paper is as follows. In Section 2, the problem under consideration is stated and background assumptions are collected. Section 3 presents some selected facts from the multiresolution theory, basic for derivation and analysis of the identification algorithms. General multiresolution non-parametric identification algorithm for recovering Hammerstein system non-linearity from noisy input-output measurements is developed in Section 4. Pointwise convergence of the algorithm and asymptotic rate of convergence are discussed in Section 5. It is shown that under mild conditions concerning the resolution level the algorithm converges to the unknown non-linearity and attains the best possible non-parametric rate of convergence. This property is insensitive to correlation of the noise i.e. the form of the algorithm, convergence conditions and asymptotic rate of convergence are the same for white and coloured noise. In Section 6, a particular version of the general identification algorithm, obtained for the Haar multiscale approximation, is discussed as an example. Advantages and disadvantages of the proposed identification routine are summarized in Section 7.

2. Statement of the problem

The non-linear time-series system under investigation is shown in Fig. 1. The system is a cascade connection of a non-linear memoryless element, with a characteristic R, and a linear output dynamics. The linear dynamic part is a discrete-time time-invariant and asymptotically stable element operating in steady state, with the unknown impulse response $\{\lambda_p; p = 0, 1, \ldots\}$ and $\sum_{p=0}^{\infty} |\lambda_p| < \infty$. The system inputs $\{x_k; k = \ldots, -1, 0, 1, 2, \ldots\}$ form by assumption a sequence of independent and identically distributed (i.i.d.) random variables with finite variance for which there exists a probability density function f, not assumed known. It is assumed, as usual for the Hammerstein systems, that the internal signal $w_k=R(x_k)$ (output of a static non-linearity), interconnecting both parts of the system, is not accessible for the measurements [2, 6, 22]. However, we can measure the input x_k and the output y_k of the overall system, the latter disturbed by an additive stationary random noise $\{z_k; k = \ldots, -1, 0, 1, 2, \ldots\}$. The following input-output equation holds

$$y_k = \sum_{p=0}^{\infty} \lambda_p R(x_{k-p}) + z_k \tag{2.1}$$

The noise z_k is either

(*a*) *white*, with zero mean, $Ez_k = 0$, and finite variance, $\operatorname{var} z_k < \infty$, or

(*b*) *coloured* - obtained as an output of a discrete-time time-invariant and asymptotically stable linear filter, with the impulse response $\{\omega_p; p = 0, 1, \ldots\}$, which operates in steady state and is driven by a zero mean stationary white noise process $\{\varepsilon_k; k = \ldots, -1, 0, 1, 2, \ldots\}$ with finite variance, i.e. $z_k = \sum_{p=0}^{\infty} \omega_p \varepsilon_{k-p}$, where $E\varepsilon_k = 0$, $\operatorname{var} \varepsilon_k < \infty$ and $\sum_{p=0}^{\infty} |\omega_p| < \infty$.

Processes $\{x_k\}$ and $\{z_k\}$ ($\{x_k\}$ and $\{\varepsilon_k\}$) are by assumption mutually independent. About the unknown non-linearity R and input density f we only assume that they are bounded functions

$$\sup_x |R(x)| = M_R < \infty, \qquad \sup_x f(x) = M_f < \infty \tag{2.2}$$

The goal is to recover the non-linear characteristic R from input-output observations $\{(x_k, y_k)\}$ of the whole system. We observe that inaccessibility of the intermediate signal w_k makes the identification problem of the characteristic R implicit (since not both the input and output signal of the element to be identified can be directly measured) and small *a priori*

information about R requires implementation of a non-parametric method for its recovering [11, 23, 28].

For ease of presentation, we shall assume in the following that $ER(x_k) = 0$. Such a condition is fulfilled, for instance, if the non-linearity R is an odd and the input probability density f is a symmetric (even) function. One can without difficulty generalize the approach to any density and any non-linearity.

3. Multiscale approximation

Let us present some properties of multiscale orthogonal approximations associated with scaling functions of compact support, relevant for further considerations. Detailed treatment of the subject can be found in [10, 8, 35, 50]. Let $\phi(x)$ denote a real function such that translations $\{ \phi(x - n) \}$, $n \in Z$, the set of all integers, form an orthonormal system in $L^2(\mathbb{R})$ and generate a subspace $V_0 = \text{span}\{\phi(x-n), n \in Z\}$ of $L^2(\mathbb{R})$. Consequently, $\phi_{mn}(x) = 2^{m/2} \phi(2^m x - n)$, $n \in Z$, i.e. scaled and translated versions of $\phi(x)$, are orthonormal

$$\int_{-\infty}^{\infty} \varphi_{ml}(x) \varphi_{mn}(x) dx = \begin{cases} 1 & \text{for } l = n \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

and let

$$q_m(x,v) = \sum_{n=-\infty}^{\infty} \varphi_{mn}(x) \varphi_{mn}(v) \tag{3.2}$$

be the summation kernel of the orthonormal family $\{\phi_{mn}(x), n \in Z\}$. Let further $V_m = \text{span}\{\phi_{mn}(x), n \in Z\}$ and suppose that for $m = 0,\pm1,...$ the V_m's form a sequence of nested subspaces:

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots \subset V_m \subset \cdots \subset L^2(\mathbb{R})$$

and that

$$\lim_{m \to -\infty} V_m = \{0\}, \qquad \overline{\lim_{m \to \infty} V_m} = L^2(\mathbb{R})$$

A root function $\phi(x)$ satisfying the above requirements is called a scaling function and the sequence of subspaces $\{ V_m \}_{m \in Z} \subset L^2(\mathbb{R})$ with orthonormal bases $\{\phi_{mn}(x), n \in Z\}$, approximating $L^2(\mathbb{R})$ with increasing precision and called the resolution spaces, constitutes a multiscale (multiresolution) analysis of $L^2(\mathbb{R})$. Due to the above properties of multiresolution analysis, each scaling function $\phi(x)$ satisfies the so-called scaling equation

$$\varphi(x) = \sqrt{2} \sum_{n=-\infty}^{\infty} c_n \varphi(2x - n)$$

for appropriate set of coefficients $\{c_n\}$. A number of practical scaling functions $\phi(x)$ have been proposed in the literature (see [9, 33, 34, 37, 47] and the aforementioned monographs). It is worth noticing that the basis functions $\phi_{mn}(x)$ of the resolution spaces V_m are generated from $\phi(x)$ by only scaling (the scale factor m) and shifting (the translation factor n).

Any function $F(x) \in L^2(\mathbb{R})$ can be approximated in the resolution space V_m (at the resolution level m) as follows

$$F(x;m) = \sum_{n=-\infty}^{\infty} \alpha_{mn} \varphi_{mn}(x) \tag{3.3}$$

where the coefficients α_{mn} are given by

$$\alpha_{mn} = \int_{-\infty}^{\infty} F(v)\,\varphi_{mn}(v)\,dv \tag{3.4}$$

This can be equivalently represented, inserting (3.4) into (3.3) and using the definition in (3.2), in terms of the kernel function

$$F(x;m) = \int_{-\infty}^{\infty} q_m(x,v)\,F(v)\,dv \tag{3.5}$$

The approximation $F(x;m)$ at the resolution level m is the orthogonal projection of a function $F(x)$ onto the resolution space V_m and the sequence $\{F_m(x) = F(x;m)\}_{m\in Z}$ provides the multiscale (multiresolution) approximation of $F(x)$ [29, 30].

