# Computational algorithms for multiscale identification of nonlinearities in Hammerstein systems with random inputs[‡]

Przemysław Śliwiński[*] and Zygmunt Hasiewicz[†]

**Abstract**

There are well known procedures for computing values of compactly supported wavelets in binary grid points. Such algorithms are inherently well suited for solving system identification tasks with fixed input design. We show that they can be also efficiently used for the solution of system identification problems with random $x$-variables.

**Index Terms**

compactly supported wavelets, computational algorithm, system identification, Hammerstein system

## I. Introduction

It is well known that compactly supported wavelet functions invented by Daubechies have no explicit formulas (except for Haar family) but are defined by recursive procedures. These procedures are computationally simple and fast, however, provide with exact values of wavelet functions only at binary grid points $2^{-H}b$, $H, b = \ldots, -1, 0, 1, \ldots$, where $H < \infty$, i.e. for arguments with finite binary representation; see Daubechies [3], Strang [18]. This is not a disadvantage when excitations (arguments) are deterministic and equidistant (e.g. in signal and image processing, Mallat [2], [14], numerical algebra, Ruskai *et al.* [17], etc.) but becomes a shortcoming when inputs are random. This problem has been already investigated only in relation to computing wavelet expansion coefficients; see Delyon and Juditsky [4], Antoniadis, Grégoire and Vial [1], Härdle *et al.* [9], Kovacz and Silverman [13] and Györfi *et al.* [7].

[‡] This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

[*][†] Przemysław Śliwiński and Zygmunt Hasiewicz are with Institute of Engineering Cybernetics, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland

In this correspondence we examine the problem in the context of nonparametric identification of nonlinear characteristics of Hammerstein systems, with the use of compactly supported wavelet scaling functions. Adequate identification procedures, worked out in [10], provide potentially efficient recovering of nonlinearities in Hammerstein systems under random inputs, however, the aforementioned limitations make it necessary in practice to create computational counterparts of the theoretical algorithms to deal efficiently with random inputs. The clue of our proposition is to round-off random input data to the neighboring binary grid points and substitute scaling functions by their proper piecewise-constant approximations. This leads to computationally simple procedures which, under easy to fulfill conditions, maintain convergence and convergence rate of the prototype.

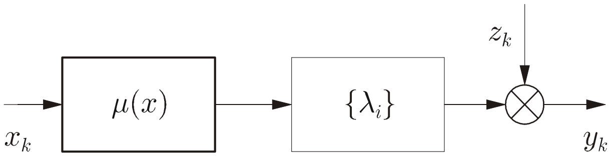

Fig. 1. Hammerstein system

We start with short presentation of the reference multiscale identification algorithm and its limit properties. Then we construct a class of the computational algorithms and examine their asymptotic behavior, as well. Specifically, we establish convergence conditions and provide rate of convergence of the practical computational algorithms in dependence of smoothness of the identified nonlinearity, input density (assumed to exist) and regularity of the wavelet scaling function.

## II. Reference multiscale identification algorithm

The reference wavelet algorithm for identifying nonlinear characteristics $\mu(x)$ in Hammerstein systems (see Fig. 1) from random input-output data $\{(x_k, y_k)\}$, $k = 1, \ldots, N$, has the form [10]:

$$\hat{\mu}(x) = \frac{\sum\limits_{n=n_{\min}(x)}^{n_{\max}(x)} \hat{\alpha}_{mn}\varphi_{mn}(x)}{\sum\limits_{n=n_{\min}(x)}^{n_{\max}(x)} \hat{a}_{mn}\varphi_{mn}(x)}, \qquad \begin{array}{l} \hat{\alpha}_{mn} = N^{-1}\sum\limits_{k=1}^{N}\varphi_{mn}(x_k)\, y_k \\ \hat{a}_{mn} = N^{-1}\sum\limits_{k=1}^{N}\varphi_{mn}(x_k) \end{array} \tag{1}$$

where

$$\varphi_{mn}(x) = 2^{m/2}\varphi(2^m x - n),\ m, n = \ldots -1, 0, 1, \ldots$$

are dilations and translations of a compactly supported wavelet scaling function $\varphi(x)$ from, e.g., Daubechies, symmlet or coiflet family. Parameter $m$ is referred to as a scaling factor and translation limits $n_{\min}(x)$

and $n_{\max}(x)$ depend on $x$ and are related to the supports of the employed scaling functions; see Table I in Appendix IV.

