Int. J. Appl. Math. Comput. Sci., 2013, Vol. 23, No. 3, 507–520
DOI: 10.2478/amcs-2013-0039

A SIMPLE SCHEME FOR SEMI–RECURSIVE IDENTIFICATION OF HAMMERSTEIN SYSTEM NONLINEARITY BY HAAR WAVELETS

PRZEMYSŁAW ŚLIWIŃSKI, ZYGMUNT HASIEWICZ, PAWEŁ WACHEL

Institute of Computer Engineering, Control and Robotics
Wrocław University of Technology, Wybrzeże Wyspiańskiego 27, Wrocław, Poland
e-mail: przemyslaw.sliwinski@pwr.wroc.pl

A simple semi-recursive routine for nonlinearity recovery in Hammerstein systems is proposed. The identification scheme is based on the Haar wavelet kernel and possesses a simple and compact form. The convergence of the algorithm is established and the asymptotic rate of convergence (independent of the input density smoothness) is shown for piecewise-Lipschitz nonlinearities. The numerical stability of the algorithm is verified. Simulation experiments for a small and moderate number of input-output data are presented and discussed to illustrate the applicability of the routine.

Keywords: Hammerstein system, non-parametric recursive identification, Haar orthogonal expansion, convergence analysis, numerical stability.

1. Introduction

A majority of natural phenomena, objects, or man-made systems have dynamic and nonlinear nature. Discovering this nature is an important and interesting yet difficult scientific problem, particularly when the prior knowledge is poor. Usually, two (very often opposite) requirements are needed to be jointly satisfied:

- a universal character of the approach, which allows finding the best (or the genuine) description of the system at hand, and
- the simplicity of the identification algorithms (and the resulting models), which makes them realizable in practice.

In the paper we apply these generic guidelines to the problem of a nonlinearity recovery in Hammerstein systems working in a stochastic environment. The Hammerstein system (Fig. 1(a)) is a cascade connection of a memoryless subsystem with a nonlinear characteristic and a linear dynamic one. Due to its simplicity, it is a popular nonlinear system modeling tool and has already found applications in various areas like, e.g., automatic control, signal processing, economy and biomedical engineering, (cf., e.g., Chen *et al.*, 1989; Coca and Billings, 2001; Capobianco, 2002; Jyothi and Chidambaram, 2000; Lortie and Kearney, 2001; Westwick and Kearney, 2003; Marmarelis, 2004; Zhou and DeBrunner, 2007; Kukreja *et al.*, 2005; Nordsjo and Zetterberg, 2001; Clancy *et al.*, 2012). Moreover, a number of systems encountered in applications, e.g., the multibranched, Uryson and MISO systems can be reduced to the equivalent canonical Hammerstein structure. Our goal is to recover a non-linear part of such systems. The rationale is twofold (cf. Hasiewicz *et al.*, 2005; Greblicki and Pawlak, 2008, Ch. 2.3):

- The linear subsystem can be recovered in a separate routine, independent of the nonlinear static one.
- The problem of linear subsystem recovery appears to be much simpler.

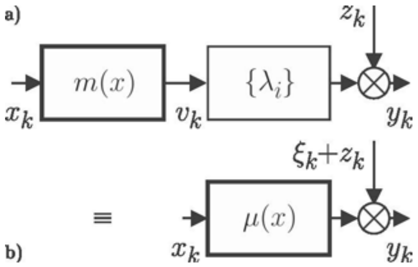

Fig. 1. Hammerstein system (a), equivalent identified memoryless nonlinear system (b).

To obtain an algorithm recovering a nonlinearity of virtually any shape, we use a nonparametric approach;

see the work of Greblicki and Pawlak (2008) for its comprehensive presentation. In this approach the measurement data are the only information about the nonlinearity and, in random environments, the number of measurements has to be large. We construct a convenient *semi-recursive* identification algorithm based on Haar wavelet functions (the simplest member of the popular family of Daubechies functions (see Mallat, 1998). It can process large amounts of data without an excessive computational overhead (cf., e.g., Skubalska-Rafajłowicz, 2001; Rutkowski, 2004; Saeedi *et al.*, 2011). We also examine the convergence conditions and the asymptotic convergence rate of the algorithm.

This locates our study in the framework of papers such as those by Greblicki and Pawlak (1987) or Krzyżak (1986; 1993), where the non-wavelet kernel recursive identification algorithms were investigated, and the ones by Greblicki (2004) or Chen (2004; 2010), where identification algorithms based on stochastic approximation were proposed and analyzed, and eventually those by Pawlak and Hasiewicz (1998) or Hasiewicz (1999; 2000), where the *batch* (non-recursive) Haar wavelet estimation algorithm was discussed. The originality and main advantages of our proposition can be summarized as follows:

1. In contrast to previously examined quotient-form wavelet algorithms (see, e.g., Hasiewicz, 1999; Hasiewicz and Śliwiński, 2002; Hasiewicz *et al.*, 2005),

$$\hat{\mu}_{K(k)}(x) = \frac{\sum_{l=1}^{k} y_l \cdot \phi_{K(k)}(x, x_l)}{\sum_{l=1}^{k} \phi_{K(k)}(x, x_l)}, \tag{1}$$

   (here $\phi_{K(k)}(x, v)$ is a wavelet kernel and $K(k)$ is the scale adjusted to the overall measurement set size k), which could have potentially been ill-posed or could have exploded for the denominator being close to zero, in the proposed Haar wavelet based approach such a menace does not exist as the scheme the possesses numerical stability property, i.e., for bounded input data it produces a bounded estimate (see Lemma 1 in Section 4).

2. The algorithm is straightforward and easy to implement. Due to the basic form of the Haar wavelet kernel, it is also computationally much simpler than other recursive orthogonal series kernel estimation algorithms (Greblicki and Pawlak, 1989; Krzyżak, 1993; Györfi *et al.*, 2002, Ch. 24).

3. The range of applicability of the algorithm is rather wide. It can be successfully used to recover virtually any nonlinearity in any stable Hammerstein system driven by a random signal having almost any probability density functions (cf. also Greblicki and Pawlak, 1987; Vörös, 2003).

4. For a class of piecewise-Lipschitz nonlinear characteristics, the asymptotic efficiency of the procedure cannot be outperformed by any other routine, since its asymptotic convergence rate is optimal (i.e., the best possible in the sense of Stone (1980)). Furthermore, this rate is independent of the input density smoothness.

There are some weaker points of the presented procedure: the estimates are computed separately for each of the *a priori* chosen estimation points[1] and the resulting estimate is discontinuous (piecewise-constant). Nevertheless, one can consider interpolation as a simple remedy to these deficiencies. It yields a global and continuous model of the nonlinearity and can be easily refined with the successively incoming measurement data; see Remark 8 (in the numerical experiments in Section 5.1 we demonstrate potential advantages of an interpolation scheme (see also, e.g., Pawlak *et al.*, 2003; Śliwiński, 2013).

2. Problem statement

Our task is to recover a characteristic of the nonlinear memoryless part of a Hammerstein system from the pairs of successively incoming measurements of the system input and output, (x_k, y_k), $k = 0, 1, \ldots$, in a *recursive* fashion, i.e., without the necessity of memorizing the measurement data. Similarly as, e.g., Greblicki and Pawlak (1989), Krzyżak (1993), Pawlak and Hasiewicz (1998), Hasiewicz (1999; 2000), Greblicki (2002) or Śliwiński (2010), we assume the following:

A1. The input signal x_k is a second order random *i.i.d.* sequence possessing a probability density function, say $f(x)$, which is bounded away from zero in the identification region.

A2. The unknown nonlinearity, $m(x)$, is an arbitrary function such that $|m(x)| \leq c_0 + c_1 |x|$ for some $c_0, c_1 > 0$.

A3. The linear dynamic part, with an impulse response $\{\lambda_i\}$, $i = 0, 1, \ldots$, is asymptotically stable, i.e.,

$$\sum_{i=0}^{\infty} |\lambda_i| < \infty, \tag{2}$$

and $\lambda_0 \neq 0$.

A4. The external output noise, z_k, is a zero-mean second order stationary process, i.e.,

$$Ez_1 = 0, \quad \operatorname{var} z_1 < \infty,$$

[1] Such local and pointwise nature of the estimate is typical for all kernel nonparametric algorithms and can be directly attributed to the lack of prior knowledge about the shape of the estimated characteristic.

with an *arbitrary correlation structure*. The signals z_k and x_k are mutually independent.

A5. The interconnecting signal v_k is not available for measurements.

A6. The input-output measurements (x_k, y_k) are not stored in a memory.

The above assumptions (typical for nonparametric identification tasks) are *qualitative* in nature. The underlying system cannot be therefore described by a parametric equation of the known form. Moreover, the assumption A1 does not impose any restriction on the smoothness of the input signal density function. The requirement that $f(x)$ be locally bounded away from zero follows from a rather obvious observation that the recovery of a nonlinearity can in general be performed only in these regions where the measurements can occur. It does not preclude that the input density vanishes elsewhere.