In this paper, we shall use a class of scaling functions which are absolutely bounded and supported in a compact set, $[s_1, s_2]$ say (equal zero outside some $[s_1, s_2]$), i.e.

$$|\varphi(x)| \le C, \quad \text{some } C > 0,\ \forall\, x; \qquad \text{supp}\ \varphi(x) = [s_1, s_2] \tag{3.6}$$

This can be jointly expressed as

$$|\varphi(x)| \le C I_{[s_1,s_2]}(x), \quad \text{some } C > 0$$

where $I_A(x)$ denotes the indicator function of A. Examples are the constant Haar scaling function $\phi_H(x)$ with the support [0,1], corresponding to the Haar orthogonal system (see Section 6) and all compactly supported Daubechies scaling functions $\phi_D^s(x)$, $s > 1$, with the supports $[0,2s\text{-}1]$ (for computation of $\phi_D^s(x)$ see [10, Chapter 6]). For each scaling function $\phi(x)$ as in (3.6) there exists an L^1 radial decreasing function $\eta(x)$ such that

$$|\varphi(x)| \le \eta(x)$$

where $\eta(x)$ is such that $\eta(x_1) = \eta(x_2)$ whenever $|x_1| = |x_2|$, $\eta(x_1) \le \eta(x_2)$ when $|x_1| \ge |x_2|$, and $\eta(x) \in L^1(\mathbb{R})$. Candidate functions are $\eta(x)$ of exponential decay, $\eta(x) = Ce^{-a|x|}$ some positive a, algebraic decay, $\eta(x) = C/(1+|x|)^b$ some $b > 1$, or $\eta(x) = CI_{[-c,c]}(x)$ some positive c. Hence, according to Definition 1.4 in [30], $\phi(x)$ is in the class RB of radially bounded scaling functions. The basis functions $\phi_{mn}(x)$ of the resolution spaces V_m associated with scaling functions as in (3.6) satisfy the conditions

$$|\varphi_{mn}(x)| \le C\,2^{m/2}, \qquad \forall\, x, n \tag{3.7a}$$

$$\text{supp}\ \varphi_{mn}(x) = [(s_1+n)/2^m, (s_2+n)/2^m] \tag{3.7b}$$

or in a concise form

$$|\varphi_{mn}(x)| \le C\,2^{m/2}\,I_{[\frac{s_1+n}{2^m}, \frac{s_2+n}{2^m}]}(x) \tag{3.8}$$

and the kernel function (3.2) fulfils then the bound (cf. Theorem 2.4 in [30])

$$|q_m(x,v)| \le C\,2^m H(2^m|x-v|)\,I_{\{v:\,2^m|x-v|\,\le\, s_2-s_1\}}(v) \tag{3.9}$$

some $C > 0$, where $H: [0,\infty) \to R_+$ is a decreasing L^1 function such that $H(0) = 1$. For bounding functions $\eta(x)$ of exponential or algebraic decay we have for example $H(v) = e^{-av/2}$ and $H(v) = 1/(1+v)^b$, $v \ge 0$, respectively (see Theorem 2.5 in [30]).

Since $\phi_{mn}(x)$ are supported in finite intervals $[(s_1+n)/2^m, (s_2+n)/2^m]$, the infinite sum in (3.3) can be truncated for each point x to finite number of terms, yielding

$$F(x;m) = \sum_{n\,=\,n_{\min}(x;m)}^{n_{\max}(x;m)} \alpha_{mn}\,\varphi_{mn}(x) \tag{3.10}$$

where the summation limits are

$$n_{\min}(x;m) = [2^m x - s_2] + 1 \quad \text{and} \quad n_{\max}(x;m) = [2^m x - s_1] \tag{3.11}$$

and where $[v]$ stands for the integer part of v, with the following modification of the weighting coefficients

$$\alpha_{mn} = \int_{(s_1+n)/2^m}^{(s_2+n)/2^m} F(x)\,\varphi_{mn}(x)\,dx \tag{3.12}$$

Respectively, taking account of (3.5) and (3.9), we obtain

$$F(x;m) = \int_{x-(s_2-s_1)/2^m}^{x+(s_2-s_1)/2^m} q_m(x,v)\,F(v)\,dv \tag{3.13}$$

and the kernel function in (3.2) takes for each point x the form

$$q_m(x,v) = \sum_{n=n_{\min}(x;m)}^{n_{\max}(x;m)} \varphi_{mn}(x)\,\varphi_{mn}(v) \tag{3.14}$$

In the sum in (3.10) the number of terms, for every scale m and every point x, does not exceed the number $S = [s_2 - s_1] + 1$ and all the appearing basis functions $\phi_{mn}(x)$ are non-zero only in a finite interval

$$\bigcup_{n=n_{\min}(x;m)}^{n_{\max}(x;m)} \operatorname{supp} \varphi_{mn}(v) \subseteq [x-(s_2-s_1)/2^m,\, x+(s_2-s_1)/2^m] \tag{3.15}$$

Consequently (see (3.9), (3.13), (3.14))

$$\operatorname{supp}\, q_m(x,v) \subseteq [x-(s_2-s_1)/2^m,\, x+(s_2-s_1)/2^m], \quad \forall\, x \tag{3.16}$$

For the scaling function $\phi(x)$ fulfilling the condition in (3.6) (and hence belonging to the class *RB*) the associated multiscale approximation $\{F_m(x) = F(x;m)\}_{m\in Z}$ converges to $F(x)$ pointwise almost everywhere in the Lebesgue measure sense, i.e. at all points x except sets of zero Lebesgue measure (Theorem 2.1(*i*) in [30]). For such $\phi(x)$ we thus have in particular the convergence

$$F(x;m) \to F(x) \quad \text{as} \quad m \to \infty \tag{3.17}$$

at each point x where $F(x)$ is a continuous function (for short: at $x \in \text{Cont}(F)$ - the set of all continuity points of F). From the convergence (3.17), substituting $F(v) = I_{[x-a,x+a]}(v)$ some $a > (s_2\text{-}s_1)/2^m$ into (3.13), one can immediately infer that

$$\int_{x-(s_2-s_1)/2^m}^{x+(s_2-s_1)/2^m} q_m(x,v)\,dv \to 1 \quad \text{as} \quad m \to \infty \tag{3.18}$$

Special versions of the scaling functions as in (3.6), called coiflets, possess vanishing moments, i.e.

$$\int_{s_1}^{s_2} x^k \varphi(x)\,dx = 0, \quad k = 1,2,\ldots,r \tag{3.19}$$

for some $r \geq 1$, and $r+1$ is then called the order of the coiflet. For construction of coiflets see [10, Chapter 8] and [35, Chapter 7].

4. Identification algorithm

The fundamental observation for derivation of our identification algorithms is that under

assumptions of Section 2, for white as well as coloured noise, it holds

$$E\{y_k \mid x_{k-d} = x\} = cR(x), \tag{4.1}$$

where c (= λ_d) is a constant (see (2.1) in Section 2). This reveals that the system non-linearity R (up to a scale constant c) can be estimated as a regression function of the system output y_k on the system input x_{k-d} [16]. In equation (4.1) we only require that $\lambda_d \neq 0$ for some $d \geq 0$. Since it is particularly permitted that $\lambda_0 = \lambda_1 = ... = \lambda_{d-1} = 0$ hence a d-step delay is admissible in the linear system dynamics. Let us now observe that for x such that $f(x) > 0$ the regression in (4.1) can be decomposed as follows [13, 15, 38]

$$cR(x) = g(x)/f(x), \tag{4.2}$$

where $g(x) = E\{y_k | x_{k-d} = x\} f(x)$ and $f(x)$ is the probability density function of the input signal. Owing to (2.2) we notice that $\int_{-\infty}^{+\infty} f^2(x)\,dx < \infty$ and $\int_{-\infty}^{+\infty} g^2(x)\,dx < \infty$. Due to this property the numerator g and denominator f of the ratio in (4.2) may be approximated with an arbitrary precision using multiscale orthogonal approximations of Section 3. Applying a multiscale approximation associated with scaling function $\phi(x)$ with the compact support $[s_1, s_2]$ (see (3.6)), we obtain at the resolution level m the following approximants ((3.10))

$$g(x;m) = \sum_{n = n_{\min}(x;m)}^{n_{\max}(x;m)} a_{mn}\varphi_{mn}(x) \quad \text{and} \quad f(x;m) = \sum_{n = n_{\min}(x;m)}^{n_{\max}(x;m)} b_{mn}\varphi_{mn}(x) \tag{4.3}$$

In the above sums $n_{\min}(x;m)$ and $n_{\max}(x;m)$ are as in (3.11) and the summation coefficients a_{mn} and b_{mn} (cf. (3.12)) are simple expectations

$$a_{mn} = \int_{(s_1+n)/2^m}^{(s_2+n)/2^m} g(x)\varphi_{mn}(x)\,dx = E\{y_d\,\varphi_{mn}(x_0)\} \tag{4.4a}$$

$$b_{mn} = \int_{(s_1+n)/2^m}^{(s_2+n)/2^m} f(x)\varphi_{mn}(x)\,dx = E\{\varphi_{mn}(x_0)\} \tag{4.4b}$$