We assume the following conditions (cf. [10]):

**A1.** The nonlinearity $\mu(x)$ (to be identified) is bounded.

**A2.** The system input $\{x_k\}$, $k = \ldots, -1, 0, 1, \ldots,$ is an i.i.d. random process with finite variance possessing bounded probability density function $f(x)$.

**A3.** Dynamic element of the system, with the impulse response $\{\lambda_i\}$, $i = 0, 1, \ldots,$ is asymptotically stable.

**A4.** The output noise $\{z_k\}$, $k = \ldots, -1, 0, 1, \ldots,$ is zero mean stationary process with finite variance, white or correlated.

*Remark 1:* Using only outer signals $\{(x_k, y_k)\}$, due to the complex structure of the system one cannot certainly identify the true system nonlinearity $\mu(x)$ but only its scaled and shifted version $\mu_0(x) = \lambda_0\mu(x) + d$, $\lambda_0 \neq 0$ by assumption, $d = \mathrm{E}\mu(x_1)\sum_{i=1}^{\infty}\lambda_i$ (cf. e.g. [6], [10] or [15]).

As was shown in [10], the following holds true.

*Corollary 1 ([10, Th.1]):* If the scaling factor $m$ satisfies the conditions

$$m \to \infty \text{ and } 2^m/N \to 0 \text{ as } N \to \infty \tag{2}$$

then

$$\hat{\mu}(x) \to \mu_0(x) \text{ in probability as } N \to \infty,$$

at all points $x$ in which $\mu(x)$ and $f(x)$ are simultaneously continuous and $f(x) > 0$.

To assure convergence, it is thus enough to grow the scale $m$ with the number of data $N$ with arbitrary rate less than $\log_2 N$. However, to make the convergence the fastest possible, the growth of $m$ must be strictly related to the smoothness of the underlying nonlinearity $\mu(x)$, input probability density function $f(x)$ and the properties of employed scaling function $\varphi(x)$. Denoting by $\mathcal{C}^{\lambda}(x-\varepsilon, x+\varepsilon)$ the class of functions having, in the $\varepsilon$-neighborhood of $x$, $r = \lceil\lambda\rceil - 1$ derivatives with the last one being Hölder continuous with exponent $\lambda - r$ (cf. e.g. [8], [14]), we get:

*Theorem 1 (cf. [10, Th. 2]):* If $\mu(x) \in \mathcal{C}^{\lambda_\mu}(x-\varepsilon, x+\varepsilon)$ and $f(x) \in \mathcal{C}^{\lambda_f}(x-\varepsilon, x+\varepsilon)$, $\lambda_\mu, \lambda_f > 0$, and if the scale $m$ is selected as

$$m = \left\lceil \frac{1}{2\gamma+1}\log_2 2\gamma N \right\rceil \text{ with } \gamma = \min\{\lambda_\mu, \lambda_f, p\} \tag{3}$$

where $p$ is a number of vanishing moments of the wavelet function $\psi(x)$ associated with the scaling function $\varphi(x)$, then the algorithm in (1) achieves the convergence rate

$$|\hat{\mu}(x) - \mu_0(x)| = O\left(N^{-\gamma/(2\gamma+1)}\right) \text{ in probability.}$$

*Proof:* See Appendix I. ■

*Remark 2:* To achieve the fastest rate of convergence for a given nonlinearity $\mu(x)$ and input probability density $f(x)$, it is necessary to apply scaling functions with $p \geq \min\{\lambda_\mu, \lambda_f\}$. However, the smoothness of $\mu(x)$ and $f(x)$ is usually unknown and, moreover, can vary for different $x$'s (e.g. for splines). Practically relevant issue of selection of proper $p$ in such circumstances is discussed in [11, Section 6.4].

## III. Computational algorithm

Since we are not able to compute easily values of scaling functions $\varphi(x)$ in arbitrary points $x$, which is needed in (1), we approximate them by the values in appropriate binary grid points, exactly calculable for any $H$; see (20) in Appendix IV. For a given $x$, these binary points can be chosen threefold:

$$b_H(x) = \begin{cases} 2^{-H}\left\lfloor 2^H x\right\rfloor \\ 2^{-H}\left\lfloor 2^H x + 1/2\right\rfloor \\ 2^{-H}\left\lceil 2^H x\right\rceil \end{cases}$$

i.e., as the left-nearest, the nearest, or the right-nearest binary grid neighbors of $x$, respectively. Denoting $\bar{x}_{Hm} = 2^{-m} b_H(2^m x)$, we propose the following approximators of $\varphi_{mn}(x)$:

$$\bar{\varphi}^H_{mn}(x) = \varphi_{mn}(\bar{x}_{Hm}) = 2^{m/2}\varphi\left(b_H(2^m x) - n\right). \tag{4}$$