It is well known that, if $E|m(x)|$ is finite, then (see, e.g., Greblicki and Pawlak, 1986)

$$E\{y_k \,|x_k = x\} = \lambda_0 m(x) + \beta, \tag{3}$$

where the shift term, $\beta = Em(x_1)\sum_{i=1}^{\infty}\lambda_i$, is a system-dependent constant. Using only the input-output data (x_k, y_k) we can retrieve $\mu(x) = \lambda_0 m(x) + \beta$, the scaled and shifted version of the true nonlinear characteristic $m(x)$. That is, from the algorithmic point of view, we recover a nonlinear characteristic $\mu(x)$ of an equivalent fictitious memoryless system shown in Fig. 1(b), disturbed by the combination of the external output noise z_k and the 'system noise' $\xi_k = \sum_{i=1}^{\infty}\lambda_i\zeta_{k-i}$ (where $\zeta_k = m(x_k) - Em(x_1)$). Note that the latter depends on the input signal x_k and is correlated due to the system dynamics.

Remark 1. The condition in the assumption A3, stating that there is no delay in the system, is imposed only to make the presentation simpler. If there is a d-step delay in the system, i.e., we have $\lambda_\iota = 0$ for all $\iota < d$, one can take in (3) any other $\lambda_\iota \neq 0$ for $\iota = d+1, \ldots$, and then use the data pairs $(x_{k-\iota}, y_k)$ in place of (x_k, y_k) in the identification routine.

3. Identification algorithm

Let $I(x)$ be the indicator function of the unit interval $[0, 1]$. Let us define the function $\phi_{K(k)}(x, u) = \phi(2^{K(k)}x, 2^{K(k)}u)$, where $\phi(x, u) = I(x - \lfloor u \rfloor)$. This function is equal to the kernel of the Haar wavelet series $2^{K(k)}\phi\left(2^{K(k)}x, 2^{K(k)}u\right)$, up to the scaling factor $2^{K(k)}$, and will further be called the kernel (Walter and Shen, 2001, Ch. 3).

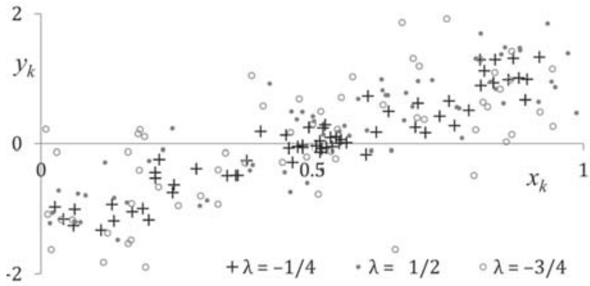

Fig. 2. Visualization of the identification problem. The only available information is carried by the heavily noised input-output measurements $\{(x_k, y_k)\}$ of the Hammerstein system (here with a quantizer-like nonlinearity as in (21) and with dynamics described by various impulse responses $\{\lambda^i\}$; monotonic with $\lambda = 1/2$, or oscillating with $\lambda = -1/4, -3/4$; cf. Fig. 5 in Section 5.1).

Algorithm 1.

At each estimation point x, and for each measurement pair (x_k, y_k), $k = 1, 2, \ldots$, the nonlinearity $\mu(x)$ is estimated by the following semi-recursive[2] formula:

$$\hat{\mu}_k(x) = \hat{\mu}_{k-1}(x) + \gamma_k(x, x_k) \times [y_k - \hat{\mu}_{k-1}(x)]\,\phi_{K(k)}(x, x_k), \tag{4}$$

where $\gamma_k(x, x_k) = 1/\kappa_k(x, x_k)$ is a weighting factor with a denominator updated recursively in a separate subroutine:

$$\kappa_k(x, x_k) = \kappa_{k-1}(x, x_{k-1}) + \phi_{K(k)}(x, x_k). \tag{5}$$

At each point x, the algorithm starts from the natural initial conditions[3]

$$\hat{\mu}_0(x) = 0, \quad K(0) = 0, \quad \kappa_0(x, x_0) = 0.$$

The scale factors $K(k)$, $k = 1, 2, \ldots$, form an increasing sequence and its proper selection rule with respect to k will be discussed in detail further on. Certainly, because of the recursive character of the routine (4)–(5), for each estimation point x only the previously estimated value $\hat{\mu}_{k-1}(x)$ and the number $\kappa_{k-1}(x, x_{k-1})$ need to be stored.

The form of the proposed routine resembles both the stochastic approximation algorithms (see, e.g., Kushner and Yin, 2003) and the classic two-step recursive

[2] The proposed routine is referred to as a semi-recursive algorithm to emphasize its two-step update procedure (4) and (5), and to distinguish it from the recursive stochastic approximation-based algorithms; (cf., e.g., Györfi *et al.*, 2002; Greblicki and Pawlak, 2008).

[3] The estimated values $\hat{\mu}_k(x)$ are kept zero for each estimation point x until the weighting factor $\gamma_k(x, x_k)$ corresponding to x is undefined, i.e., until the first measurement pair (x_k, y_k) with x_k inside the support of the kernel $\phi_{K(k)}(x, x_k)$ appears.

least-squares methods. Moreover, one can easily see that (4) can be written in the following *convex combination* form:

$$\hat{\mu}_k(x) = [1 - \rho_k(x)] \cdot \hat{\mu}_{k-1}(x) + \rho_k(x) \cdot y_k, \tag{6}$$

with $\rho_k(x) = \phi_{K(k)}(x, x_k) / \kappa_k(x, x_k) \in [0, 1)$. This links our procedure with the standard recursive formula for computing an empirical mean $\hat{x}_k = (1/k)\sum_{l=1}^{k} x_l$, i.e.,

$$\hat{x}_k = (1 - \rho_k) \cdot \hat{x}_{k-1} + \rho_k \cdot x_k \tag{7}$$

where $\rho_k = 1/k$.

Remark 2. One can find yet another link between the routine considered and the least-squares approach by recalling that the batch-form estimate in (1) is in fact the solution to the optimization problem

$$\hat{\mu}_{K(k)}(x) = \arg\min_{c \in \mathbb{R}} \Phi(c; x),$$

where

$$\Phi(c; x) = \sum_{l=1}^{k} [y_l - c]^2 \phi_{K(k)}(x, x_l).$$

From the formulas in (6) and (7), one can derive an intuitive interpretation of Algorithm 1 for each estimation point x, the associated kernel function $\phi_{K(k)}(x, u)$ 'picks' only these new measurement pairs (x_k, y_k) in which inputs x_k have fallen into a proper neighborhood of this point. The term $\kappa_k(x, x_k)$ in (5) acts as a counter of these measurement pairs and the estimated value $\hat{\mu}_k(x)$ is just a locally weighted empirical mean of the corresponding outputs y_k. With the growing number of measurements k (and the subsequently growing scale factor $K(k)$; see Section 4), the kernel support is successively contracted

$$|\operatorname{supp}\phi_{K(k)}(x, u)| = 2^{-K(k)}, \tag{8}$$

and the estimate (4) takes into account the measurements from the narrowing neighborhood of x, thus becoming more and more localized around the estimation points.

Remark 3. In Algorithm 1, the explicit constant-form and simple kernel function of Haar wavelet series is used. We emphasize this feature because this computationally desired property of a kernel function is unique amongst the Daubechies wavelet family (which Haar wavelets belong to) and not shared by other orthogonal series counterparts either. For instance, the Dirichlet kernel (i.e., the kernel of the Fourier trigonometric series) and the kernels of polynomial series are of different (and rather complicated) forms for different scale factors K (cf. Sansone, 1959; Szego, 1974; Greblicki and Pawlak, 2008).

4. Convergence of the algorithm

Algorithm 1 does not require the bulk of measurements to be stored in memory and hence is much more convenient in use than its off-line counterpart (1). Nevertheless, if properly tuned, it still maintains the asymptotic properties of the latter. Our first theorem characterizes its convergence conditions.

Theorem 1. *Let the assumptions A1–A4 be in force. If the scale factor $K(k)$ successively grows with the growing number of processed data k in such a way that*

$$K(k) \to \infty \tag{9a}$$

and

$$\sum_{l=1}^{k} 2^{-K(l)} \to \infty, \tag{9b}$$

as $k \to \infty$, then the estimate converges to the identified nonlinearity,

$$\hat{\mu}_k(x) \to \mu(x) \quad \text{as } k \to \infty \text{ in probability,}$$

at all continuity points of $\mu(x)$.

Proof. As it follows directly from the way of obtaining (4)–(5) described in Section 3, and from derivation of the identification scheme in Appendix A, the recursive estimate (4) can be written in the equivalent batch form:[4]

$$\hat{\mu}_k(x) = \frac{\sum_{l=1}^{k} y_l \cdot \phi_{K(l)}(x, x_l)}{\sum_{l=1}^{k} \phi_{K(l)}(x, x_l)}. \tag{10}$$

We shall further use the following, equivalent to (10) and hence to (4), form of the estimate (this idea is borrowed from Greblicki and Pawlak (1987)):

$$\hat{\mu}_k(x) = \frac{\hat{\vartheta}_k(x)}{\hat{\eta}_k(x)} = \frac{\dfrac{1}{\sum_{l=1}^{k} E\phi_{K(l)}(x, x_l)} \sum_{l=1}^{k} y_l \phi_{K(l)}(x, x_l)}{\dfrac{1}{\sum_{l=1}^{k} E\phi_{K(l)}(x, x_l)} \sum_{l=1}^{k} \phi_{K(l)}(x, x_l)}. \tag{11}$$

[4] Note that the estimation formula (10) significantly differs from the *batch* version (1) (worked out and discussed earlier by, e.g., Pawlak and Hasiewicz (1998), Hasiewicz (1999) or Hasiewicz and Śliwiński (2002)) in the sense that, in the former, the scale $K = K(l)$ is not determined *a priori* but gradually adapts to the current number of processed data $\{(x_l, y_l)\}$, $l = 1, \ldots, k, \ldots$, while in the latter it is fixed and selected as $K = K(k)$ either *a posteriori* just as the whole set of data $\{(x_l, y_l)\}$, $l = 1 \ldots, k$ is collected, and the data length k is established, or in advance, for the *a priori* assumed length of data which are planned to be collected in an identification experiment. In the latter case, the *batch* estimates suffer from *undersmoothing* (when the actual measurements number is smaller than the designed one), or from *oversmoothing* in the opposite situation).