The latter fact can be easily deduced from (3.4), (3.7b), (3.12), definition of g and f and stationarity of the processes $\{x_k\}$ and $\{y_k\}$ (Section 2). Based on the factorization in (4.2) and (4.3) along with (4.4), we can propose the following natural fractional-form estimate of $cR(x)$ at the resolution level m (in the resolution space V_m):

$$R_N(x;m) = g_N(x;m)/f_N(x;m) = \sum_{n = n_{\min}(x;m)}^{n_{\max}(x;m)} a_{mn,N}\varphi_{mn}(x) \Big/ \sum_{n = n_{\min}(x;m)}^{n_{\max}(x;m)} b_{mn,N}\varphi_{mn}(x) \tag{4.5}$$

where $a_{mn,N}$ and $b_{mn,N}$ are estimates of a_{mn}'s and b_{mn}'s computed from N (random) observations of the whole system input and output $\{(x_k, y_{k+d}); k = 1, 2, ..., N\}$ as sample means:

$$a_{mn,N} = \frac{1}{N}\sum_{k=1}^{N} y_{k+d}\,\varphi_{mn}(x_k)\ ; \qquad b_{mn,N} = \frac{1}{N}\sum_{k=1}^{N} \varphi_{mn}(x_k) \tag{4.6}$$

The empirical coefficients $a_{mn,N}$ and $b_{mn,N}$ in the ratio estimate (4.5) may be simplified to the form

$$a_{mn,N} = \sum_{\{k:\, x_k \in [\frac{s_1+n}{2^m}, \frac{s_2+n}{2^m}]\}} y_{k+d}\,\varphi_{mn}(x_k)\,; \qquad b_{mn,N} = \sum_{\{k:\, x_k \in [\frac{s_1+n}{2^m}, \frac{s_2+n}{2^m}]\}} \varphi_{mn}(x_k) \tag{4.7a}$$

or further

$$a_{mn,N} = \sum_{\{k:\, u_k \in [s_1, s_2]\}} y_{k+d}\,\varphi(u_k)\,; \qquad b_{mn,N} = \sum_{\{k:\, u_k \in [s_1, s_2]\}} \varphi(u_k) \tag{4.7b}$$

where $u_k = 2^m x_k - n$ after omitting, respectively, the common factors $1/N$ and $2^{m/2}$.

Remark 4.1: Due to compactness of the support of $\phi(x)$ there is a finite number of terms in the sums in (4.5). For each point x and each scale m the numerator and denominator of (4.5) contains at most S components (cf. Section 3). If $cR(x)$ is estimated for x in a certain interval $[x_{\min},x_{\max}]$ (which is a typical situation) then at the resolution level m the translation factor n varies in the estimate over the range of integers $[2^m x_{\min} - s_2] + 1 \le n \le [2^m x_{\max} - s_1]$ and the total number of coefficients $(a_{mn,N},b_{mn,N})$ needed by $R_N(x;m)$ does not exceed $[2^m x_{\max} - s_1] - [2^m x_{\min} - s_2]$. The resolution level m can be various for various subintervals $[x_{\min},x_{\max}]$.

Remark 4.2: Substituting (4.6) into (4.5) and invoking the definition in (3.14) of the kernel $q_m(x,v)$, we get equivalently

$$R_N(x;m) = \frac{g_N(x;m)}{f_N(x;m)} = \frac{\frac{1}{N}\sum_{k=1}^{N} q_m(x,x_k)\, y_{k+d}}{\frac{1}{N}\sum_{k=1}^{N} q_m(x,x_k)} \tag{4.8}$$

This form of the estimate recalls the traditional Nadaraya-Watson kernel regression estimator

$$\hat{R}(x;h) = \frac{\frac{1}{N}\sum_{k=1}^{N} K_h(x-x_k)\, y_{k+d}}{\frac{1}{N}\sum_{k=1}^{N} K_h(x-x_k)}$$

originally with $d = 0$, where $K_h(u) = h^{-1}K(u/h)$, K is a kernel function and h is a bandwidth parameter (see, e.g., the source papers [36, 49] and monographs [4, 48, 23]), studied for the Hammerstein systems in [16-18, 20], among others. However, our estimate is quite different from the computational viewpoint. In contrast to the kernel algorithms, for the recalculation (at each individual estimation point x) the estimate does not need the whole set of measurements $\{(x_k,y_{k+d}); k = 1, 2, \dots, N\}$ but exploits instead only the set of coefficients $\{(a_{mn,N},b_{mn,N})\}$. For a given resolution level m the coefficients $a_{mn,N}$ and $b_{mn,N}$ are computed from the measurements only ones and are smaller in the number than the number of raw measurement data, i.e. occupy smaller amount of computer memory (Remark 4.1 and condition in (5.1) in Section 5).

Remark 4.3: In the sample means in (4.6) the measurements x_k are independent and identically distributed random variables (see assumptions in Section 2) whereas y_k are dependent quantities as the outputs of a non-linear dynamical system (they form an infinite non-linear moving average process - see (2.1)). This locates the problem of the identification of $cR(x)$ beyond the standard tasks of non-parametric estimation of a regression function from independent input-output measurements [11, 23]. The problem of estimation of a regression from dependent data has been studied in the statistical literature by many authors and mainly the so-called mixing conditions for the data have been there assumed as a model of dependence (see [21] and the references therein). In general this model is however different from ours (cf. the discussion in Section II in [14]).

Remark 4.4: Using the algorithm in (4.5)-(4.6) we can only estimate the system non-linearity with accuracy to the scaling constant c, which depends on the linear dynamic part of the system ($c = \lambda_d$). This limitation is an unavoidable consequence of the cascade complex structure of the system and assembling nature of the data $\{(x_k,y_{k+d})\}$ used for identification, independent of the applied identification algorithm (e.g. [13, 14, 20]).

We note that the form of the identification algorithm in (4.5)-(4.6) is the same for white and

coloured noise.

5. Convergence and rate of convergence

The choice of the parameter m in (4.5)-(4.6) plays the crucial role for the asymptotic behaviour of the estimate. Too small values of m result in excessive bias whereas too large values of m cause the increase of variance. If the scale factor (resolution level) m is appropriately fitted to the number N of measurement data $\{(x_k, y_{k+d})\}_{k=1}^{N}$, i.e. treated as a function of the sample size $m = m(N)$ and selected in such a way that

$$m(N) \to \infty, \qquad 2^{m(N)}/N \to 0 \qquad \text{as} \qquad N \to \infty \tag{5.1}$$

then estimate in (4.5)-(4.6) converges to the unknown characteristic as $N \to \infty$. The following theorem holds.

Theorem 1: Let the assumptions of Section 2 be in force and let $ER(x) = 0$. Let the scaling function $\phi(x)$ of a multiscale approximation satisfy conditions (3.6) in Section 3 (i.e. (3.7a) and (3.7b) hold for the multiresolution basis functions $\{\phi_{mn}(x)\}$) and let the scale parameter $m = m(N)$ fulfil conditions in (5.1). Then for white as well as coloured noise with general correlation properties (cf. Section 2)

$$R_N(x_0; m) \to c\,R(x_0) \qquad \textit{in probability} \qquad \text{as} \qquad N \to \infty \tag{5.2}$$

at every point $x_0 \in \mathrm{Cont}(R,f)$ - the set of continuity points of R and f, where $f(x_0) > 0$.

Proof: See Appendix A.

The above result shows that under mild conditions on the system dynamics and system noise, specified in Section 2, the identification algorithm of Section 4 is convergent for a large class of non-linear characteristics and input densities. Observe that if the input observations $x_k \in [x_{\min}, x_{\max}]$ and accordingly $cR(x)$ is to be estimated for x in the interval $[x_{\min}, x_{\max}]$ (Remark 4.1), in Theorem 1 virtually no conditions are imposed on R and f. The estimate converges, in a pointwise manner, under relatively weak boundedness requirements (2.2) and convergence takes place at the points where R and f are both continuous functions and the input density f does not vanish. Such a range of convergence is practically not worse than the range of convergence in probability (at almost all points) of the conventional orthogonal series ratio estimates, employing classical trigonometric or Hermite orthogonal polynomials [15, 13, 31], however our algorithm can offer faster rate of convergence, which will be shown below. Additional assumption that $ER(x) = 0$ is not essential for convergence. If it does not hold, the estimate converges to scaled and shifted non-linearity $cR(x)+e$, where $e = ER(x)\sum_{p \neq d}\lambda_p$ ((2.1), (4.1), (4.2)). In the light of the above, the only important demand for convergence of the estimate $R_N(x; m)$ is fulfilment of the condition in (5.1) imposed on the resolution level m, i.e. proper selection of the strategy $m = m(N)$.