*Remark 3:* Since $\operatorname{supp}\varphi_{mn}(x) = \left[2^{-m}(s_1 + n), 2^{-m}(s_2 + n)\right]$, some integer $s_1, s_2$ (see Table I in Appendix IV) and $|x - \bar{x}_{Hm}| < 2^{-(m+H)}$ for any $b_H(x)$ under consideration thus if $x \in \operatorname{supp}\varphi_{mn}(x)$ then also $\bar{x}_{Hm} \in \operatorname{supp}\varphi_{mn}(x)$ for any $b_H(x)$ and $H \geq 0$.

Our idea is to use the according approximators in place of their prototypes $\varphi_{mn}(x)$ in the reference algorithm (1), getting the following plug-in generic computational algorithm:

$$\tilde{\mu}(x) = \frac{\sum\limits_{n=n_{\min}(x)}^{n_{\max}(x)} \tilde{\alpha}_{mn}\bar{\varphi}^H_{mn}(x)}{\sum\limits_{n=n_{\min}(x)}^{n_{\max}(x)} \tilde{a}_{mn}\bar{\varphi}^H_{mn}(x)}, \qquad \begin{array}{l} \tilde{\alpha}_{mn} = N^{-1}\sum\limits_{k=1}^{N}\bar{\varphi}^H_{mn}(x_k)\, y_k \\ \tilde{a}_{mn} = N^{-1}\sum\limits_{k=1}^{N}\bar{\varphi}^H_{mn}(x_k) \end{array} \tag{5}$$

Observe that approximators are used both for computing empirical wavelet coefficients and the values of the estimate $\tilde{\mu}(x)$ at each point $x$. It is clear that usage of these approximators instead of genuine scaling functions will cause an additional bias error in comparison with the reference algorithm. However, with growing $H$ this influence is reduced.

*Theorem 2:* Let the assumptions about $\mu(x)$ and $f(x)$ of Theorem 1 hold and let $\varphi(x) \in C^{\lambda_\varphi}$ $(x-\varepsilon, x+\varepsilon), \lambda_\varphi > 0$. If (2) is in force and

$$H \to \infty \tag{6}$$

then

$$\tilde{\mu}(x) \to \mu_0(x) \text{ in probability as } N \to \infty.$$

*Proof:* See Appendix II. ■

Note that to ensure convergence, the rate at which $H$ tends to infinity can be arbitrarily slow (see (18) and (19) in Appendix II). However, to maintain the original (as in Theorem 1) convergence rate for the computational algorithm, the choice of the binary grid factor $H$ must be more precise.

*Theorem 3:* Let all assumptions of Theorem 2 hold. If $m$ is selected as in (3) and $H$ is set according to the rule

$$H = \left\lceil \frac{\gamma m}{\eta} \right\rceil \text{ where } \eta = \min\{\lambda_\varphi, 1\} \tag{7}$$

then the computational algorithm (5) preserves the convergence rate of the prototype (1), i.e.

$$|\tilde{\mu}(x) - \mu_0(x)| = O\left(N^{-\gamma/(2\gamma+1)}\right) \text{ in probability.}$$

*Proof:* See Appendix III. ■

The choice of a suitable grid factor $H$ depends therefore on three elements: the scale $m$, the number $p$ of vanishing moments of wavelet $\psi(x)$ associated with the replaced scaling function $\varphi(x)$, and the smoothnesses of $\varphi(x)$, $\mu(x)$ and $f(x)$. Even if the latter are unknown, $H$ can still be 'safely' determined as follows, based merely on the properties of scaling functions

$$H = \left\lceil \frac{pm}{\eta} \right\rceil. \tag{8}$$

This rule is certainly too pessimistic when $\min\{\lambda_\mu, \lambda_f\} < p$ (i.e., produces proper, but 'too big', factors $H$), but equal to (7) otherwise. It is moreover obvious that all $H$, greater than those determined by (7) or (8), are, in view of Theorem 3, admissible as well.