For notational simplicity we employ the shortened symbols $\phi_l = \phi_{K(l)}(x, x_l)$, $\mu_l = \mu(x_l)$, and

$$\kappa \triangleq \sum_{l=1}^{k} E\phi_{K(l)}(x, x_l) = \sum_{l=1}^{k} E\phi_l.$$

Furthermore, to make the proofs less tedious (in particular, the covariance analysis part in Appendix B), we will consider the case when the noise signal z_k is white (the analysis with the correlated external noise resembles the one performed for the correlated 'system noise' ξ_k; see Appendix B). The proof has two main steps: (i) we show the MSE convergence of the numerator $\hat{\vartheta}_k(x)$ to $\mu(x)$ and of the denominator $\hat{\eta}_k(x)$ to 1, and (ii) we conclude the convergence of the whole quotient $\hat{\mu}_k(x) = \hat{\vartheta}_k(x)/\hat{\eta}_k(x)$ to $\mu(x)$ in probability.

Bias error analysis. Consider the expectation of the numerator $\hat{\vartheta}_k(x)$ in (11). Recalling that $y_l = \mu_l + \xi_l + z_l$, we have

$$E\hat{\vartheta}_k(x) = \frac{\sum_{l=1}^{k} Ey_l\phi_l}{\sum_{l=1}^{k} E\phi_l} = \frac{\sum_{l=1}^{k} E\mu_l\phi_l}{\sum_{l=1}^{k} E\phi_l},$$

since ξ_l and z_l are zero-mean and independent of ϕ_l for a given l. For the bias error, defined here for the numerator $\hat{\vartheta}_k(x)$ as follows, we have that

$$\begin{aligned}\text{bias}\hat{\vartheta}_k(x) &= E\left[\hat{\vartheta}_k(x) - \mu(x)\right] \\ &= \frac{\sum_{l=1}^{k} E\mu_l\phi_l - \sum_{l=1}^{k} E\mu(x)\phi_l}{\sum_{l=1}^{k} E\phi_l} \\ &= \frac{\sum_{l=1}^{k} E\phi_l \cdot \overbrace{E\left\{[\mu_l - \mu(x)]\frac{\phi_l}{E\phi_l}\right\}}^{=b_l}}{\sum_{l=1}^{k} E\phi_l}. \quad (12)\end{aligned}$$

To see that bias $\hat{\vartheta}_k(x)$ vanishes as $K(l) \to \infty$ (cf. (9a)), observe that (cf. (8))

$$b_l = \int_{\text{supp}\phi_{K(l)}} |\mu(x) - \mu(u)| \frac{f(u)}{\int_{\text{supp}\phi_{K(l)}} f(u)\,\mathrm{d}u} du \to 0,$$

as $K(l), l \to \infty$, i.e., that b_l vanishes in all continuity points of $\mu(x)$, i.e., almost everywhere (by virtue of Luzin's theorem) (see, e.g., Wheeden and Zygmund, 1977, Theorem 10.49; Greblicki and Pawlak, 2008, Lemma A.10).

Observe now that

(i) under the assumption A1, the sequence $\{2^{K(l)}E\phi_l\}$ is uniformly lower bounded in l (cf., e.g., Gomes and Cortina, 1995; Wheeden and Zygmund, 1977, Theorem 10.49).

(ii) From the condition in (9b), we obtain

$$\kappa = \sum_{l=1}^{k} E\phi_l = \sum_{l=1}^{k} 2^{-K(l)}\left[2^{K(l)}E\phi_l\right] \to \infty,$$

since the term in square brackets is no less than $\inf f(x) \cdot \sum_{l=1}^{k} 2^{-K(l)}$, where the infimum is taken over all $x \in \text{supp}\,\phi_1$. By virtue of these facts, the bias error (12) vanishes in all continuity points of $\mu(x)$ since, by the Toeplitz lemma (see, e.g., Van der Vaart, 2000; Greblicki and Pawlak, 2008) we have that

$$\frac{\sum_{l=1}^{k} b_l E\phi_l}{\sum_{l=1}^{k} E\phi_l} \to 0 \quad \text{as } k \to \infty.$$

Variance error. Examining the variance of the numerator $\hat{\vartheta}_k(x)$ in (11) we get

$$\begin{aligned}\text{var}\hat{\vartheta}_k(x) &= \text{var}\left\{\frac{\sum_{l=1}^{k} y_l\phi_{K(l)}(x, x_l)}{\kappa}\right\} \\ &= \frac{1}{\kappa^2}\text{var}\left\{\sum_{l=1}^{k} y_l\phi_l\right\}, \quad (13)\end{aligned}$$

where

$$\begin{aligned}\text{var}\left\{\sum_{l=1}^{k} y_l\phi_l\right\} &= \sum_{l=1}^{k} \text{var}\{y_l\phi_l\} \\ &\quad + \sum_{i=1}^{k}\sum_{\substack{j=1\\ i\neq j}}^{k} \text{cov}\{y_i\phi_i, y_j\phi_j\} \\ &= \sum_{l=1}^{k} \text{var}\{y_l\phi_l\} \\ &\quad + 2\sum_{i=1}^{k}\sum_{j=i+1}^{k} \text{cov}\{y_i\phi_i, y_j\phi_j\}. \end{aligned} \quad (14)$$

After rather cumbersome derivations (see Appendix B) we conclude that

$$\text{var}\hat{\vartheta}_k(x) \le \frac{c_{\text{var}} + c_{\text{cov}}}{\kappa^2}\sum_{l=1}^{k} E\phi_l = \frac{c_{\text{var}} + c_{\text{cov}}}{\kappa},$$

where c_{var} and c_{cov} are some positive constants (cf. (B1) and (B3) in Appendix B). This leads to the final ascertainment that there exists some $c_\vartheta = c_{\text{var}} + c_{\text{cov}} > 0$, such that

$$\text{var}\hat{\vartheta}_k(x) \le c_\vartheta\kappa^{-1}.$$

Recalling now that $\phi_l = \phi_{K(l)}(x, x_l)$ and that in all continuity points of $\mu(x)$ the quantity $2^{K(l)}E\phi_l$ is uniformly lower bounded in l, we realize that if the sequence $K(l)$ satisfies the condition in (9b), then the variance of the estimate numerator $\mathrm{var}\hat{\vartheta}_k$ vanishes as k grows, since we have

$$\kappa = \sum_{l=1}^{k} 2^{-K(l)} \cdot \left[2^{K(l)} E\phi_l\right] \to \infty \quad \text{as } k \to \infty$$

at all these points x. The above mean-square error analysis can immediately be repeated for $\hat{\eta}_k(x)$, the denominator in (11), since $\hat{\eta}_k(x) = \hat{\vartheta}_k(x)$ for $y_l \equiv 1, l = 1, \ldots, k$. In particular, it is straightforward to observe that $\hat{\eta}_k(x)$ is an unbiased estimate of unity, i.e.,

$$E\hat{\eta}_k(x) = E\left\{\frac{\sum_{l=1}^{k} \phi_{K(l)}(x, x_l)}{\sum_{l=1}^{k} E\phi_{K(l)}(x, x_l)}\right\} = 1$$

for all k while for the variance, because of the whiteness of the inputs, we have that

$$\mathrm{var}\hat{\eta}_k(x) \le \frac{\sum_{l=1}^{k} E\phi_l^2}{\kappa^2} = \frac{\sum_{l=1}^{k} E\phi_l}{\kappa^2} = \kappa^{-1}.$$

Since both bias and variance errors of $\hat{\vartheta}_k(x)$ and $\hat{\eta}_k(x)$ vanish almost everywhere, by applying the Slutsky theorem (cf., e.g., Serfling, 1980), we see that the quotient $\hat{\mu}_k(x) = \hat{\vartheta}_k(x)/\hat{\eta}_k(x)$ converges to $\mu(x)$ in probability. ■

We have thus shown that the estimate $\hat{\mu}_k(x)$ converges to $\mu(x)$ almost everywhere and that the convergence holds independently of the shape of the input probability density function (provided that the assumption 1 is fulfilled) and of the particular system dynamics as well as correlation structure of the external output noise.

As an example of the scale factor $K(k)$ satisfying (9a)–(9b) one can take $\lfloor \alpha \log_2 k \rfloor$ with any $0 < \alpha < 1$.