Remark 5.1: Denoting $1/2^{m(N)} = h_N$ we observe that (5.1) can be restated as follows

$$h_N \to 0, \qquad N h_N \to \infty \qquad \text{as} \qquad N \to \infty$$

which coincides with the standard demand imposed on the bandwidth parameter $h = h_N$ for ensuring pointwise convergence in probability of kernel regression estimates (see e.g. [23, Proposition 3.1.1, p. 29] and [16, 18, 20] for Hammerstein systems). Observe that the resolution $1/2^m$ can be interpreted as the bandwidth factor of the summation kernel $q_m(x,v)$ in the alternative representation of the estimate $R_N(x; m)$ in (4.8) (cf. (3.16)).

Assuming some local smoothness of R and f around x_0, we can prove the following result establishing the asymptotic rate of convergence in (5.2) and yielding some recommendations for the optimal choice of the resolution level $m = m(N)$.

Theorem 2: Let ϕ be a scaling function with compact support, such as in (3.6), and let all the remaining assumptions of Theorem 1 hold. Let $f(x_0) > 0$ and let moreover

A) $R \in Lip(x_0;\alpha)$, $f \in Lip(x_0;\beta)$, $\alpha,\beta \in (0,1]$, i.e. $|R(x) - R(x_0)| \le L_R|x - x_0|^{\alpha}$, $|f(x) - f(x_0)| \le L_f|x - x_0|^{\beta}$ for some positive L_R, L_f in the neighbourhood of x_0.

B1) $R \in C^p(x_0)$, $f \in C^q(x_0)$, $p,q \ge 1$.

B2) $R \in C^p(x_0)$, $f \in C^q(x_0)$, $p,q \ge 1$, and ϕ be a coiflet of the order r+1 i.e. a scaling function as in (3.6) possessing in addition r vanishing moments (see (3.19)).

Then asymptotically, without any distinction for white and coloured noise, it holds

$$| R_N(x_0;m) - cR(x_0) | = O\left(N^{-\delta/(2\delta+1)}\right) \quad in\ probability \tag{5.3}$$

where

$$\delta = \begin{cases} \min(\alpha,\beta) & for\ the\ case\ A \\ 1 & for\ the\ case\ B1 \\ \min(p,q,r+1) & for\ the\ case\ B2 \end{cases}$$

provided that the resolution level m is selected as

$$m(N) = \left\lfloor \frac{1}{2\delta+1} \log_2 N \right\rfloor \tag{5.4}$$

Proof: See Appendix B.

Remark 5.2: In the above theorem, for a sequence of random variables $\{\zeta_N\}$ the statement $\zeta_N = O(c_N)$ *in probability* means that $d_N\zeta_N / c_N \to 0$ *in probability* as $N \to \infty$ for any number sequence $\{d_N\}$ convergent to zero (see, e.g., [20], p. 140).

All rates given above are in an asymptotic sense, i.e. take place for large numbers N of data. As it follows from the proof in Appendix B, the local choice (5.4) of the resolution level m is asymptotically best possible and thus the rate in (5.3) is asymptotically optimal. This rate depends, through the index δ, on local smoothness of more crude function from among $R(x)$ and $f(x)$ (i.e. a function with smaller Lipschitz exponent or smaller number of bounded derivatives: case *A* and *B2*) and is sensitive to the smoothness of the scaling function $\phi(x)$ (case *B1* and *B2*). It does not however depend on the particular linear dynamics in the Hammerstein system. Increase of smoothness of R and f and adequate choice of ϕ, matching the smoothness of the functions R and f, result in improvement of the convergence speed. Since however $\delta/(2\delta+1) < 1/2$ for each index δ, the rate $O(N^{-\delta/(2\delta+1)})$ guaranteed in (5.3) is always smaller than $O(N^{-1/2})$, i.e. the best possible parametric rate of convergence in probability [3]. The deterioration of convergence speed in non-parametric inference, where uncertainty in the identified characteristics is incomparably greater than in the standard parametric estimation, is a typical occurrence and achievable rate of convergence in probability is generally of smaller order $O(N^{-r})$, $0 < r < 1/2$ ([23]). Nevertheless, the following expedient properties of our non-parametric identification algorithm can be observed from (5.3).

Remark 5.3: In the case *A*, for $\alpha = \beta = 1$ ($\delta = 1$) and the optimal resolution level selection law

$m(N) = \lceil (1/3)\log_2 N \rceil$ we attain the rate $O(N^{-1/3})$ in probability which agrees with the best possible non-parametric rate of convergence for the class of Lipschitz functions and independent input-output data [44]. If however the scaling function ϕ of a given multiscale approximation is characterized only by the basic requirement in (3.6) the guaranteed convergence rate in (5.3) is not improved for smoother differentiable functions R and f (case *B1*).

Remark 5.4: In the case *B2*, for $q,r \geq p$ ($\delta = p$) and the resolution level choice $m(N) = \lceil (1/(2p+1))\log_2 N \rceil$ we obtain the rate $O(N^{-p/(2p+1)})$ i.e. the algorithm achieves the best possible non-parametric rate of convergence in probability for differentiable characteristics derived in [44] in the context of independent data (memoryless system). This rate approaches the optimal parametric rate $O(N^{-1/2})$ as $p \to \infty$. Using instead the conventional trigonometric or Hermite orthogonal series algorithms, for $q = p$ we obtain worse rate of convergence $O(N^{-(2p-1)/4p})$ in probability [13] (see also Theorem 3 in [31]).

Consequently, the multiscale identification algorithm in (4.5)-(4.6) can have a better rate of convergence than traditional orthogonal series algorithms worked out earlier. For merely Lipschitz functions R and f around x_0 the attainable rate $O(N^{-1/3})$ is faster than $O(N^{-1/4})$ in probability guaranteed for $R,f \in C^1(x_0)$ by the trigonometric or Hermite orthogonal series routines. The rate $O(N^{-1/4})$ is in turn achieved by the multiscale algorithm for $R,f \in Lip(x_0;1/2)$. For $R,f \in C^2(x_0)$ and the scaling function with $r = 1$ vanishing moments we obtain the rate $O(N^{-2/5})$ which is not much worse than $O(N^{-1/2})$, compared with $O(N^{-3/8})$ for the trigonometric or Hermite orthogonal series estimates. In the particular case of locally constant functions R and f around x_0, which may be considered as the fulfilment of the conditions as in the case A of Theorem 2 with arbitrarily large exponents α and β for x in some interval (x_0-a,x_0+a), $0<a<1$, the estimate attains the rate $O(N^{-1/2})$ in probability. Such covergence rate is achieved for each scale factor m for which $a > (s_2-s_1)/2^m$, where $[s_1,s_2]$ is the support of ϕ (cf. (5.3), (3.16) and proof for the case A in Appendix B). It is worth emphasizing that the estimate converges under the same conditions and all the above rates remain the same for white and correlated noise of a rather general correlation structure (Section 2).

6. Example

For specific multiscale bases $\{\phi_{mn}(x)\}$, we obtain particular versions of the algorithm in (4.5)-(4.6) and theorems of Section 5. Here we shall give an example obtained for a popular choice of the scaling function corresponding to the Haar multiscale basis $\{\phi_{H,mn}(x),\ n \in Z\}$, considered in more detail in [39, 25]. The Haar scaling function has the form

$$\varphi_H(x) = I_{[0,1)}(x), \qquad x \in (-\infty,\infty)$$

i.e. is supported in $[s_1,s_2] = [0,1]$, and the Haar resolution spaces V_m are spanned, for each resolution level $m \in Z$, by the functions

$$\varphi_{H,mn}(x) = 2^{m/2}\,\varphi_H(2^m(x - \frac{n}{2^m})) = 2^{m/2}\,I_{[\frac{n}{2^m},\frac{n+1}{2^m})}(x), \qquad n \in Z \tag{6.1}$$

supported in $[n/2^m, (n+1)/2^m]$. Hence the subspace V_m at the resolution level m consists of all functions being piecewise constant on all intervals $\{[n/2^m, (n+1)/2^m), n \in Z\}$. For the Haar basis $\{\phi_{H,mn}(x), n \in Z\}$ the correspondent summation kernel has the form (cf. (3.11), (3.14))

$$q_m(x,v) = 2^m I_{[\frac{[2^m x]}{2^m},\frac{[2^m x]+1}{2^m})}(v)$$

and this kernel satisfies condition (3.9) for $H(v) = I_{[0,1)}(v)$ since at present we have $s_2 - s_1 = 1$ and $2^m x - 1 \le [2^m x] \le 2^m x$.

