Recursive Nonparametric Identification of Hammerstein Systems

by WŁODZIMIERZ GREBLICKI

Institute of Engineering Cybernetics, Technical University of Wrocław, Wrocław, Poland

and MIROSŁAW PAWLAK

Department of Electrical Engineering, University of Manitoba, Winnipeg, Manitoba, Canada, R3T 2N2

ABSTRACT: *A discrete-time nonlinear Hammerstein system is identified. The dynamical subsystem is recovered using the standard correlation method. The main results concern the estimation of the nonlinear memoryless subsystem. The class of nonlinearities considered in the paper consists of a broad class of functions which cannot be parametrized. Two new algorithms of a recursive form for estimating the nonlinear characteristic are proposed. It is shown that they converge at all continuity points of the characteristic. The integrated absolute error also converges to zero. The algorithms are equivalent with respect to the asymptotical rate of convergence. The efficiency of the algorithms is discussed, and numerical examples are presented.*

I. Introduction

We identify a discrete-time nonlinear system, shown in Fig. 1, usually called a system of the Hammerstein type. It consists of two subsystems connected in a cascade. The first is nonlinear and memoryless and its characteristic is denoted by m. The second is linear and dynamic and has a weighting function $\{g_i\}$, $i = 0, 1, 2, \ldots$. The input of the whole system is a stationary white noise, whereas the output is disturbed by a zero mean stationary white noise. The system of that form has been employed in many areas as a model of nonlinear distortions. For example:

- (i) In communication theory, where it is known as the bandpass nonlinear network, the system corresponds, for example, to the final detector-video amplifier portion of a communication receiver, see DeBoer (**1**).
- (ii) In auditory theory, the system has been proposed as a descriptor for nonlinear distortion occurring in the inner ear, see Pfeiffer (**2**) and references cited therein.
- (iii) The human visual system as well as neural networks can be modeled in this form. Here usually $m(u) = \text{const } u^{\alpha}$, $0 < \alpha < 1$, see Hall and Hall (**3**) and Korenberg (**4**).
- (iv) In image processing, the model of such form is referred to as an input non linearity (**5**). A common example of that is radiography, where the quantity of interest, the X-ray absorption, is proportional approximately to the logarithm of the transmitted X-ray intensity. This, in turn, is passed through the linear system.

(v) The Hammerstein system has been employed for modeling the noise process. That is, to obtain dependent non-Gaussian noise from Gaussian independent background ($\{U_n\}$ are i.i.d. $N(0,1)$ random variables). This is called "transformation noise technique", see Martinez *et al.* (**6**).

Further applications are in adaptive control (**7**) and adaptive noise cancellation problems (**8**).

The problem in the present paper is to identify both the subsystems from the input-output observations of the whole system. Since the signal interconnecting both the subsystems is inaccessible to measurement, it is clearly impossible to estimate consistently $m(u)$ or $\{g_i\}$. It is obvious that we can, if only, estimate merely $\alpha m(u)$ and $\{\gamma g_i\}$, where α and γ are some unknown factors.

Since the weighting function can be estimated by the standard correlation method; recovering the nonlinear characteristic is the main problem of the paper. Thus far, all authors studying the problem have assumed that the nonlinearity is of a polynomial form, see e.g., the original work of Narendra and Gallman (**9**), as well as later papers by Billings and Fakhouri (**10–12**), Chang and Luus (**13**), Gallman (**14**), (**15**), Tchatachar and Ramaswamy (**16**), Kaminskas (**17**), Hsia (**18**), Stoica (**19**), Stoica and Söderstrom (**20**), as well as Kung and Shih (**21**).

It is clear that the class of characteristics of the nonlinear subsystem considered by all these authors is not wide for it does not even contain any discontinuous functions. It has been noted (**11**), that due to the Weierstrass theorem, we can approximate very closely any continuous function defined on a compact set by a polynomial. However, there is usually no prior information about the smoothness of the underlying function. Moreover, even accepting the continuity condition, we do not know how many terms in the approximating polynomial should be taken. A large number of the terms leads to very computationally extensive algorithms, whereas a small number of the terms can poorly represent the function and yields the substantial identification error. In any case, the identification methods based on such approximations are not consistent, i.e. the estimation error is always positive even if a number of the input-output data tends to infinity. Furthermore, the usage of the Volterra representation of the system is restricted, roughly speaking, to the class of systems with analytic nonlinearities, see Sandberg (**22**).

Recently, observing that the nonlinear characteristic can be expressed as a regression function, Greblicki and Pawlak (**23-25**) proposed a new approach to the problem of recovering the nonlinearity. Their idea, based on the nonparametric estimation of the regression function, made it possible for them to introduce an algorithm which is considerably simpler than those known in the literature. The algorithm is not only completely independent of procedures which can be used for identification of the linear subsystem, but, first of all, recovers nonlinear characteristics even if they are not continuous. The class of nonlinearities considered by the authors, and also studied in this paper, consists of all Borel functions increasing not faster than linear functions.