*Remark 4:* In spite of the fact that all the results above are asymptotic (i.e., true for large $N$), one can infer from (3), (7) and (8), a practical rule for selection of $H$:

$$H = \left\lceil \frac{p}{\eta(2p+1)} \log_2 2pN \right\rceil. \tag{9}$$

*Remark 5:* To be concise, our considerations have been confined to the Hammerstein class of nonlinear systems. However, Corollary 1 and Theorems 1-3 are also valid for other nonlinear block-oriented systems, e.g. multichannel systems; see [6], [11], [12], [15], [16].

## IV. Simulation study

The performance of the computational algorithm (5) for small and moderate number of data has been investigated by means of computer simulations. During simulations, system input $\{x_k\}$ and output noise $\{z_k\}$ were white and uniformly distributed on the intervals $[-0.6, 0.6]$ and $[-0.1, 0, 1]$, respectively. Two nonlinearities were included into test: smooth, $\mu_1(x) = \arctan(x)$ and nonsmooth, $\mu_2(x) = \operatorname{sgn}(x)$. Dynamic element with finite impulse response $\{\lambda_i\} = \{1, 1/2, 1/4, 1/8\}$ was selected as output dynamics. The computational algorithm was based on the approximations of the third Daubechies scaling function $\varphi(x)$ (i.e. for $p = 3$ and $\lambda_\varphi \cong 1.018$; see [3]). The scales $m$ and $H$ were set according to the rules in (3) and (9), yielding $m = \lceil 1/7 \log_2 6N \rceil$ and $H = \lceil 3/7 \log_2 6N \rceil$, respectively. As the reference algorithm we took the computational one with $H = 15$ (which ensures high quality approximation of the scaling function but through (9) corresponds to rather impractical number of data $N \cong 5.72 \times 10^9$). The performance of the algorithm was measured globally by numerically computed MISE error (determined in the interval $[-0.5, 0.5]$ to reduce the influence of boundary effects).

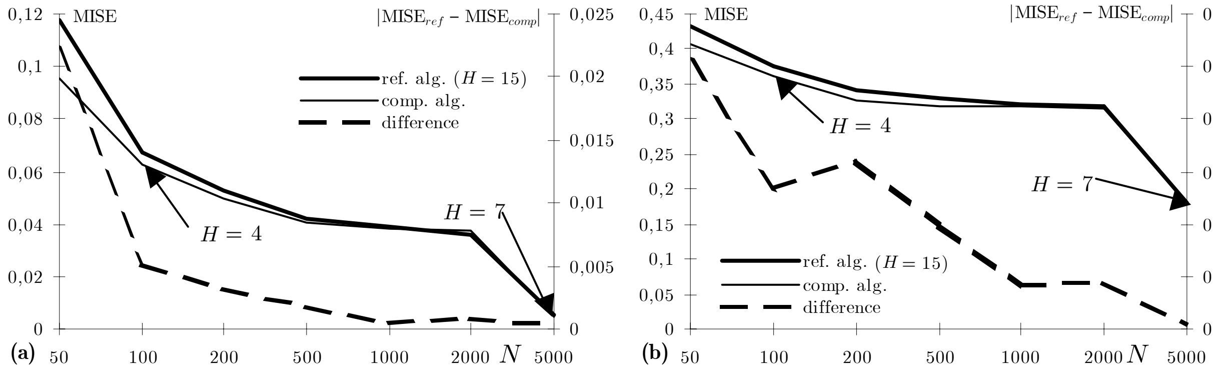

Fig. 2. MISE errors of the reference and computational algorithms (left $y$-axes) and differences between the MISE errors (right $y$-axes) for nonlinearities **(a)** $\mu_1(x)$ and **(b)** $\mu_2(x)$

The results are shown in Fig. 2. They reveal only a little difference in quality between the computational and reference algorithm, not exceeding 10% of the MISE error of the latter for $N > 100$. Since for $N = 100$ we have $H = 4$ and for $N = 5000$ we get merely $H = 7$, the plots reveal also that high precision in computation of wavelet function values is not necessary for efficiency of the identification algorithm, particularly for smaller number of data.

## V. Final remarks

We have proposed and examined a computational method which enables wavelet scaling function-based identification algorithm to work efficiently with random inputs. In our approach, hardly computable scaling functions in the original algorithm were replaced by their approximations which can be easily computed by the standard method (as shown in Appendix IV). Our algorithm possesses the same limit properties as the theoretical prototype and performs well also for moderate number of data.