Application of the Haar basis functions $\{\phi_{H,mn}(x)\}$ in the algorithm (4.5)-(4.6) yields at the resolution level m the following estimate

$$R_{H,N}(x;m) = a_{H,mn,N} / b_{H,mn,N} \tag{6.2}$$

for each point $x \in [n/2^m, (n+1)/2^m)$, where ((4.7a))

$$a_{H,mn,N} = \sum_{\{k:\, x_k \in [\frac{n}{2^m}, \frac{n+1}{2^m})\}} y_{k+d}\ ; \qquad b_{H,mn,N} = \#\{ x_k \in [\frac{n}{2^m}, \frac{n+1}{2^m}) \}$$

and where # denotes the cardinality of a collection. This is a straightforward consequence of the fact that two Haar basis functions as in (6.1) of the same scale m do not overlap and $n_{\min}(x;m) = n_{\max}(x;m) = [2^m x]$. Since in the ratio in (6.2) the denominator counts the number of measurements x_k in the interval $[n/2^m, (n+1)/2^m)$ and the numerator selects and sums up the corresponding output measurements y_{k+d}, including a possible d-step delay in the system, we obtain at present the sample mean of shifted output measurements as the estimate of $cR(x)$ at x. This sample mean is computed for x_k from the interval containing the point x at which the estimation is carried out, of the length $1/2^m$ for the resolution level m.

Obviously, the Haar scaling function $\phi_H(x)$ fulfils condition (3.6) of Section 3. This observation along with Theorems 1 and 2 yields the following (compare Theorem 1 and 2 in [39, 25])

Corollary: If the assumptions of Theorem 1 hold and in particular $R \in Lip(x_0;1)$, $f \in Lip(x_0;1)$ and the resolution level $m(N)$ is selected according to the rule $m(N) = \lceil (1/3)\log_2 N \rceil$ then for both white and coloured noise

$$| R_{H,N}(x_0;m) - cR(x_0) | = O\left(N^{-1/3}\right) \qquad \textit{in probability}$$

provided that $f(x_0) > 0$, i.e. the Haar estimate $R_{H,N}(x_0; m)$ is convergent to $cR(x_0)$ and attains asymptotically the optimal non-parametric rate of convergence for Lipschitz functions (Remark 5.3).

As the scaling function $\phi_H(x)$ does not possess vanishing moments (i.e. $r = 0$; cf. (3.19) in Section 3), the convergence rate $O(N^{-1/3})$ of the Haar estimate does not increase in the case of smoother differentiable functions $R \in C^p(x_0)$, $f \in C^q(x_0)$ with arbitrary $p,q \ge 1$ (Theorem 2 and Remark 5.3).

To illustrate the behaviour of estimate in (6.2) for finite number of measurements a numerical experiment has been performed in which the exciting input signal $\{x_k; k = ..., -1, 0, 1, 2, ...\}$ and the white noise process $\{\varepsilon_k; k = ..., -1, 0, 1, 2, ...\}$ (see Fig. 1) were distributed uniformly, over the intervals [-0.5,0.5] and [-l,l] respectively. Then $E\varepsilon_k = 0$, see Section 2, and the noise dispersion $\sigma_\varepsilon = l/\sqrt{3}$. In the example, a quantizer function $R(x) = 0.2\lceil 10(x+0.05)\rceil$, $x \in [-0.5, 0.5]$ has been selected as system non-linearity while *FIR*(4) element with one-step delay $v_k = 0.5w_{k-1} + 0.25w_{k-2} + 0.125w_{k-3} + 0.0625w_{k-4}$ was chosen as system dynamics (i.e. $\lambda_0 = 0$, $d = 1$ and scaling constant $c = \lambda_1 = 0.5$ in our example). The correlated output noise $\{z_k; k = ..., -1, 0, 1, 2, ...\}$ was generated from the white noise $\{\varepsilon_k\}$ as *MA*(1) or *MA*(3) process, using the filters $z_k = \varepsilon_k + 0.6\varepsilon_{k-1}$ and $z_k = \varepsilon_k + 0.6\varepsilon_{k-1} + 0.4\varepsilon_{k-2} + 0.2\varepsilon_{k-3}$, respectively. The level of the noise (white or correlated) was selected, by changing l, so as to give the noise dispersion value (σ_ϵ or σ_z) of 5% of the magnitude, $\max|v_k|$, of the noiseless output signal $|v_k|$. The length of data sequence was chosen equal to $N = 50, 100, 150, ..., 500$. As the estimate accuracy index, the mean integrated squared error $MISE(m;N) = E\int_{-0.5}^{0.5}\left[cR(x) - R_{H,N}(x;m)\right]^2 dx$ has been calculated

numerically, yielding as well an average pointwise estimation error in the interval [-0.5,0.5] of the unit length. For each number N of data the resolution level m for the estimate has been selected according to the law $m = m(N) = \lceil C\,(1/3)\log_2 N \rceil$ where the constant $C = 1.5$, minimizing the estimation error, has been established experimentally. The run of the *MISE* error versus the number N of data is shown in Fig. 2. From the plots we see that estimation error quickly decreases with growing N and that the error does not visibly depend on correlation of the noise. For the white noise, *MA*(1) noise and *MA*(3) noise the corresponding plots differ by a small percentage only, so that no essential difference in the estimate accuracy can be observed for $N > 250$.

7. Conclusions

In this paper, we have studied a class of non-parametric identification algorithms for reconstruction of non-linearities in the discrete-time Hammerstein systems. The algorithms are based on the theory of multiscale orthogonal expansions associated with scaling functions of compact support. They can be used when *a priori* knowledge about the system is very small, and in particular no parametric representation of unknown non-linear characteristic is given.

Advantages of the algorithms can be summarized as follows:

1. The algorithms are relatively simple, converge under weak requirements for a wide class of non-linear characteristics and offer faster rate of convergence than traditional orthogonal series identification routines elaborated earlier ([15, 13, 31]).

2. The form of the algorithms, convergence conditions and asymptotic rate of convergence are the same for white and correlated noise with general correlation properties. This is in contrast to the traditional parametric approach where correlation of the noise requires, even in the case of linear systems, a deep revision of identification routines and construction of general identification algorithms, effective for a large class of noise models, is still the actual problem [52, 43].

3. The proposed algorithms are convenient for computer implementations. They need only elementary computations and occupy moderate amount of computer memory, much smaller than conventional kernel non-parametric identification algorithms (Remark 4.2). For the use of the algorithms, only a set of coefficients $(a_{mn,N}, b_{mn,N})$ must be calculated from experimental data and in the case of a delay in the system merely appropriately shifted output measurements have to be used to compute the coefficients. In order to memorize the estimate at the resolution level m just a number of $O(2^m)$ of pairs $(a_{mn,N}, b_{mn,N})$ is sufficient to be stored in a computer in practical cases (Remark 4.1). We do not need in particular to keep in a memory the whole set of N input-output measurements $\{(x_k, y_{k+d}); k = 1, 2, \ldots, N\}$ which is the case for non-parametric kernel estimates. This yields savings in computer load which rapidly grow with growing N as we require $2^m/N \to 0$ as $N \to \infty$ (condition (5.1) in Section 5).

As disadvantages of the approach one can recognize the following:

1. We can estimate non-linearity R only up to a constant c. However, if we possess a bit more information about R than in Section 2 and we know in particular its value in only one point, x_0 say, such that $R(x_0) \neq 0$ then we can estimate c as $R_N(x_0;m)/R(x_0)$, using our identification algorithm, and next $(R(x_0)/R_N(x_0;m))R_N(x;m)$ may be used for estimation of $R(x)$, provided that $R_N(x_0;m) \neq 0$.

2. For the multiscale orthogonal bases $\{\phi_{mn}(x)\}$ of compact support, except the Haar basis, the underlying scaling function $\phi(x)$ is not given in the explicit form but its values must be computed numerically for each point x. This is in reality not a serious drawback as we may employ to this

end ready for use computation procedures described in the literature [9, 35, 37].