Remark 4. The convergence condition in (9a) can be replaced by the following one (cf. (12)):

$$\frac{\sum_{l=1}^{k} 2^{-K(l)} I_{(K(l)<\delta)}}{\sum_{l=1}^{k} 2^{-K(l)}} \to 0, \tag{15}$$

as $k \to \infty$ for any $\delta > 0$, which is weaker than ours and admits, for instance, non-monotonic sequences, of the form (see, e.g., Greblicki and Pawlak, 1987; Krzyżak, 1993, Remark 2)

$$K(k) = \begin{cases} 0 & \text{if } k \text{ is a dyadic integer,} \\ \lfloor \alpha \log_2 k \rfloor & \text{otherwise.} \end{cases}$$

Remark 5. In the paper we are focused on the *in probability* convergence properties of the algorithm. It is, however, interesting to note that, for a memoryless system and white output noise, the conditions (9a) and (15) are sufficient (and necessary) for the estimate $\hat{\mu}_k(x)$ to converge not only in probability, but also with probability 1 (Greblicki and Pawlak, 1987; Krzyżak and Pawlak, 1984; Rutkowski, 1984). In turn, if (in the case of a Hammerstein system) the output signal is a stationary process of order $s > 2$, then (as claimed by Krzyżak, 1993, Theorem VI) the algorithm converges with probability 1 to the nonlinearity $\mu(x)$ almost everywhere if the condition in (15) and the following one:

$$\frac{k^{-\frac{s+2}{2s}}}{\sqrt{\log k}} \sum_{l=1}^{k} 2^{-K(l)} \to \infty \quad \text{as } k \to \infty \tag{16}$$

hold true; see the works of Krzyżak (1992; 1993) for technical details. Note that the condition (16) is more stringent than (9b): an example of a scale factor sequence $K(k)$ satisfying (15)–(16) is $\lfloor \alpha \log_2 k \rfloor$ with $0 < \alpha < (s-2)/2s$.

4.1. Convergence rate.

Assume now that the nonlinearity $\mu(x)$ is piecewise-Lipschitz, that is, it has an unknown (but finite) number of step discontinuities (jumps) and is Lipschitz continuous between them, i.e.,

$$|\mu(x) - \mu(v)| \le c_m |x - v| > 0 \tag{17}$$

for some c_m. The asymptotic rate of convergence of the estimate $\hat{\mu}_k(x)$ to such nonlinearities is the subject of the next theorem.

Theorem 2. *Let (17) hold together with the assumptions A1, A3–A4. If*

$$K(k) = \left\lfloor \tfrac{1}{3} \log_2 k \right\rfloor, \tag{18}$$

then, for $k \to \infty$, the estimate converges with the rate

$$|\hat{\mu}_k(x) - \mu(x)| = \mathcal{O}(k^{-\frac{1}{3}}) \quad \textit{in probability,}$$

at all continuity points of $\mu(x)$.

Proof. After a suitable reasoning (see Appendix C), we obtain that

$$\mathrm{MSE}\hat{\vartheta}_k(x) = \mathrm{bias}^2\hat{\vartheta}_k(x) + \mathrm{var}\hat{\vartheta}_k(x) \le c_{\mathrm{MSE}} k^{-\frac{2}{3}} \tag{19}$$

for some $c_{\mathrm{MSE}} > 0$. Similarly, for the denominator $\hat{\eta}_k(x)$, we get

$$\mathrm{MSE}\hat{\eta}_k(x) = \mathrm{var}\hat{\eta}_k(x) \le c'_{\mathrm{MSE}} k^{-\frac{2}{3}}, \tag{20}$$

some $c'_{\mathrm{MSE}} > 0$. Based on (19) and (20), and employing Lemma C.8 of Greblicki and Pawlak (2008), we conclude the proof. ■

The theorem demonstrates that for piecewise-Lipschitz nonlinearities the estimate $\hat{\mu}_k(x)$ with the scale selection rule (18) attains the best possible convergence rate in the framework of nonparametric inference (cf. Stone, 1980) and is robust to the smoothness of the input probability density function, the character of the system dynamics, and the external output noise. In particular, the convergence rate is preserved at points where the density function $f(x)$ is discontinuous. It should be, however, noted, that—because of the limited approximation properties of the Haar functions—the rate will not be faster when the nonlinearity is smoother than Lipschitz, e.g., when it has $p = 1, 2, \ldots$ continuous derivatives.

Remark 6. The asymptotic rate $\mathcal{O}(k^{-1/3})$ is the same as for the Haar batch identification algorithm (1) (cf. Pawlak and Hasiewicz, 1998) in spite of the fact that the recursive version has no immediate access to the whole data set but only to the consecutively incoming single measurements (x_k, y_k), and that the scale factor $K(k)$ is not kept fixed but changes (step-wise) with increasing k.

Remark 7. For any scale selection rule other than (18), i.e., for any $\alpha \neq 1/3$, the resulting asymptotic convergence rate will be slower than $\mathcal{O}(k^{-1/3})$. For example, taking $\alpha = 1/2$ or $\alpha = 1/4$ would yield the rate $\mathcal{O}(k^{-1/2})$.

5. Numerical properties

The quotient form of the equivalent representation (10) and the randomness of its denominator may put in question the estimate numerical stability, especially when the number of the processed measurements is small. The following lemma states that $\hat{\mu}_k(x)$ in (4) is bounded provided that the measurements (x_k, y_k) are bounded, too.

Lemma 1. *Assume that (x_k, y_k) are bounded. Then the estimate $\hat{\mu}_k(x)$ is also bounded for any $k = 1, 2, \ldots$.*

Proof. To verify the boundedness of the estimate, $\hat{\mu}_k(x)$, it suffices to observe that, since $\phi_{K(k)}(x, x_k)$ is non-negative for any k, then

$$|\hat{\mu}_k(x)| = \frac{\left|\sum_{l=1}^{k} y_l \cdot \phi_{K(l)}(x, x_l)\right|}{\left|\sum_{l=1}^{k} \phi_{K(l)}(x, x_l)\right|} \leq \frac{\sum_{l=1}^{k} |y_l| \cdot \phi_{K(l)}(x, x_l)}{\sum_{l=1}^{k} \phi_{K(l)}(x, x_l)} \leq \max_{l=1,\ldots,k} |y_l| .$$

■

In view of the assumptions A2–A3, the requirement that $\max_{l=1,\ldots,k} |y_l| < \infty$ is fulfilled when, along with bounded x_k, also the external noise z_k is bounded.

5.1. Numerical experiments.

To illustrate the properties of the recursive Algorithm 1(4) for small or moderate (viz. $k = 1, \ldots, 512$) numbers of data, and compare it with its batch prototype (1), some numerical tests were performed. Two nonlinear characteristics $m(x)$ of the Hammerstein system, i.e., the piecewise-polynomial with a jump at the point $x = 0.5$, and the piecewise-constant quantizer-like one:

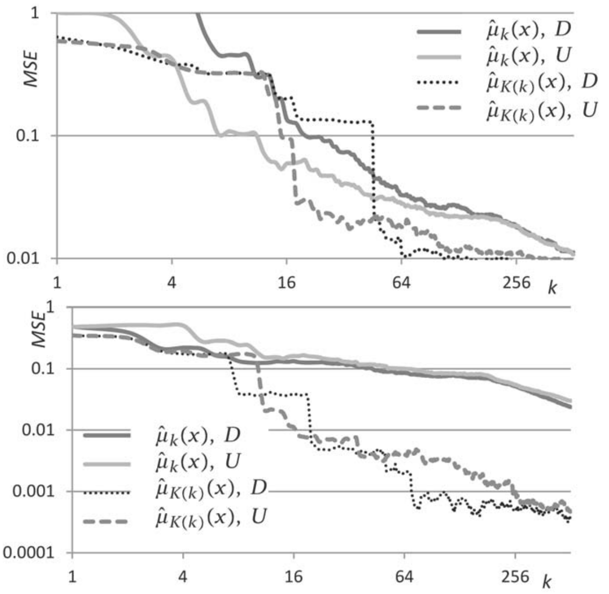

Fig. 3. *Semi-recursive*, $\hat{\mu}_k(x)$, and *batch*, $\hat{\mu}_{K(k)}(x)$, estimates of the piecewise-polynomial (left) and piecewise-constant (right) nonlinearities for $\lambda = -1/4$; 'U' and 'D' stand for the uniform and discontinuous input density functions, respectively.

$$m(x) = -5\left(2x^3 - 3x^2 + x\right) - \frac{1}{2}\operatorname{sgn}\left(x - \frac{1}{2}\right) \times \frac{1}{2}\left\lfloor 2(2x-1) + \frac{1}{2}\right\rfloor \tag{21}$$

were estimated in the unit interval $[0, 1]$. The scale selection rule (18) was employed. The inputs x_k were drawn from that interval assuming either a uniform (i.e., smooth), $f(x) = I_{[0,1)}(x)$, or a piecewise-constant (i.e., discontinuous) input density,

$$f(x) = \sum_{i=0}^{8} f_i I_{[0,1/9)}(x - i/9),$$

$f_i \in \{.08, .14, .08, .1, .2, .1, .08, .14, .08\}$. The dynamics with infinite length impulse responses, $\{\lambda_i = \lambda^i, i = 0, 1, \ldots\}$ for $\lambda = -1/4$ or $-3/4$, were used to model systems with small and large damped oscillations, respectively. The external zero-mean white noise, z_k[5], was uniformly distributed and set to give $\max_k |z_k| / \max_x |m(x)| = 10\%$. A numerically evaluated MSE, averaged over 128 equidistant estimation points, was an empirical measure of the pointwise quality of the

[5] Note that, in such a setting, the identity $\mu(x) = m(x)$ holds in the experiments.

algorithms. The value of the MSE was computed for an increasing number of data points $k = 1, \ldots, 512$, and, purposefully, only for one random data sequence (i.e., with no typical averaging of the experiment results over a number of independent runs) to mimic the realistic conditions where only one particular data set is processed by the recursive identification procedure. Moreover, the cubic spline interpolation based on the interpolation knots $\{(x_i, \hat{\mu}_k(x_i))\}_{i=0}^{7}$, where $x_i = i/8 + 1/16$ and $k = 512$ was performed to assess the capabilities of the interpolation scheme.