We emphasize simplicity and efficiency of the algorithm. Simplicity comes from piecewise-constant approximations, which are particularly convenient to program or to embed in a hardware. In turn, efficiency is the consequence of the fact that the prospective grid factor $H$ can be quite small in practice, and therefore, the numerical overhead imposed by the conventional procedure (20) reported in Appendix IV can have small impact on the overall computations in the identification algorithm. This is especially the case when $N$, the number of data points $\{(x_k, y_k)\}$, is known in advance. Then the approximators (i.e. $\varphi$-values at the binary grid points with adequate $H$) can be calculated beforehand and stored in a computer memory.

## Appendix I
## Proof of Theorem 1

Denote by $\hat{g}(x)$ and $\hat{f}(x)$ the numerator and denominator of the algorithm $\hat{\mu}(x)$, respectively.

For fixed $x$, we have

$$\operatorname{MSE} \hat{g}(x) = \operatorname{bias}^2 \hat{g}(x) + \operatorname{var} \hat{g}(x) .$$

Under assumptions A1-A4, for the variance term it holds that [10, Appendix A].

$$\operatorname{var} \hat{g}(x) = O\left(2^m N^{-1}\right) . \tag{10}$$

for each $x$. To compute bias error observe that $\operatorname{E} \hat{\alpha}_{mn} = \langle g(x), \varphi_{mn}(x) \rangle$ , see (1), i.e. empirical coefficients $\hat{\alpha}_{mn}$ are unbiased estimators of inner products $\alpha_{mn} = \langle g(x), \varphi_{mn}(x) \rangle$ where $g(x) = \mu_0(x) f(x)$. Hence

$$\operatorname{bias} \hat{g}(x) = \sum_{k=m}^{\infty} \sum_{l=l_{\min}(x)}^{l_{\max}(x)} \beta_{kl} \psi_{kl}(x) , \quad \beta_{kl} = \langle g(x), \psi_{kl}(x) \rangle \tag{11}$$

where summation limits $l_{\min}(x)$, $l_{\max}(x)$ are related to supports of the wavelets $\psi_{kl}(x)$ associated with the scaling function $\varphi(x)$ (see Table I). Since $g(x) \in \mathcal{C}^{\lambda_g}(x-\varepsilon, x+\varepsilon), \lambda_g = \min\{\lambda_\mu, \lambda_f\}$, we can expand $g(x)$ in the neighborhood $(x-\varepsilon, x+\varepsilon)$ into the following Taylor series; see Mallat [14, Section 6.1.1]:

$$g(x) = \sum_{r=0}^{\lceil\lambda_g\rceil-1} G_r \cdot (x-v)^r + \epsilon(x,v), \quad v \in (x-\varepsilon, x+\varepsilon)$$

where $G_r = g^{(r)}(v)/r!$ and $\epsilon(x,v) \leq L_g |x-v|^{\lambda_g}$, some $L_g > 0$. Hence

$$\beta_{kl} = \langle G_0 \cdot (x-v), \psi_{kl}(x)\rangle + \cdots + \langle G_{\lceil\lambda_g\rceil-1} \cdot (x-v)^{\lceil\lambda_g\rceil-1}, \psi_{kl}(x)\rangle + \langle \varepsilon(x,v), \psi_{kl}(x)\rangle$$

As by assumption $\psi(x)$ has $p$ vanishing moments, we get (cf. [14, Section 6.1.3 and Th. 9.7])

$$|\beta_{kl}| = O\left(2^{-k\left(\gamma_g+1/2\right)}\right), \quad \gamma_g = \min\{\lambda_g, p\}.$$

This along with (11) leads to

$$|\mathrm{bias}\, \hat{g}(x)| = O\left(2^{-\gamma_g m}\right). \tag{12}$$

Including (10), (12) and using the scale selection rule (3) results in

$$\mathrm{MSE}\, \hat{g}(x) = O\left(N^{-2\gamma_g/\left(2\gamma_g+1\right)}\right). \tag{13}$$

The same routine applied to the denominator $\hat{f}(x)$ gives

$$\mathrm{MSE}\, \hat{f}(x) = O\left(N^{-2\gamma_f/\left(2\gamma_f+1\right)}\right), \quad \gamma_f = \min\{\lambda_f, p\} \tag{14}$$

Since $\min\left\{\gamma_g, \gamma_f\right\} = \gamma_g = \gamma$, the conclusion is obtained by virtue of the following lemma (see Greblicki and Pawlak [5, Th. 3]).