3. The algorithm performs merely pointwise estimation of non-linearity, i.e. the estimate $R_N(x;m)$ must be recomputed for each point x separately. This disadvantage is a common feature of all non-parametric methods which provide in fact only a graph of a non-linear characteristic rather than a closed-form representation of non-linearity. This is a natural consequence of poor prior knowledge [23].

To sum up, the proposed non-parametric algorithms may be used for estimation of non-linearities when traditional parametric methods fail, i.e. when *a priori* information about non-linearity is small and in the presence of a disturbance whose correlation is not explicitly modeled. An extension of the approach towards more general composite systems and application of more general wavelet orthogonal bases has been recently discussed in [26, 40].

Acknowledgements

The author thanks the reviewers for their helpful comments and M.Sc. P. Śliwiński for his assistance in preparing the numerical example.

Appendix A. Proof of Theorem 1

I. Bias Error (limit properties): For a random sample $\{(X_k\,,\,Y_{k+d})\}_{k=1}^{N}$ (further on we denote random variables by capitals, for clarity) and $x = x_0$, a fixed point, we have (cf. (4.5))

$$g_N(x_0;m) = \sum_{n=n_{\min}(x_0;m)}^{n_{\max}(x_0;m)} a_{mn,N}\,\varphi_{mn}(x_0) \quad \text{and} \quad f_N(x_0;m) = \sum_{n=n_{\min}(x_0;m)}^{n_{\max}(x_0;m)} b_{mn,N}\,\varphi_{mn}(x_0) \qquad \text{(A.1)}$$

where (cf. (4.6))

$$a_{mn,N} = \frac{1}{N}\sum_{k=1}^{N} Y_{k+d}\,\varphi_{mn}(X_k)\;; \qquad b_{mn,N} = \frac{1}{N}\sum_{k=1}^{N} \varphi_{mn}(X_k) \qquad \text{(A.2)}$$

and $g_N(x_0\,;m)$ and $f_N(x_0\,;m)$ play the role of estimates of $g(x_0) = cR(x_0)f(x_0)$ and $f(x_0)$, assuming that $ER(X_k) = 0$ ((4.2)). Since (see definitions in (4.4a)-(4.4b))

$$E\{a_{mn,N}\} = a_{mn}\,; \qquad E\{b_{mn,N}\} = b_{mn}$$

we note that for white as well as coloured noise, and for each N, it holds ((4.3))

$$E\{g_N(x_0;m)\} = \sum_{n=n_{\min}(x_0;m)}^{n_{\max}(x_0;m)} a_{mn}\,\varphi_{mn}(x_0) = g(x_0;m) \qquad \text{(A.3)}$$

$$E\{f_N(x_0;m)\} = \sum_{n=n_{\min}(x_0;m)}^{n_{\max}(x_0;m)} b_{mn}\,\varphi_{mn}(x_0) = f(x_0;m) \qquad \text{(A.4)}$$

i.e. the above expectations are approximations, at $x = x_0$, of $g(x)$ and $f(x)$ in the resolution space V_m. Hence, if $x_0 \in \mathrm{Cont}(R,f) = \mathrm{Cont}(g)$ - the set of continuity points of R and f, we have the convergence ((3.17) in Section 3)

$$\text{bias}\{g_N(x_0;m)\} = E\{g_N(x_0;m)\} - g(x_0) \to 0 \quad \text{as} \quad m \to \infty \qquad \text{(A.5)}$$

$$\text{bias}\{f_N(x_0;m)\} = E\{f_N(x_0;m)\} - f(x_0) \to 0 \quad \text{as} \quad m \to \infty \qquad \text{(A.6)}$$

for each N and for both white and coloured noise.

II. Variance Error: Taking account of (4.8), we have equivalently

$$\begin{aligned} g_N(x_0;m) &= \frac{1}{N}\sum_{k=1}^{N} q_m(x_0,X_k)\,Y_{k+d} \\ f_N(x_0;m) &= \frac{1}{N}\sum_{k=1}^{N} q_m(x_0,X_k) \end{aligned} \tag{A.7}$$

where $q_m(x_0,x)$ is the summation kernel in (3.14) associated with the system $\{\phi_{mn}(x),\ n \in \mathrm{Z}\}$. Noticing that (cf. (3.9), (3.16))

$$\int_{-\infty}^{\infty} q_m^2(x_0,x)\,dx \;\le\; 2C^2(s_2-s_1)2^m \;\triangleq\; \bar{C}2^m$$

and including (2.2), we get as a consequence

$$E\,q_m^2(x_0,X_1) = \int_{-\infty}^{\infty} q_m^2(x_0,x)\,f(x)\,dx \;\le\; M_f\bar{C}\,2^m$$

$$E\left[R^2(X_1)\,q_m^2(x_0,X_1)\right] = \int_{-\infty}^{\infty} q_m^2(x_0,x)\,R^2(x)\,f(x)\,dx \;\le\; M_R^2 M_f\bar{C}\,2^m$$

Now, exploiting the above bounds and repeating the steps as in Appendix I in [25], one can ascertain that for asymptotically stable dynamics/noise filter in the Hammerstein system and for large values of m the following bound holds

$$\mathrm{var}\,\{\,g_N(x_0;m)\,\} = O\left(2^m/N\right) \tag{A.8}$$

for each point x_0 and for both white and coloured noise as in Section 2 (the rather vast but straightforward proof of this fact is here omitted for shortness; necessary guidelines for its derivation can be found in [25]). Similarly, for each x_0, we obtain

$$\mathrm{var}\,\{\,f_N(x_0;m)\,\} = O\left(2^m/N\right) \tag{A.9}$$

Convergence in (A.5) and (A.6) and variance bounds in (A.8) and (A.9) yield together the mean square convergence of $g_N(x_0;m)$ and $f_N(x_0;m)$ to $g(x_0) = cR(x_0)f(x_0)$ and $f(x_0)$ as $N \to \infty$, provided that $x_0 \in \mathrm{Cont}(R,f)$ and conditions in (5.1) on the resolution level m are satisfied. If $f(x_0) > 0$, this yields convergence in probability of the ratio estimate $R_N(x_0;m) = g_N(x_0;m)/f_N(x_0;m)$ to $cR(x_0)$ as $N \to \infty$. Such convergence takes place equally for white and coloured noise. □

Appendix B. Proof of Theorem 2

Let $x_0 \in \mathrm{Cont}(R,f) = \mathrm{Cont}(g)$. Owing to (A.3) and (3.13), one can easily recognize that

$$E\,\{\,g_N(x_0;m)\,\} = \int_{x_0-(s_2-s_1)/2^m}^{x_0+(s_2-s_1)/2^m} q_m(x_0,x)\,g(x)\,dx \tag{B.1}$$

and hence

$$\mathrm{bias}\,\{\,g_N(x_0;m)\,\} = \int_{x_0-(s_2-s_1)/2^m}^{x_0+(s_2-s_1)/2^m} q_m(x_0,x)\,g(x)\,dx \;-\; g(x_0)$$

Due to convergence in (3.18) and the inequality $\big|\,|a| - |b|\,\big| \le |a-b|$ we ascertain that vanishing of the bias error in (A.5), $|\mathrm{bias}\,\{\,g_N(x_0;m)\,\}| \to 0$ as $m \to \infty$, is not slower than

convergence of the expression

$$|G(x_0;m)| \triangleq \left| \int_{x_0-(s_2-s_1)/2^m}^{x_0+(s_2-s_1)/2^m} q_m(x_0,x)\,[g(x) - g(x_0)]\,dx \right| \tag{B.2}$$