The results, presented in Figs. 3–5, confirm several advantageous properties of the routine (4) established formally in previous sections. Namely, one can observe the following:

1. the (almost monotonous for less oscillating dynamics) decrease in the estimation errors with the growing number k of measurements, which confirms the convergence and to the established stability of the algorithm (Section 4, Theorem 1 and Lemma 1);

2. the robustness of the estimate behavior to the smoothness of the input probability density function supporting the 'density-free' convergence rate shown in Section 4 (Theorem 2);

3. comparable performances of the proposed recursive and the earlier batch algorithms. Note, however, a larger bias error for the *recursive* algorithm in the case of the piecewise-continuous nonlinearity, and a larger variance error ('wiggles' of the error plot) of the *batch* algorithm when the nonlinearity is polynomial. They seem to be the obvious consequences of the adaptive way the scale K is selected in the recursive identification algorithm.

In turn, Figs. 5(a)–(b) confirm rather good quality of the produced estimates and reveal that incorporating cubic spline interpolation can actually be a useful **tool** for modelling unknown nonlinearities when their true values can be recovered (estimated) at only a finite and rather small number of points, particularly when the nonlinearity is supposed to be a (piecewise-)smooth function.

Remark 8. The location and number of estimation points is arbitrary. Nevertheless, having in mind a prospective application of an interpolation scheme, the most (computationally) convenient solution is to put them on an equidistant grid (see also the work of Śliwiński (2013) for an alternative, density-adaptive, approach). A number of estimation points

$$q(k) = 2^{K(k)},$$

for a given number of measurements k, can, e.g., be derived from the scale selection rule $K(k)$ in (18) and from the length of the scaled Haar kernel support in (8). Note further that the scale factors $K(k) = 0, 1, 2, \ldots$ form an integer number sequence. The number of new estimation points $q(x)$ is thus doubled each time the scale $K(k)$ increases, and the new estimation points appear in the midpoints between the old ones. The estimate values at the new points can be zero-initialized or interpolated (see, e.g., Śliwiński, 2010).

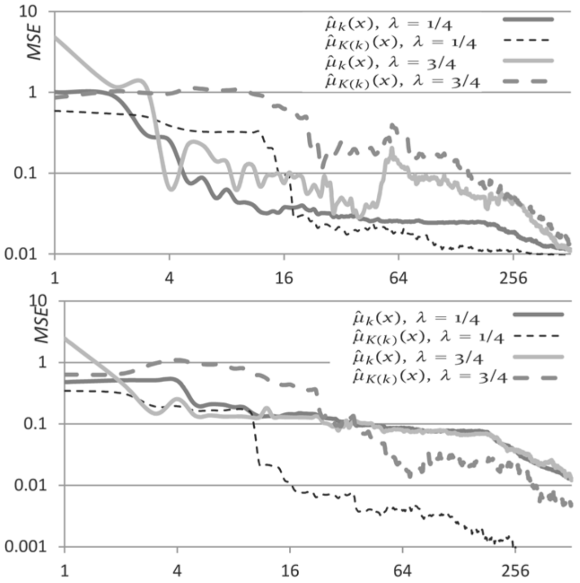

Fig. 4. *Semi-recursive*, $\hat{\mu}_k(x)$, and *batch*, $\hat{\mu}_{K(k)}(x)$, estimates of the piecewise-polynomial (left) and piecewise-constant nonlinearities (right) for a uniformly distributed input signal and for dynamics with $\lambda = -1/4$ or $-3/4$. The larger bias error of the *recursive* estimate for the latter is not as apparent as for the former one.

6. Conclusions

The simple semi-recursive Haar wavelet scheme for estimating the nonlinearity in Hammerstein systems has been proposed and examined. The routine considered has a computationally convenient form and exploits raw data, i.e., it does not need collecting and preprocessing of measurements. Asymptotic analysis of the routine shows its efficiency and a wide range of applicability due to weak, and rather only theoretical in nature, limitations imposed on the input probability density and unknown system characteristic. It is shown that for piecewise-Lipschitz nonlinearities the estimate converges to the target nonlinearity with the optimal convergence rate regardless of the smoothness of the input density function. Furthermore, the limit properties are robust to the correlation structure of the external noise and the structure of the system dynamics. Combination of these asymptotic properties with implementation-relevant numerical stability and computational simplicity makes the presented algorithm an interesting offer in the system identification area, e.g., fault detection (Chen *et al.*, 2011; Patan and Korbicz, 2012).

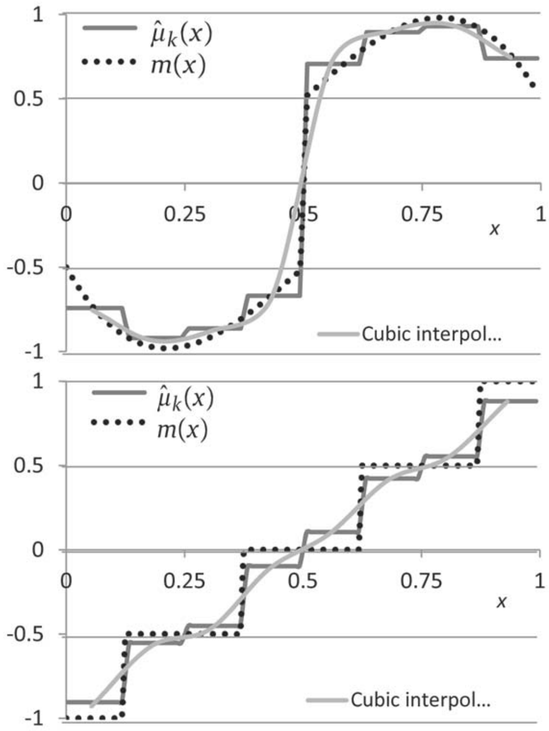

Fig. 5. Estimates of the piecewise-polynomial and piecewise-constant quantizer-type nonlinearities and their cubic spline interpolations built upon the values of the estimates in the equidistant interpolation knots $x_i = i/8 + 1/16$, $i = 0, \ldots, 7$ (evaluated from $k = 512$ measurements for $\lambda = -1/4$).

6.1. Examples of Hammerstein-type structures. As mentioned in Introduction, the Hammerstein system is the simplest instance of Uryson or MISO (i.e., multiple-input single-output) dynamic systems—the multi-branch structures composed of Hammerstein systems connected in parallel; see Fig. 6. In the former, the input signal, x_k, is common for all subsystems, while in the latter, each of the U branches is driven independently.

We shortly examine the problem of nonlinearity recovery in these systems beginning with the Uryson one (cf. Gallman, 1975). Let all the branch subsystems satisfy the assumptions A2–A3, and let $m(x)$ be a nonlinearity of interest. The input-output description equation of the Uryson system has the form

$$y_k = \sum_{i=0}^{\infty} \lambda_i m(x_{k-i}) + \sum_{u=1}^{U}\sum_{i=0}^{\infty} \omega_{u,i}\eta_u(x_{k-i}) + z_k,$$

and the regression function of the system output on the system input is a weighted sum of scaled nonlinearities from all the system branches (shifted by a constant factor β), cf. (3),

$$\begin{aligned} E\{y_k \,|x_k = x\} &= \lambda_0 E\{m(x_k)\,|x_k = x\} \\ &+ \sum_{u=1}^{U} \omega_{u,0} E\{\eta_u(x_k)\,|x_k = x\} \\ &+ \underbrace{\sum_{i=1}^{\infty} \lambda_{u,i} Em(x_{k-i}) + \sum_{u=1}^{U}\sum_{i=1}^{\infty} \omega_{u,i} E\eta_u(x_{k-i})}_{=\beta} \\ &= \underbrace{\lambda_0 m(x) + \beta}_{=\mu(x)} + \underbrace{\sum_{u=1}^{U} \omega_{u,0}\eta_u(x)}_{=\mu_u(x)} = \mu_U(x). \end{aligned} \tag{22}$$

Nevertheless, there exist specific situations of practical significance[6], when

$$E\{y_k\,|x_k = x\} = \mu_U(x) = \mu(x)$$

holds and the nonlinearity $\mu(x)$ can be separately estimated (as in the canonical Hammerstein system):

- if $\omega_{u,0} = 0$ for all $u = 1, \ldots, U$, i.e., if all other dynamic subsystems, $\{\omega_{u,i}\}$, have a non-zero delay;
- if $\operatorname{supp}\mu(x) \cap \operatorname{supp}\eta_u(x) = \emptyset$ for all $u = 1, \ldots, U$, i.e., if all the branch nonlinearities are active in the input signal ranges non-overlapping with the active input range of $\mu(x)$.

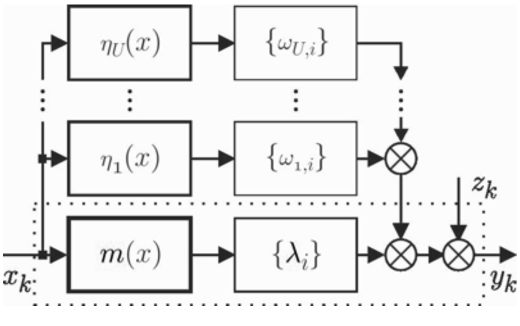

Fig. 6. Uryson system.