However, the estimates examined by them need to be recomputed entirely when an additional observation is combined with previous ones, i.e. they are not recursive. In this paper, we propose two new algorithms for estimating the nonlinear

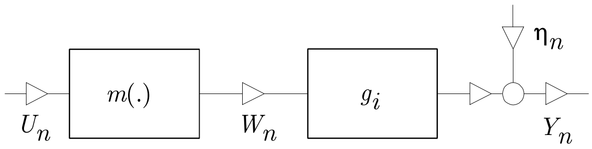

Figure 1: The identified nonlinear system.

characteristic. They are of a recursive form and extremely easy in calculations. Besides saving computation effort, the recursive estimates possess some better asymptotic properties, i.e. they can have a smaller variance. We examine not only pointwise, but also global properties of the algorithms and show that the integrated absolute error converges to zero as the number of observations increases to infinity. They are equivalent with respect to the asymptotical rate of convergence. However, they can have different asymptotic variance values. The convergence of the algorithms is also illustrated numerically.

The recursive identification algorithms of the Hammerstein system, in the case of the polynomial nonlinear characteristic, have been examined in a number of papers, (**9**), (**14**), (**15**), (**17**). Stoica (**19**) shows, however, that they cannot be convergent even in the case of the quadratic nonlinear characteristic.

II. Identification Algorithms

We assume that the input signal $\{U_n\}$, $n = ..., -1, 0, 1, ...$, is a stationary white noise and that

$$EU_n^2 < \infty. \tag{1}$$

Throughout the paper we assume that the input signal has a probability density f.

The characteristic of the first subsystem is a Borel function and satisfies the following condition:

$$|m(u)| \leq c_1 + c_2|u|. \tag{2}$$

In turn, the linear dynamic subsystem is described by the following state equation:

$$\begin{cases} X_{n+1} &= AX_n + bW_n \\ Y_n &= c^T X_n + \eta_n, \end{cases} \tag{3}$$

where $n = ..., -1, 0, 1, ..$ and $W_n = m(U_n)$. X_n is a state vector, $\{\eta_n\}$ is a stationary white noise with zero mean and finite variance, whereas b and c are some vectors. A is a matrix assumed to be asymptotically stable. We assume, moreover, that $\{U_n\}$ and $\{\eta_n\}$ are mutually independent. Let $\{g_i = c^T A^{i-1} b\}$ be the pulse response of the dynamical subsystem.

It is clear that (1) and (2) imply

$$EW_n^2 < \infty. \tag{4}$$

From (4), the asymptotical stability of A and Anderson (**26**), it follows that both X_n and Y_n are random variables. We would like to emphasize here that restrictions (1) and (2) are totally independent of the identification method presented in the paper. They simply guarantee (4) to be satisfied and in this way make the problem well posed in the sense that the output of the system is a random variable. We note that the condition (2) may be relaxed to $|m(u)| < c_1 + c_2|u|^q$, $q > 0$. Then, in order to have (4) in force we have to assume that $E|Y|^{2q} < \infty$.

Furthermore, we note that both $\{X_n\}$ and $\{Y_n\}$, $n = ..., -1, 0, 1, ...$ are strictly stationary stochastic processes, see also Anderson (**26**).

The class of Borel functions satisfying (2) is so wide that it can not be parametrized. Due to that the problem of recovering the characteristic is nonparametric. We would like to bring out that this is the main difference between the problem discussed here, as well as in Greblicki and Pawlak (**23**, **24**), and those considered by other authors. The case considered by them is much simpler since they assumed that the characteristic is of a polynomial form of a known order. Therefore, the problem investigated by them is parametric for they estimate only a finite number of coefficients.

From (3), it follows that

$$E\{Y_{n+1}|U_n = u\} = \alpha m(u) + \beta, \tag{5}$$

where

$$\alpha = c^T b \text{ and } \beta = c^T AEX_n = E\{m(U)\}\sum_{i=2}^{\infty} g_i.$$

In order to estimate $\alpha m(u)$, we first do the regression function in (5). For estimating the regression in (5), we propose two recursive algorithms, μ'_n and and $\bar{\mu}'_n$. The first algorithm is defined in the following way:

$$\left.\begin{aligned}
\mu'_n(u) &= \mu'_{n-1}(u) + \frac{1}{F_n(u)}\left[Y_n - \mu'_{n-1}(u)\right]\frac{1}{h(n-1)}K\left(\frac{u - U_{n-1}}{h(n-1)}\right),\\
F_n(u) &= F_{n-1}(u) + \frac{1}{h(n-1)}K\left(\frac{u - U_{n-1}}{h(n-1)}\right),\\
\mu'_0(u) &= F_0(u) = 0.
\end{aligned}\right\} \tag{6}$$