A. Bias Error for Lipschitz R and f: For $R \in Lip(x_0;\alpha)$ and $f \in Lip(x_0;\beta)$ the function $g(x) = cR(x)f(x)$ is also locally Lipschitz, $g \in Lip(x_0;\gamma)$, and around the point x_0 we have

$$|g(x) - g(x_0)| \le L_g |x - x_0|^{\gamma}$$

where $L_g = |c|(M_f L_R + M_R L_f)$ and $\gamma = \min(\alpha,\beta)$ (see (2.2) and assumptions of case *A* in the theorem). Since for the scaling function ϕ satisfying condition (3.6) the corresponding kernel $q_m(x_0,x)$ fulfils condition (3.9) (see Section 3), we obtain

$$|G(x_0;m)| \le C L_g \int_{x_0-(s_2-s_1)/2^m}^{x_0+(s_2-s_1)/2^m} 2^m H(2^m|x-x_0|)\,|x-x_0|^{\gamma}\,dx$$

$$= 2CL_g 2^{-m\gamma} \int_0^{s_2-s_1} H(v)\,v^{\gamma}\,dv$$

This yields

$$|\text{bias}\{g_N(x_0;m)\}| = O\left(2^{-\gamma m}\right) \tag{B.3}$$

By similar reasoning one can infer that convergence in (A.6) is of order

$$|\text{bias}\{f_N(x_0;m)\}| = O\left(2^{-\beta m}\right) \tag{B.4}$$

B1. Bias Error for differentiable R and f: If $R \in C^p(x_0)$ and $f \in C^q(x_0)$ then (by the Leibnitz formula) $g = cRf \in C^t(x_0)$, where $t = \min(p,q)$. Applying the Taylor series expansion, we get around x_0:

$$\begin{aligned} g(x) - g(x_0) &= C_1(x-x_0) + C_2(x-x_0)^2 + \cdots \\ &\cdots + C_{t-1}(x-x_0)^{t-1} + \frac{1}{(t-1)!}\int_{x_0}^{x}(x-v)^{t-1} g^{(t)}(v)\,dv \end{aligned} \tag{B.5}$$

whence

$$|g(x) - g(x_0)| \le D\left\{|x-x_0| + |x-x_0|^2 + \cdots + |x-x_0|^t\right\}$$

for some positive constant *D*. Hence, owing to (B.2), for ϕ such as in (3.6) we obtain

$$|G(x_0;m)| \le CD \sum_{k=1}^{t} \int_{x_0-(s_2-s_1)/2^m}^{x_0+(s_2-s_1)/2^m} 2^m H(2^m|x-x_0|)\,|x-x_0|^k\,dx$$

$$= 2CD \sum_{k=1}^{t} 2^{-km} \int_0^{s_2-s_1} H(v)\,v^k\,dv$$

which for large values of *m* results in

$$|\text{bias}\{g_N(x_0;m)\}| = O\left(2^{-m}\right) \tag{B.6}$$

for each *p,q*. By the same arguments

$$|\text{bias}\{f_N(x_0;m)\}| = O\left(2^{-m}\right) \tag{B.7}$$

B2. Bias Error for differentiable R and f and ϕ with vanishing moments: For ϕ with *r* vanishing moments (cf. (3.19)) one can ascertain that for each $n_{\min}(x_0;m) \le n \le n_{\max}(x_0;m)$ and large values of *m* it holds

$$\int_{x_0-(s_2-s_1)/2^m}^{x_0+(s_2-s_1)/2^m} (x-x_0)^k \varphi_{mn}(x)\,dx = 0, \qquad k = 1,2,\ldots,r$$

where $n_{\min}(x_0;m)$, $n_{\max}(x_0;m)$ are as in (3.11). Hence, for the correspondent kernel $q_m(x_0,x)$ we have (see (3.14))

$$\int_{x_0-(s_2-s_1)/2^m}^{x_0+(s_2-s_1)/2^m} (x-x_0)^k q_m(x_0,x)\,dx = 0, \qquad k = 1,2,\ldots,r$$

Consequently, using (B.2) and (B.5) and next including the bound in (3.9) and the fact that

$$\int_{x_0-(s_2-s_1)/2^m}^{x_0+(s_2-s_1)/2^m} |q_m(x_0,x)|\,|x-x_0|^k\,dx \le 2C(s_2-s_1)^{k+1}\,2^{-km}$$

$$\int_{x_0-(s_2-s_1)/2^m}^{x_0+(s_2-s_1)/2^m} |q_m(x_0,x)|\,\left|\frac{1}{(t-1)!}\int_{x_0}^{x}(x-v)^{t-1}g^{(t)}(v)\,dv\right|dx \le 2CD(s_2-s_1)^{t+1}\,2^{-tm}$$

we obtain for $R \in C^p(x_0)$, $f \in C^q(x_0)$ the following

$$|G(x_0;m)| \le \begin{cases} 2CD\sum_{k=r+1}^{t}(s_2-s_1)^{k+1}\,2^{-km} & \text{for } r < t \\ 2CD(s_2-s_1)^{t+1}\,2^{-tm} & \text{for } r \ge t \end{cases}$$

where *D* is a positive constant and $t = \min(p,q)$. This, for large *m*, implies that

$$|\text{bias}\{g_N(x_0;m)\}| = O\left(2^{-\rho m}\right) \tag{B.8}$$

where $\rho = \min(t,r+1) = \min(p,q,r+1)$. By the same reasoning we find that

$$|\text{bias}\{f_N(x_0;m)\}| = O\left(2^{-\tau m}\right) \tag{B.9}$$

with $\tau = \min(q,r+1)$.

Combining (B.3), (B.6), (B.8) with (B.4), (B.7), (B.9) we jointly obtain that

$$|\text{bias}\{g_N(x_0;m)\}| = O\left(2^{-\delta m}\right); \qquad |\text{bias}\{f_N(x_0;m)\}| = O\left(2^{-\nu m}\right) \tag{B.10}$$

where respectively

$$\delta = \begin{cases} \min(\alpha,\beta) \\ 1 \\ \min(p,q,r+1) \end{cases}; \qquad \nu = \begin{cases} \beta & \textit{for the case A} \\ 1 & \textit{for the case B1} \\ \min(q,r+1) & \textit{for the case B2} \end{cases} \tag{B.11}$$

Proof of the Theorem: Using the above bias asymptotic bounds along with the variance asymptotic expressions (A.8) and (A.9), for white as well as coloured noise we get

$$E[g_N(x_0;m) - g(x_0)]^2 = O\left(2^{-2\delta m}\right) + O\left(2^m/N\right) \tag{B.12}$$

and

$$E[f_N(x_0;m) - f(x_0)]^2 = O\left(2^{-2\nu m}\right) + O\left(2^m/N\right) \tag{B.13}$$

Further few steps are based on the following lemma referring to the ratio estimates and resulting

from Chebyshev's inequality (see Lemma 2 in [15]).

Lemma: If $R_N(x;m) = g_N(x;m)/f_N(x;m)$ is an estimate of $cR(x) = g(x)/f(x)$ and at the point x_0 such that $f(x_0) > 0$ it holds

$$E[g_N(x_0;m) - g(x_0)]^2 = O(a_N)$$

$$E[f_N(x_0;m) - f(x_0)]^2 = O(b_N)$$

then

$$|R_N(x_0;m) - cR(x_0)| = O\left(\sqrt{\max(a_N, b_N)}\right) \qquad \textit{in probability}$$

where for a sequence of random variables $\zeta_N = O(c_N)$ *in probability* means that $d_N\zeta_N / c_N \to 0$ *in probability* as $N \to \infty$ for any number sequence $\{d_N\}$ convergent to zero.

Taking account of (B.12)-(B.13), noticing that in each case under investigation $\delta \le \nu$ ((B.11)) and employing the lemma, we conclude that for x_0 where $f(x_0) > 0$ and for a given resolution level selection rule $m = m(N)$ the convergence rate is

$$|R_N(x_0;m) - cR(x_0)| = O\left(\sqrt{2^{-2\delta m(N)} + 2^{m(N)}/N}\right) \qquad \textit{in probability} \qquad \text{(B.14)}$$

Minimization of the right hand side of (B.14) yields directly the optimum selection law in (5.4) and the asymptotic rate of convergence in (5.3). □

References

[1] A. Antoniadis and G. Oppenheim (Ed.), Wavelets and Statistics, Springer-Verlag, New York, 1995.

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

[3] P.J. Bickel and K.A. Doksum, Mathematical Statistics: Basic Ideas and Selected Topics, Hoden-Day, San Francisco, 1977.

[4] H.J. Bierens, Kernel Estimators of Regression Functions, Cambridge University Press, Cambridge, 1987.

[5] S.A. Billings, "Identification of Non-Linear Systems: A Survey", Proc. IEE, Vol. 127, No. 6, 1980, pp. 272-285.

[6] S.A. Billings and S.Y. Fakhouri, "Identification of Systems Containing Linear Dynamic and Static Non-Linear Elements", Automatica, Vol. 18, No. 1, 1982, pp. 15-26.

[7] H.W. Chen, "Modeling and Identification of Parallel Non-Linear Systems: Structural Classification and Parameter Estimation Methods", Proc. IEEE, Vol. 83, No. 1, 1995, pp. 39-66.

[8] C.K. Chui, An Introduction to Wavelets, Academic Press, New York, 1992.