Compared with the Uryson system case, the nonlinearity identification conditions in the MISO structure as in Fig. 7 are much less stringent: if the

[6] In the original paper by Gallman (1975), it was assumed that the nonlinearities in each branch were known orthogonal functions (Hermite polynomials).

input signals x_k and $x_{u,k}$, $u = 1, \ldots, U$, are stochastically independent, then (cf. (3) and (22))

$$\begin{aligned} &E\left\{y_k \left| x_k = x\right.\right\} \\ &= \lambda_0 m(x) \\ &\underbrace{+ \sum_{i=1}^{\infty} \lambda_{u,i} E m(x_{k-i}) + \sum_{u=1}^{U} \sum_{i=0}^{\infty} \omega_{u,i} E \eta_u(x_{u,k-i})}_{=\beta} \\ &= \mu(x), \end{aligned}$$

and, by estimating the regression function from the measurement pairs (x_k, y_k), the nonlinearity $\mu(x)$ can be recovered independently of other nonlinearities and properties of the component dynamic subsystems.

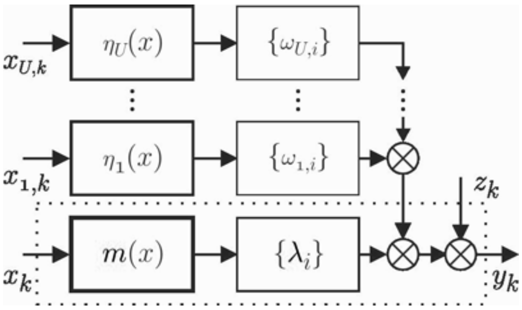

Fig. 7. MISO system.

Acknowledgment

The authors wish to thank the anonymous reviewers for their comments and helpful suggestions.

References

Capobianco, E. (2002). Hammerstein system representation of financial volatility processes, *The European Physical Journal B: Condensed Matter* **27**(2): 201–211.

Chen, H.-F. (2004). Pathwise convergence of recursive identification algorithms for Hammerstein systems, *IEEE Transactions on Automatic Control* **49**(10): 1641–1649.

Chen, H.-F. (2010). Recursive identification for stochastic Hammerstein systems, *in* F. Giri and E.W. Bai (Eds.), *Block-oriented Nonlinear System Identification*, Lecture Notes in Control and Information Sciences, Vol. 404, Springer-Verlag, Berlin/Heidelberg, pp. 69–87.

Chen, S., Billings, S.A. and Luo, W. (1989). Orthogonal least squares methods and their application to non-linear system identification, *International Journal of Control* **50**(5): 1873–1896.

Chen, W., Khan, A.Q., Abid, M. and Ding, S.X. (2011). Integrated design of observer based fault detection for a class of uncertain nonlinear systems, *International Journal of Applied Mathematics and Computer Science* **21**(3): 423–430, DOI: 10.2478/v10006-011-0031-0.

Clancy, E.A., Liu, L., Liu, P. and Moyer, D.V.Z. (2012). Identification of constant-posture EMG-torque relationship about the elbow using nonlinear dynamic models, *IEEE Transactions on Biomedical Engineering* **59**(1): 205–212.

Coca, D. and Billings, S.A. (2001). Non-linear system identification using wavelet multiresolution models, *International Journal of Control* **74**(18): 1718–1736.

Gallman, P. (1975). An iterative method for the identification of nonlinear systems using a Uryson model, *IEEE Transactions on Automatic Control* **20**(6): 771–775.

Gomes, S.M. and Cortina, E. (1995). Some results on the convergence of sampling series based on convolution integrals, *SIAM Journal on Mathematical Analysis* **26**(5): 1386–1402.

Greblicki, W. (2002). Stochastic approximation in nonparametric identification of Hammerstein systems, *IEEE Transactions on Automatic Control* **47**(11): 1800–1810.

Greblicki, W. (2004). Hammerstein system identification with stochastic approximation, *International Journal of Modelling and Simulation* **24**(2): 131–138.

Greblicki, W. and Pawlak, M. (1986). Identification of discrete Hammerstein system using kernel regression estimates, *IEEE Transactions on Automatic Control* **31**(1): 74–77.

Greblicki, W. and Pawlak, M. (1987). Necessary and sufficient consistency conditions for a recursive kernel regression estimate, *Journal of Multivariate Analysis* **23**(1): 67–76.

Greblicki, W. and Pawlak, M. (1989). Recursive nonparametric identification of Hammerstein systems, *Journal of the Franklin Institute* **326**(4): 461–481.

Greblicki, W. and Pawlak, M. (2008). *Nonparametric System Identification*, Cambridge University Press, New York, NY.

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

Hasiewicz, Z. (1999). Hammerstein system identification by the Haar multiresolution approximation, *International Journal of Adaptive Control and Signal Processing* **13**(8): 697–717.

Hasiewicz, Z. (2000). Modular neural networks for non-linearity recovering by the Haar approximation, *Neural Networks* **13**(10): 1107–1133.

Hasiewicz, Z., Pawlak, M. and Śliwiński, P. (2005). Non-parametric identification of non-linearities in block-oriented complex systems by orthogonal wavelets with compact support, *IEEE Transactions on Circuits and Systems I: Regular Papers* **52**(1): 427–442.

Hasiewicz, Z. and Śliwiński, P. (2002). Identification of non-linear characteristics of a class of block-oriented non-linear systems via Daubechies wavelet-based models, *International Journal of Systems Science* **33**(14): 1121–1144.

Jyothi, S.N. and Chidambaram, M. (2000). Identification of Hammerstein model for bioreactors with input multiplicities, *Bioprocess Engineering* **23**(4): 323–326.

Krzyżak, A. (1986). The rates of convergence of kernel regression estimates and classification rules, *IEEE Transactions on Information Theory* **32**(5): 668–679.

Krzyżak, A. (1992). Global convergence of the recursive kernel regression estimates with applications in classification and nonlinear system estimation, *IEEE Transactions on Information Theory* **38**(4): 1323–1338.

Krzyżak, A. (1993). Identification of nonlinear block-oriented systems by the recursive kernel estimate, *Journal of the Franklin Institute* **330**(3): 605–627.

Krzyżak, A. and Pawlak, M. (1984). Distribution-free consistency of a nonparametric kernel regression estimate and classification, *IEEE Transactions on Information Theory* **30**(1): 78–81.

Kukreja, S., Kearney, R. and Galiana, H. (2005). A least-squares parameter estimation algorithm for switched Hammerstein systems with applications to the VOR, *IEEE Transactions on Biomedical Engineering* **52**(3): 431–444.

Kushner, H.J. and Yin, G.G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd Edn., Stochastic Modelling and Applied Probability, Springer, New York, NY.

Lortie, M. and Kearney, R.E. (2001). Identification of time-varying Hammerstein systems from ensemble data, *Annals of Biomedical Engineering* **29**(2): 619–635.

Mallat, S.G. (1998). *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, CA.

Marmarelis, V.Z. (2004). *Nonlinear Dynamic Modeling of Physiological Systems*, IEEE Press Series on Biomedical Engineering, Wiley-IEEE Press, Piscataway, NJ.

Nordsjo, A. and Zetterberg, L. (2001). Identification of certain time-varying nonlinear Wiener and Hammerstein systems, *IEEE Transactions on Signal Processing* **49**(3): 577–592.

Patan, K. and Korbicz, J. (2012). Nonlinear model predictive control of a boiler unit: A fault tolerant control study, *International Journal of Applied Mathematics and Computer Science* **22**(1): 225–237, DOI: 10.2478/v10006-012-0017-6.

Pawlak, M. and Hasiewicz, Z. (1998). Nonlinear system identification by the Haar multiresolution analysis, *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* **45**(9): 945–961.

Pawlak, M., Rafajłowicz, E. and Krzyżak, A. (2003). Postfiltering versus prefiltering for signal recovery from noisy samples, *IEEE Transactions on Information Theory* **49**(12): 3195–3212.

Rutkowski, L. (1984). On nonparametric identification with prediction of time-varying systems, *IEEE Transactions on Automatic Control* **29**(1): 58–60.

Rutkowski, L. (2004). Generalized regression neural networks in time-varying environment, *IEEE Transactions on Neural Networks* **15**(3): 576 – 596.

Saeedi, H., Mollahasani, N., Moghadam, M.M. and Chuev, G.N. (2011). An operational Haar wavelet method for solving fractional Volterra integral equations, *International Journal of Applied Mathematics and Computer Science* **21**(3): 535–547, DOI: 10.2478/v10006-011-0042-x.

Sansone, G. (1959). *Orthogonal Functions*, Interscience, New York, NY.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*, Wiley, New York, NY.

Skubalska-Rafajłowicz, E. (2001). Pattern recognition algorithms based on space-filling curves and orthogonal expansions, *IEEE Transactions on Information Theory* **47**(5): 1915–1927.

Śliwiński, P. (2010). On-line wavelet estimation of Hammerstein system nonlinearity, *International Journal of Applied Mathematics and Computer Science* **20**(3): 513–523, DOI: 10.2478/v10006-010-0038-y.

Śliwiński, P. (2013). *Nonlinear System Identification by Haar Wavelets*, Lecture Notes in Statistics, Vol. 210, Springer-Verlag, Heidelberg.

Stone, C.J. (1980). Optimal rates of convergence for nonparametric regression, *Annals of Statistics* **8**(6): 1348–1360.

Szego, G. (1974). *Orthogonal Polynomials*, 3rd Edn., American Mathematical Society, Providence, RI.

Van der Vaart, A. (2000). *Asymptotic Statistics*, Cambridge University Press, Cambridge.

Vörös, J. (2003). Recursive identification of Hammerstein systems with discontinuous nonlinearities containing dead-zones, *IEEE Transactions on Automatic Control* **48**(12): 2203–2206.