*Lemma:* If $\mathrm{MSE}\, \hat{g}(x) = O\left(N^{-a}\right)$ and $\mathrm{MSE}\, \hat{f}(x) = O\left(N^{-b}\right)$, then for $\hat{\mu}(x) = \hat{g}(x)/\hat{f}(x)$ and $\mu_0(x) = g(x)/f(x)$ it holds that

$$|\hat{\mu}(x) - \mu_0(x)| = O\left(N^{-\min\{a,b\}/2}\right) \text{ in probability.}$$

## APPENDIX II

## PROOF OF THEOREM 2

Denote by $\tilde{g}(x)$ and $\tilde{f}(x)$ the numerator and denominator of $\tilde{\mu}(x)$ in (5). Consider mean square error of $\tilde{g}(x)$ for fixed $x$:

$$\mathrm{MSE}\, \tilde{g}(x) = \mathrm{bias}^2\, \tilde{g}(x) + \mathrm{var}\, \tilde{g}(x)$$

Since each approximator $\bar{\varphi}^{H}_{mn}(x)$ in (4) is bounded and compactly supported (as the original $\varphi_{mn}(x)$), the variance component has the same order upper bound as the variance of the reference algorithm, i.e.

$$\operatorname{var}\tilde{g}(x) = O\left(2^{m}N^{-1}\right) \tag{15}$$

Bias error can be split into two parts

$$\begin{aligned} \operatorname{bias}\tilde{g}(x) &= [g(x) - \operatorname{E}\hat{g}(x)] + [\operatorname{E}\hat{g}(x) - \operatorname{E}\tilde{g}(x)] \\ &= \operatorname{bias}\hat{g}(x) + \overline{\operatorname{bias}}\,\tilde{g}(x) \end{aligned}$$

where $\operatorname{bias}\hat{g}(x)$ is as in (11) and

$$\overline{\operatorname{bias}}\,\tilde{g}(x) = \sum_{n=n_{\min}(x)}^{n_{\max}(x)} \left[\alpha_{mn}\varphi_{mn}(x) - \bar{\alpha}_{mn}\bar{\varphi}^{H}_{mn}(x)\right] \tag{16}$$

with $\bar{\alpha}_{mn} = \left\langle g(x), \bar{\varphi}^{H}_{mn}(x)\right\rangle$ is the approximation error induced by replacing $\varphi_{mn}(x)$ with $\bar{\varphi}^{H}_{mn}(x)$ in the computational algorithm. To bound the expression in square brackets, we need the bound of the approximation error of $\varphi_{mn}(x)$ by $\bar{\varphi}^{H}_{mn}(x)$. Since by assumption, $\varphi(x) \in C^{\lambda_{\varphi}}(x-\varepsilon, x+\varepsilon)$, by the Taylor series expansion in the neighborhood $(x-\varepsilon, x+\varepsilon)$ of $x$ we have asymptotically (for large $m$; see Remark 3)

$$\varphi_{mn}(x) = \sum_{r=0}^{\lceil\lambda_{\varphi}\rceil-1} \Phi_{r}\cdot\left[2^{m}x - b_{H}(2^{m}x)\right]^{r} + \epsilon\left(2^{m}x, b_{H}(2^{m}x)\right)$$

where $\Phi_{r} = \varphi^{(r)}_{mn}(\bar{x}_{Hm})/r! = 2^{m/2}\varphi^{(r)}(b_{H}(2^{m}x) - n)/r!$ (see (4)) and where $|\epsilon(2^{m}x, b_{H}(2^{m}x))| \leq 2^{m/2}L_{\varphi}\left|2^{m}x - b_{H}(2^{m}x)\right|^{\lambda_{\varphi}}$, some $L_{\varphi} > 0$. Hence, for $\lambda_{\varphi} \in (0,1]$ we get that

$$\left|\varphi_{mn}(x) - \bar{\varphi}^{H}_{mn}(x)\right| \leq 2^{m/2}L_{\varphi}\left|2^{m}x - b_{H}(2^{m}x)\right|^{\lambda_{\varphi}}$$

and, for $\lambda_{\varphi} > 1$, that

$$\begin{aligned} \left|\varphi_{mn}(x) - \bar{\varphi}^{H}_{mn}(x)\right| \leq\; & |\Phi_{1}|\cdot\left|2^{m}x - b_{H}(2^{m}x)\right| + \cdots \\ & + \left|\Phi_{\lceil\lambda_{\varphi}\rceil-1}\right|\cdot\left|2^{m}x - b_{H}(2^{m}x)\right|^{\lceil\lambda_{\varphi}\rceil-1} + 2^{m/2}L_{\varphi}\left|2^{m}x - b_{H}(2^{m}x)\right|^{\lambda_{\varphi}} \end{aligned}$$