[9] C.K. Chui, Wavelets: A Mathematical Tool for Signal Processing, SIAM Edition, Philadelphia, 1997.

[10] I. Daubechies, Ten Lectures on Wavelets, SIAM Edition, Philadelphia, 1992.

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

[12] E. Eskinat, S.H. Johnson and W.L. Luyben, "Use Hammerstein Models in Identification of Non-Linear Systems", Amer. Instit. Chem. Eng., Vol. 37, 1991, pp. 255-268.

[13] W. Greblicki, "Nonparametric Orthogonal Series Identification of Hammerstein Systems", Int. J. Systems Sci., Vol. 20, No. 12, 1989, pp. 2355-2367.

[14] W. Greblicki, "Nonlinearity Estimation in Hammerstein Systems Based on Ordered

Observations", IEEE Trans. Signal Process., Vol. 44, No. 5, 1996, pp. 1224-1233.
[15] W. Greblicki and M. Pawlak, "Fourier and Hermite Series Estimates of Regression Functions", Ann. Inst. Statist. Math., Vol. 37A, No. 3, 1985, pp. 443-454.
[16] W. Greblicki and M. Pawlak, "Identification of Discrete Hammerstein Systems Using Kernel Regression Estimates", IEEE Trans. Automat. Control, Vol. 31, No. 1, 1986, pp. 74-77.
[17] W. Greblicki and M. Pawlak, "Hammerstein System Identification by Non-Parametric Regression Estimation", Int. J. Control, Vol. 45, No. 1, 1987, pp. 343-354.
[18] W. Greblicki and M. Pawlak, "Non-Parametric Identification of Hammerstein Systems", IEEE Trans. Inform. Theory, Vol. 35, No. 2, 1989, pp. 409-418.
[19] W. Greblicki and M. Pawlak, "Non-Parametric Identification of a Cascade Non-Linear Time Series System", Signal Processing, Vol. 22, No. 1, 1991, pp. 61-75.
[20] W. Greblicki and M. Pawlak, "Cascade Non-Linear System Identification by a Non-Parametric Method", Int. J. Systems Sci., Vol. 25, No. 1, 1994, pp. 129-153.
[21] L. Györfi, W. Härdle, P. Sarda and P. Vieu, Nonparametric Curve Estimation from Time Series, Springer-Verlag, Berlin, 1989.
[22] R. Haber and H. Unbehauen, "Structure Identification of Non-Linear Dynamic Systems: A Survey on Input/Output Approaches", Automatica, Vol. 26, No. 4, 1990, pp. 651-677.
[23] W. Härdle, Applied Nonparametric Regression, Cambridge University Press, Cambridge, 1990.
[24] W. Härdle, G. Kerkyacharian, D. Picard and A. Tsybakov, Wavelets, Approximation, and Statistical Applications, Springer-Verlag, New York, 1998.
[25] Z. Hasiewicz, "Hammerstein System Identification by the Haar Multiresolution Approximation", Int. J. Adapt. Control Signal Process., Vol. 13, 1999, in press.
[26] Z. Hasiewicz and W. Greblicki, "Non-Linearity Recovering with the Help of Wavelets", Proc. European Control Conf., Karlsruhe, Germany, Aug. 31-Sep. 3, 1999, paper # F1031-1 (on CD).
[27] I.W. Hunter and M.J. Korenberg, "The Identification of Non-Linear Biological Systems: Wiener and Hammerstein Cascade Models", Biol. Cybern., Vol. 55, 1986, pp. 135-144.
[28] A. Juditsky, H. Hjalmarsson, A. Benveniste, B. Delyon, L. Ljung, J. Sjöberg and Q. Zhang, "Nonlinear Black-Box Models in System Identification: Mathematical Foundations", Automatica, Vol. 31, No. 12, 1995, pp. 1725-1750.
[29] S. Kelly, M. Kon and L. Raphael, "Local Convergence of Wavelet Expansions", J. Funct. Analys., Vol. 126, 1994, pp. 102-138.
[30] S. Kelly, M. Kon and L. Raphael, "Pointwise Convergence of Wavelet Expansions", Bull. Amer. Math. Soc., Vol. 30, No. 1, 1994, pp. 87-94.
[31] A. Krzyzak, "Identification of Discrete Hammerstein Systems by the Fourier Series Regression Estimate", Int. J. Systems Sci., Vol. 20, No. 9, 1989, pp. 1729-1744.
[32] A. Krzyzak, "On Estimation of a Class of Non-Linear Systems by the Kernel Regression Estimate", IEEE Trans. Inform. Theory, Vol. 36, No. 1, 1990, pp. 141-152.
[33] S. Mallat, "Multiresolution Approximations and Wavelet Orthonormal Bases of $L^2(R)$", Trans. Amer. Math. Soc., Vol. 315, No. 1, 1989, pp. 69-87.
[34] S. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", IEEE Trans. Pattern Analysis Machine Intell., Vol. 11, No. 7, 1989, pp. 674-693.
[35] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, San Diego, 1998.
[36] E.A. Nadaraya, "On Estimating Regression", Theory Probab. Appl., Vol. 9, 1964, pp. 141-142.
[37] R.T. Ogden, Essential Wavelets for Statistical Applications and Data Analysis, Birkhäuser,

Boston, 1997.
[38] M. Pawlak, "On the Series Expansion Approach to the Identification of Hammerstein Systems", IEEE Trans. Automat. Control, Vol. 36, 1991, pp. 763-767.
[39] M. Pawlak and Z. Hasiewicz, "Non-Linear System Identification by the Haar Multiresolution Analysis", IEEE Trans. Circuits Systems, Vol. 45, No. 9, 1998, pp. 945-961.
[40] M. Pawlak and Z. Hasiewicz, "Non-Parametric System Identification of Non-Linear Block-Oriented Systems by Multiscale Expansions", Proc. European Control Conf., Karlsruhe, Germany, Aug. 31-Sep. 3, 1999, paper # F1065-6 (on CD).
[41] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delyon, P.-Y. Glorennec, H. Hjalmarsson and A. Juditsky, "Nonlinear Black-Box Modeling in System Identification: A Unified Overview", Automatica, Vol. 31, No. 12, 1995, pp. 1691-1724.
[42] T. Söderström and P. Stoica, System Identification, Prentice-Hall, New York, 1989.
[43] T. Söderström, W.X. Zheng and P. Stoica, "Comments on 'On a Least-Squares-Based Algorithm for Identification of Stochastic Linear Systems'", IEEE Trans. Signal Process., Vol. 47, No. 5, 1999, pp. 1395-1396.
[44] C.J. Stone, "Optimal Rates of Convergence for Nonparametric Estimators", Ann. Statist., Vol. 8, No. 6, 1980, pp. 1348-1360.
[45] N. Sureshbabu and J.A. Farrell, "Wavelet-Based System Identification for Nonlinear Control", IEEE Trans. Automat. Control, Vol. 44, No. 2, 1999, pp. 412-417.
[46] M. Vetterli and J. Kovacevic, Wavelets and Subband Coding, Prentice-Hall, Englewood Cliffs, 1995.
[47] G.G. Walter, Wavelets and Other Orthogonal Systems with Applications, CRC Press, Boca Raton, 1994.
[48] M.P. Wand and M.C. Jones, Kernel Smoothing, Chapman and Hall, London, 1995.
[49] G.S. Watson, "Smooth Regression Analysis", Sankhya Ser. A, Vol. 26, 1964, pp. 359-372.
[50] P. Wojtaszczyk, A Mathematical Introduction to Wavelets, Cambridge University Press, Cambridge, 1997.
[51] J. Zhang, G.G. Walter, Y. Miao and W.N.W. Lee, "Wavelet Neural Networks for Function Learning", IEEE Trans. Signal Process., Vol. 43, No. 6, 1995, pp. 1485-1497.
[52] W.X. Zheng, "On a Least-Squares-Based Algorithm for Identification of Stochastic Linear Systems", IEEE Trans. Signal Process., Vol. 46, No. 6, 1998, pp. 1631-1638.

CAPTIONS FOR FIGURES

Figure 1. The Hammerstein system disturbed by the noise z_k: (*a*) - white noise case, (*b*) - coloured noise case.

Figure 2. *MISE* versus *N*; quantizer non-linearity, *FIR*(4) dynamics, $\sigma_z = 5\%\max|y_k|$, various correlation of noise.