Walter, G.G. and Shen, X. (2001). *Wavelets and Other Orthogonal Systems With Applications*, 2nd Edn., Chapman & Hall, Boca Raton, FL.

Westwick, D.T. and Kearney, R.E. (2003). *Identification of Nonlinear Physiological Systems*, IEEE Press Series on Biomedical Engineering, Wiley-IEEE Press, Piscataway, NJ.

Wheeden, R. L. and Zygmund, A. (1977). *Measure and Integral: An Introduction to Real Analysis*, Pure and Applied Mathematics, Marcel Dekker Inc., New York, NY.

Zhou, D. and DeBrunner, V.E. (2007). Novel adaptive nonlinear predistorters based on the direct learning algorithm, *IEEE Transactions on Signal Processing* **55**(1): 120–133.

Przemysław Śliwiński received the M.Sc. and Ph.D. degrees in computer engineering from the Wrocław University of Technology, Poland, in 1996 and 2000, respectively. In 2000 he joined the Institute of Engineering Cybernetics, Wrocław University of Technology, where he is currently an assistant professor. His research interests are wavelets and their applications to nonparametric system identification, signal and image processing, analysis and compression. In 2006 he was a visiting research professor at the University of Arizona, Tucson, USA.

Zygmunt Hasiewicz received the M.Sc. and Ph.D. degrees in control engineering from the Wrocław University of Technology, Poland, in 1971 and 1974, respectively, and the D.Sc. degree from the Warsaw University of Technology, Poland, in 1993. In 1971 he joined the Institute of Engineering Cybernetics, Wrocław University of Technology, where he is currently a full professor. In 1976 he held visiting appointments at Lille University, France, and in 1995 and 2001 he was a visiting professor at the University of Manitoba, Winnipeg, Canada. His research interests include nonlinear system modeling and statistical methods in composite system identification.

Paweł Wachel received M.Sc. and Ph.D. degrees in control engineering from the Wrocław University of Technology respectively in 2004 and 2008. In 2007 he was a visiting scientist at the Department of Electrical and Computer Engineering, University of Manitoba, Canada, where he was developing identification algorithms for a class of nonlinear systems. Currently, Dr. Wachel is with the Institute of Computer Engineering, Control and Robotics of the Wrocław University of Technology. His research area is concerned with nonparametric system identification (including MISO systems driven by signals on manifolds) and optical profilometry algorithms.

Appendix A

Algorithm derivation

Denote for shortness the estimate in (1) with the fixed scale $K(k)$ and for the k data length by $\hat{\mu}_k(x)$ instead of $\hat{\mu}_{K(k)}(x)$. We have

$$\begin{aligned}
\hat{\mu}_k(x) &= \frac{\sum_{l=1}^{k} \phi_{K(k)}(x, x_l)\, y_l}{\sum_{l=1}^{k} \phi_{K(k)}(x, x_l)} \\
&= \frac{1}{\sum_{l=1}^{k} \phi_{K(k)}(x, x_l)} \sum_{l=1}^{k-1} \phi_{K(k)}(x, x_l)\, y_l \\
&\quad + \frac{\phi_{K(k)}(x, x_k)}{\sum_{l=1}^{k} \phi_{K(k)}(x, x_l)} y_k \\
&= \frac{\sum_{l=1}^{k-1} \phi_{K(k)}(x, x_l)}{\sum_{l=1}^{k} \phi_{K(k)}(x, x_l)} \hat{\mu}_{k-1}(x) \\
&\quad + \frac{\phi_{K(k)}(x, x_k)}{\sum_{l=1}^{k} \phi_{K(k)}(x, x_l)} y_k \\
&= \left(\hat{\mu}_{k-1}(x) - \hat{\mu}_{k-1}(x)\right) \\
&\quad + \frac{\sum_{l=1}^{k-1} \phi_{K(k)}(x, x_l)}{\sum_{l=1}^{k} \phi_{K(k)}(x, x_l)} \hat{\mu}_{k-1}(x) \\
&\quad + \frac{\phi_{K(k)}(x, x_k)}{\sum_{l=1}^{k} \phi_{K(k)}(x, x_l)} y_k
\end{aligned}$$

$$\begin{aligned}
&= \hat{\mu}_{k-1}(x) - \left[\hat{\mu}_{k-1}(x) - y_k\right] \\
&\quad \times \frac{\phi_{K(k)}(x, x_k)}{\sum_{l=1}^{k-1} \phi_{K(k)}(x, x_l) + \phi_{K(k)}(x, x_k)} \\
&= \hat{\mu}_{k-1}(x) \\
&\quad + \gamma_k(x, x_k)\left[y_k - \hat{\mu}_{k-1}(x)\right] \phi_{K(k)}(x, x_k),
\end{aligned}$$

where $\gamma_k(x, x_k) = 1/\kappa_k(x, x_k)$ with $\kappa_k(x, x_k) = \kappa_{k-1}(x, x_{k-1}) + \phi_{K(k)}(x, x_k)$. Now, varying $K(k)$, i.e., taking $K(l)$ for each incoming new data point (x_l, y_l) instead of the fixed scale $K(k)$, we get (10) and, in consequence, Algorithm 1.

Appendix B

Convergence analysis

Variance term. Our intermediate goal is to show that, for some $c > 0$, the following bound holds:

$$\operatorname{var}\left\{\sum_{l=1}^{k} y_l \phi_l\right\} \le c\kappa.$$

For the variance term in (14), we have (recall that ξ_l and z_l are zero-mean and independent of each other and of ϕ_l and μ_l)

$$\begin{aligned}
\operatorname{var}\{y_l \phi_l\} &= E\{y_l \phi_l\}^2 - E^2\{y_l \phi_l\} \\
&= E\{[\mu_l + \xi_l + z_l]\,\phi_l\}^2 \\
&\quad - E^2\{[\mu_l + \xi_l + z_l]\,\phi_l\} \\
&= E\mu_l^2 \phi_l^2 + E\phi_l^2 E\xi_l^2 \\
&\quad + E\phi_l^2 E z_l^2 - E^2\{\mu_l \phi_l\} \\
&\quad + 2E\mu_l \phi_l^2 E\xi_l + 2E\mu_l \phi_l^2 E z_l \\
&\quad + 2E\phi_l^2 E\xi_l E z_l
\end{aligned}$$

$$\begin{aligned}
&= E\mu_l^2 \phi_l^2 - E^2\{\mu_l \phi_l\} + E\phi_l^2 E\xi_l^2 + E\phi_l^2 E z_l^2 \\
&= \operatorname{var}\{\mu_l \phi_l\} + E\phi_l^2 \operatorname{var}\xi_l + E\phi_l^2 \operatorname{var} z_l.
\end{aligned}$$

Observing that in our case $\phi_{K(l)}(x, u) = \phi_{K(l)}^2(x, u)$ for each l, we get

$$\operatorname{var}\{y_l \phi_l\} \le E\phi_l^2 \cdot c_{\text{var}} = E\phi_l \cdot c_{\text{var}}, \tag{B1}$$

where $c_{\text{var}} = \max_x \mu^2(x) + \operatorname{var}\xi_l + \operatorname{var} z_l$.

Covariance term. By definition

$$\operatorname{cov}\{y_i \phi_i, y_j \phi_j\} = E y_i \phi_i y_j \phi_j - E y_i \phi_i E y_j \phi_j,$$

where for the latter two expectations we have

$$\begin{aligned}
E y_i \phi_i &= E\mu_i \phi_i + E\xi_i E\phi_i + E z_i E\phi_i \\
&= E\mu_i \phi_i \quad \text{and} \quad E y_j \phi_j = E\mu_j \phi_j,
\end{aligned} \tag{B2}$$

after recalling again that ξ_i and z_i are zero-mean and independent of each other and for a given i they are also independent of ϕ_i and μ_i. Next, observe that

$$Ey_i\phi_i y_j\phi_j = E\left\{(\mu_i+\xi_i+z_i)(\mu_j+\xi_j+z_j)\phi_i\phi_j\right\}$$

and that, after some reordering, we get

$$\begin{aligned}Ey_i\phi_i y_j\phi_j = {} & E\mu_i\phi_i\mu_j\phi_j + E\xi_j\mu_i\phi_i\phi_j \\ & + E\mu_j\xi_i\phi_j\phi_i + E\phi_i\xi_i\phi_j\xi_j \\ & + E\mu_i\phi_i\phi_j z_j + E\mu_j\phi_j\phi_i z_i \\ & + E\phi_i\xi_i\phi_j z_j + E\phi_i\phi_j\xi_j z_i + E\phi_i\phi_j z_i z_j.\end{aligned}$$