respectively. Since for each $x$ and arbitrary $b_{H}(x)$ from (4) it holds that $|x - b_{H}(x)| < 2^{-H}$ thus $|2^{m}x - b_{H}(2^{m}x)| < 2^{-H}$ for any $m$, and consequently we obtain the bound

$$\left|\varphi_{mn}(x) - \bar{\varphi}^{H}_{mn}(x)\right| = O\left(2^{m/2}2^{-\eta H}\right) \tag{17}$$

where $\eta = \min\{\lambda_{\varphi}, 1\}$. Using this bound and the following identity

$$\alpha_{mn}\varphi_{mn}(x) - \bar{\alpha}_{mn}\bar{\varphi}^{H}_{mn}(x) = \alpha_{mn}\left[\varphi_{mn}(x) - \bar{\varphi}^{H}_{mn}(x)\right] + \bar{\varphi}^{H}_{mn}(x)\left[\alpha_{mn} - \bar{\alpha}_{mn}\right]$$

and including that $|\alpha_{mn}| = O\left(2^{-m/2}\right)$, $\left|\bar{\varphi}^H_{mn}(x)\right| = O\left(2^{m/2}\right)$ and $|\alpha_{mn} - \bar{\alpha}_{mn}| = O\left(2^{-m/2}2^{-\eta H}\right)$, any $n$, yields immediately $\left|\alpha_{mn}\varphi_{mn}(x) - \bar{\alpha}_{mn}\bar{\varphi}^H_{mn}(x)\right| = O\left(2^{-\eta H}\right)$. The latter and (16) along with $n_{\max}(x) - n_{\min}(x) + 1 \leq c$, some $c$, for each $x$ (see Table I in Appendix IV) gives $\left|\overline{\text{bias}}\,\tilde{g}(x)\right| = O\left(2^{-\eta H}\right)$. Thus, owing to (12) we get $\text{bias}^2\,\tilde{g}(x) = O\left(2^{-2\gamma_g m}\right) + O\left(2^{-2\eta H}\right)$, and eventually (see (15))

$$\text{MSE}\,\tilde{g}(x) = O\left(2^{-2\gamma_g m}\right) + O\left(2^{-2\eta H}\right) + O\left(2^m N^{-1}\right) \tag{18}$$

The same routine applied to the error $\text{MSE}\,\tilde{f}(x)$ results in the bound

$$\text{MSE}\,\tilde{f}(x) = O\left(2^{-2\gamma_f m}\right) + O\left(2^{-2\eta H}\right) + O\left(2^m N^{-1}\right) \tag{19}$$

where $\gamma_f$ is as in (14). These errors tend to zero as $N \to \infty$ for $m$ and $H$ satisfying the convergence conditions (2) and (6), respectively, which concludes the proof. ■

## Appendix III

### Proof of Theorem 3

It is enough to observe that for $m$ and $H$ selected according to the rules (3) and (7) the MSE errors (18)-(19) of the computational algorithm (5) are of the same order as the corresponding MSE errors (13)-(14) of the prototype (1). ■

## Appendix IV

### Computing Daubechies scaling functions

To compute values of the $p$th Daubechies scaling function $\varphi$ in binary grid points

$$\mathbf{b} = \begin{bmatrix} b & b+1 & \cdots & b+(2p-2) \end{bmatrix}^T$$

where $b = 2^{-H}n, n = 0, 1, \ldots, 2^H - 1$, the following conventional formula can be employed (see Daubechies [3, Chapter 7.2] or Strang [18, Section 1.1]):

$$\varphi(\mathbf{b}) = \prod_{h=1}^{H} \left[(1-b_h)\,\mathbf{A} + b_h\mathbf{B}\right]\varphi(\mathbf{0}) \tag{20}$$

where $\varphi(\mathbf{b})$ is the resulting vector of exact values, $b_h \in \{0,1\}$ are consecutive bits of $b$, and where $\mathbf{A}$ and $\mathbf{B}$ are square, block-Toeplitz matrices $(2p-1)\times(2p-1)$, composed of the Daubechies scaling function coefficients $\{c_t\}$, $t = 0, \ldots, 2p-1$ (see e.g. Table 6.1 in Daubechies [3, Section 6.4]):

$$\mathbf{A} = [a_{ij}] = c_{2i-j} \quad \text{and} \quad \mathbf{B} = [b_{ij}] = c_{2i-j+1}, \quad i, j = 0, 1, \ldots, 2p-2$$

Vector $\varphi(\mathbf{0})$, providing with the values of $\varphi$ at integers, is the eigenvector associated with eigenvalue $1$ of the matrix with rows formed by terms of the right-hand side of dilations equations $\varphi(x) = \sum_{t=0}^{2p-1} c_t\varphi(2x-t)$ yielded for $x = 1, \ldots, 2p-2$, respectively; see [18, Section 1.1]. The algorithm, with adequately constructed matrices $\mathbf{A}$ and $\mathbf{B}$ and evaluated vectors $\varphi(\mathbf{b})$, can be used to compute symmlets and coiflets approximators as well. MATLAB implementation can be found e.g. at G. Strang's homepage: `http://www-math.mit.edu/~gs/`.

| | Daubechies/symmlet | coiflet |
|---|---|---|
| Support of $\varphi(x)$ | $[0, 2p-1]$ | $[-2p, 4p-1]$ |
| Support of $\psi(x)$ | $[1-p, p]$ | $[1-3p, 3p]$ |
| $n_{\min}(x)$ | $\lfloor 2^m x\rfloor - 2p + 2$ | $\lfloor 2^m x\rfloor - 4p + 2$ |
| $n_{\max}(x)$ | $\lceil 2^m x\rceil - 1$ | $\lceil 2^m x\rceil + 2p - 1$ |
| $l_{\min}(x)$ | $\lfloor 2^m x\rfloor - p + 1$ | $\lfloor 2^m x\rfloor - 3p + 1$ |
| $l_{\max}(x)$ | $\lceil 2^m x\rceil + p - 2$ | $\lceil 2^m x\rceil + 3p - 2$ |
| Vanishing moments of $\psi(x)$ | $p$ | $2p$ |

TABLE I

BASIC PROPERTIES OF WAVELET FUNCTIONS ($p$ – WAVELET NUMBER)

## REFERENCES

[1] A. Antoniadis, G. Grégoire, and P. Vial. Random design wavelet curve smoothing. *Statistics and Probability Letters*, 35:225–232, 1997.

[2] C. K. Chui. *An Introduction to Wavelets (Wavelet Analysis and Its Applications)*, volume 1. Academic Press, San Diego, CA, January 1992.

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

[4] B. Delyon and A. Juditsky. Estimating wavelet coefficients. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statistics*, pages 151–168. Springer-Verlag, New York, 1995.

[5] W. Greblicki and M. Pawlak. Fourier and Hermite series estimates of regression function. *Annals of The Institute of Statistical Mathematics*, 37:443–455, 1985.

[6] W. Greblicki and M. Pawlak. Hammerstein system identification by non-parametric regression estimation. *International Journal of Control*, 45:343–354, 1987.

[7] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.

[8] W. Härdle. *Applied Nonparametric Regression*. Cambridge University Press, Cambridge, 1990.

[9] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, Approximation, and Statistical Applications*. Springer-Verlag, New York, 1998.

[10] Z. Hasiewicz. Non-parametric estimation of non-linearity in a cascade time series system by multiscale approximation. *Signal Processing*, 81:791–807, 2001.

[11] Z. Hasiewicz, M. Pawlak, and P. Śliwiński. Non-parametric identification of non-linearities in block-oriented complex systems by orthogonal wavelets with compact support. Submitted for publication to *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 2003.

[12] Z. Hasiewicz and P. Śliwiński. Identification of non-linear characteristics of a class of block-oriented non-linear systems via Daubechies wavelet-based models. *International Journal of Systems Science*, 14:1121–1144, 2002.

[13] A. Kovacz and B. W. Silverman. Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *Journal of the American Statistical Association*, 95:172–183, 2000.

[14] S. G. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, 1998.

[15] M. Pawlak and Z. Hasiewicz. Nonlinear system identification by the Haar multiresolution analysis. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 45:945–961, 1998.

[16] M. Pawlak and Z. Hasiewicz. Non-parametric identification of multi-channel systems by multiscale expansions. In *Proceedings of International Conference on Acoustic Speech and Signal Processing ICASSP 2002*, pages 1721–1724, Orlando, Florida, 2002.

[17] M. B. Ruskai, G. Beylin, R. Coifman, I. Daubechies, S. Mallat, Y. Meyer, and L. Rafael. *Wavelets and Their Applications*. Jones and Barlett, Boston, 1992.

[18] G. Strang. Wavelet transforms vs Fourier transforms. *Bulletin of American Mathematical Society*, 28:288–305, 1993.