Exploiting the assumed whiteness of the external noise, z_k, and the fact that $Ez_1 = 0$, we get

$$\begin{aligned}Ey_i\phi_i y_j\phi_j &= E\mu_i\phi_i\mu_j\phi_j + E\xi_j\mu_i\phi_i\phi_j \\ &\quad + E\mu_j\xi_i\phi_j\phi_i + E\phi_i\xi_i\phi_j\xi_j \\ &\quad + E\mu_i\phi_i\phi_j Ez_j + E\mu_j\phi_j\phi_i Ez_i \\ &\quad + E\phi_i\xi_i\phi_j Ez_j + E\phi_i\phi_j\xi_j Ez_i \\ &\quad + E\phi_i\phi_j Ez_i Ez_j \\ &= E\mu_i\phi_i\mu_j\phi_j + E\xi_j\mu_i\phi_i\phi_j \\ &\quad + E\mu_j\xi_i\phi_j\phi_i + E\phi_i\xi_i\phi_j\xi_j,\end{aligned}$$

while recalling that $\xi_i = \sum_{l=1}^{\infty}\lambda_l\zeta_{i-l}$ and using the fact that $E\xi_i = 0$ (as, by definition, $\zeta_i = \mu_i - E\mu_i$) and that ϕ_j, μ_j and ϕ_i are independent of all ξ_i (since $j > i$; see the formula in (14)), we get

$$\begin{aligned}E\left\{y_i\phi_i y_j\phi_j\right\} &= E\mu_i\phi_i E\mu_j\phi_j + E\xi_j\mu_i\phi_i E\phi_j \\ &\quad + E\mu_j\phi_j\phi_i E\xi_i + E\phi_j E\phi_i\xi_i\xi_j \\ &= E\mu_i\phi_i E\mu_j\phi_j + E\xi_j\mu_i\phi_i E\phi_j \\ &\quad + E\phi_j E\phi_i\xi_i\xi_j.\end{aligned}$$

The latter combined with (B2) yields

$$\operatorname{cov}\left\{y_i\phi_i, y_j\phi_j\right\} = E\phi_j E\mu_i\phi_i\xi_j + E\phi_j E\phi_i\xi_i\xi_j.$$

Assume now for simplicity that $E\mu_i = 0$, i.e., $\zeta_i = \mu_i$. We get

$$\begin{aligned}&\operatorname{cov}\left\{y_i\phi_i, y_j\phi_j\right\} \\ &= E\phi_j\sum_{l=1}^{\infty}\lambda_l E\left\{\mu_i\phi_i\mu_{j-l}\right\} \\ &\quad + E\phi_j\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}E\left\{\phi_i\lambda_k\mu_{i-k}\lambda_l\mu_{j-l}\right\}.\end{aligned}$$

Since μ_i's are i.i.d. and zero-mean (hence $E\mu_i\mu_j = 0$ for $i \neq j$), the former sum reduces to the single term

$$\sum_{l=1}^{\infty}\lambda_l E\mu_i\phi_i\mu_{j-l} = \lambda_{j-i}E\mu_i^2 E\phi_i,$$

while the latter double sum to the single one

$$\begin{aligned}&\sum_{k=1}^{\infty}\sum_{l=1}^{\infty}E\left\{\phi_i\lambda_k\mu_{i-k}\lambda_l\mu_{j-l}\right\} \\ &\quad = \sum_{k=1}^{\infty}\lambda_k\lambda_{k-j-i}E\left\{\phi_i\mu_{i-k}\mu_{j-(j-(i-k))}\right\} \\ &\quad = \sum_{k=1}^{\infty}\lambda_k\lambda_{k-j-i}E\phi_i\mu_{i-k}^2 \\ &\quad = E\phi_i\mu_i^2\sum_{k=1}^{\infty}\lambda_k\lambda_{k-j-i}.\end{aligned}$$

Hence

$$\operatorname{cov}\left\{y_i\phi_i, y_j\phi_j\right\} = E\phi_j E\mu_i^2\phi_i\left(\lambda_{j-i} + \sum_{k=1}^{\infty}\lambda_k\lambda_{k-j-i}\right).$$

Recalling (14), we get

$$\begin{aligned}&\sum_{i=1}^{k}\sum_{j=i+1}^{k}\operatorname{cov}\left\{y_i\phi_i, y_j\phi_j\right\} \\ &\quad \leq \sum_{i=1}^{k}E\mu_i^2\phi_i\sum_{j=i+1}^{k}\left(\lambda_{j-i} + \sum_{k=1}^{\infty}\lambda_k\lambda_{k-j-i}\right),\end{aligned}$$

since $E\phi_j \leq 1$. Moreover, the following term is bounded, i.e., for some $c > 0$ we have that

$$\begin{aligned}&\sum_{j=i+1}^{k}\left(\lambda_{j-i} + \sum_{k=1}^{\infty}\lambda_k\lambda_{k-j-i}\right) \\ &\qquad = \sum_{j=i+1}^{k}\lambda_{j-i} + \sum_{k=1}^{\infty}\lambda_k\sum_{j=i+1}^{k}\lambda_{k-j-i} < c,\end{aligned}$$

assuming stability. Thus

$$2\sum_{i=1}^{k}\sum_{j=i+1}^{k}\operatorname{cov}\left|y_i\phi_i, y_j\phi_j\right| \leq c_{\text{cov}}\sum_{i=1}^{k}E\phi_i, \tag{B3}$$

where $c_{\text{cov}} = 2c \cdot \max_x \mu_i^2$. Finally, combining (B1) and (B3), we obtain

$$\operatorname{var}\left\{\sum_{l=1}^{k}y_l\phi_l\right\} \leq c_{\text{var}}\sum_{i=1}^{k}E\phi_i + c_{\text{cov}}\sum_{i=1}^{k}E\phi_i,$$

and recalling that $\kappa = \sum_{l=1}^{k}E\phi_l$ (as in (13)) we get

$$\operatorname{var}\hat{\vartheta}_k(x) \leq \frac{c_{\text{var}} + c_{\text{cov}}}{\kappa^2}\sum_{l=1}^{k}E\phi_l = \frac{c_{\text{var}} + c_{\text{cov}}}{\kappa}.$$

Appendix C
Convergence rate

From the assumption A1 we easily conclude that for each estimation point x there exist some constants $\Delta_x \geq \delta_x > 0$ being, respectively, the upper and lower bounds of $f(x)$ in the neighborhood of x. Hence, for positive scale $K(l)$, the following bound holds for all $l = 1, 2, \ldots$:

$$\Delta_x \geq \frac{E\phi_{K(l)}(x, x_l)}{2^{-K(l)}} \geq \delta_x.$$

Exploring now the bias formula in (12) under the assumption that the nonlinearity $\mu(x)$ satisfies the Lipschitz condition in a neighborhood of x, we get that for sufficiently large $K(l)$ it holds that

$$\begin{aligned} 0 \leq b_l &= \int_{\mathrm{supp}\phi_{K(l)}} |\mu(x) - \mu(u)| \frac{f(u)}{\int_{\mathrm{supp}\phi_{K(l)}} f(u)\,\mathrm{d}u}\,\mathrm{d}u \\ &\leq c_m \int_{\mathrm{supp}\phi_{K(l)}} |x - u| \frac{f(u)}{\int_{\mathrm{supp}\phi_{K(l)}} f(u)\,\mathrm{d}u}\,\mathrm{d}u \\ &\leq 2^{-K(l)} c_m \end{aligned}$$

for some Lipschitz constant $c_m > 0$ (cf. the assumption A2), and hence that

$$\begin{aligned} 0 \leq \mathrm{bias}\hat{\vartheta}_k(x) &\leq c_m \frac{\sum_{l=1}^k 2^{-K(l)} \cdot E\phi_l}{\sum_{l=1}^k E\phi_l} \\ &\leq c_m \frac{\sum_{l=1}^k 2^{-K(l)} \cdot 2^{-K(l)} \Delta_x}{\sum_{l=1}^k 2^{-K(l)} \delta_x}. \end{aligned}$$

Observe that the obvious identity $\sum_{l=1}^k l^{-\alpha} = \int_0^k \lceil x \rceil^{-\alpha}\,\mathrm{d}x$ implies (for sufficiently large k) the lower and upper bounds

$$\begin{aligned} c_\alpha k^{1-\alpha} &< \int_0^k (x+1)^{-\alpha}\,dx \leq \sum_{l=1}^k l^{-\alpha} \leq \int_0^k x^{-\alpha} dx \\ &= C_\alpha k^{1-\alpha} \end{aligned}$$

for some $c_\alpha, C_\alpha > 0$, and taking $K(l) = 3^{-1}\log_2 l$ (as in (18)) results in the following bound for the bias error:

$$0 \leq \mathrm{bias}\hat{\vartheta}_k(x) \leq c_m \frac{\Delta_x}{\delta_x} \frac{\sum_{l=1}^k l^{-\frac{2}{3}}}{\sum_{l=1}^k l^{-\frac{1}{3}}} \leq c_{\mathrm{bias}} k^{-\frac{1}{3}} \qquad \text{(C1)}$$

for some $c_{\mathrm{bias}} > 0$. Passing to the variance error we obtain, for some $c'_\vartheta > 0$, that

$$\mathrm{var}\hat{\vartheta}_k(x) \leq c_\vartheta \frac{1}{\sum_{l=1}^k E\phi_l} \leq \frac{c_\vartheta}{\delta_x} \frac{1}{\sum_{l=1}^k 2^{-K(l)}} \leq c'_\vartheta k^{-\frac{2}{3}}.$$

Combining the above results we get, for some $c_{\mathrm{MSE}} > 0$, that

$$\mathrm{MSE}\hat{\vartheta}_k(x) = \mathrm{bias}^2\hat{\vartheta}_k(x) + \mathrm{var}\hat{\vartheta}_k(x) \leq c_{\mathrm{MSE}} k^{-\frac{2}{3}}.$$

In a similar way one can derive for the denominator $\hat{\eta}_k(x)$ that

$$\mathrm{MSE}\hat{\eta}_k(x) = \mathrm{var}\,\hat{\eta}_k(x) \leq c'_{\mathrm{MSE}} k^{-\frac{2}{3}},$$

some $c'_{\mathrm{MSE}} > 0$.

Received: 9 May 2012
Revised: 30 December 2012