The second algorithm proposed here is even simpler, since

$$\left.\begin{aligned}
\bar{\mu}'_n(u) &= \bar{\mu}'_{n-1}(u) + \frac{1}{\bar{F}_n(u)}\left[Y_n - \bar{\mu}'_{n-1}(u)\right]K\left(\frac{u - U_{n-1}}{h(n-1)}\right),\\
\bar{F}_n(u) &= \bar{F}_{n-1}(u) + K\left(\frac{u - U_{n-1}}{h(n-1)}\right),\\
\bar{\mu}'_0(u) &= \bar{F}_0(u) = 0.
\end{aligned}\right\} \tag{7}$$

In both the estimates K is a Borel kernel function while $\{h(n)\}$ is a positive number sequence.

We note that implementing $\mu'_n(u)$ and $\bar{\mu}'_n(u)$ require only saving $\mu'_{n-1}(u)$ and $\bar{\mu}'_{n-1}(u)$ at all u of interest.

We then make use of the only restriction of the paper strictly connected with the identification method presented by us, i.e. the assumption that

$$m(u=0)=0 \tag{8}$$

In the authors' opinion (8) is not very restrictive and is satisfied in most applications.

Owing to (8), we estimate $\alpha m(u)$ with $\mu_n(u)$ and $\mu'_n(u)$ which are defined in the following way:

$$\mu_n(u)=\mu'_n(u)-\mu'_n(u=0)$$

and

$$\bar{\mu}_n(u)=\bar{\mu}'_n(u)-\bar{\mu}'_n(u=0).$$

As far as the linear subsystem, in order to estimate $\{\gamma g_i\}$, $i=0,1,2,\ldots$, we make use of the following equality:

$$\mathrm{cov}[Y_{n+i},U_n]=\gamma g_i, \tag{9}$$

$i=0,1,2,\ldots$, $\gamma=$cov(U,W). Having observed that, we estimate γg_i with

$$\rho_{in}=\frac{1}{n}\sum_{j=1}^{n-1}(Y_{i+j}-\bar{Y})(U_j-\bar{U}),$$

where

$$\bar{U}=\frac{1}{n}\sum_{j=1}^{n}U_j$$

and

$$\bar{Y}=\frac{1}{n}\sum_{j=1}^{n}Y_{i+j}.$$

Recalling that $\{Y_n\}$ is strictly stationary and ergodic (**27**), we can easily show that

$$\rho_{in}\overset{n}{\to}\gamma g_i \text{ in probability,}$$

see, e.g., (**27**). It is now also apparent that, since $\{W_n\}$ is not measured, γ cannot be estimated.

The identification of the linear subsystem may be carried out in the frequency domain. Indeed, taking the Fourier transform of the relationship (9) gives

$$H_{YU}(\omega)=\gamma G(\omega),\ |\omega|<\pi,$$

where $H_{YU}(\omega)$ is the cross-spectral density function of the processes $\{U_n\}$ and $\{Y_n\}$, whereas $G(\omega)$ is the transfer function of the subsystem.

III. Convergence of the Algorithms

Estimates (6) and (7) can be rewritten in the following form :

$$\mu_n'(u) = \frac{\sum_{i=0}^{n-1} Y_{i+1} \frac{1}{h(i)} K\left(\frac{u-U_i}{h(i)}\right)}{\sum_{i=0}^{n-1} \frac{1}{h(i)} K\left(\frac{u-U_i}{h(i)}\right)} \tag{6a}$$

and

$$\bar{\mu}_n'(u) = \frac{\sum_{i=0}^{n-1} Y_{i+1} K\left(\frac{u-U_i}{h(i)}\right)}{\sum_{i=0}^{n-1} K\left(\frac{u-U_i}{h(i)}\right)}. \tag{7a}$$

Both of them can be regarded as recursive versions of the following one :

$$\hat{\mu}_n'(u) = \frac{\sum_{i=0}^{n-1} Y_{i+1} K\left(\frac{u-U_i}{h(n)}\right)}{\sum_{i=0}^{n-1} K\left(\frac{u-U_i}{h(n)}\right)}. \tag{10}$$

applied by Greblicki and Pawlak (**23**, **24**).

All three estimates given above have been studied in the statistical literature since the original works of Watson (**28**), and Nadaraya (**29**), see also Stone (**30**). More recent results were presented by Devroye and Wagner (**31**), Devroye (**32**), Greblicki, Krzyżak and Pawlak (**33**) as well as Greblicki and Pawlak (**34**, **35**). All of those authors studied, however, a simpler case since they assumed that pairs (U_i, Y_{i+1})'s are mutually independent. In this paper, we deal with a more complicated problem since $\{Y_n\}$, being the output of a dynamic system, is a sequence of dependent random variables.

The bounded and nonnegative Borel kernel satisfies the following two conditions:

$$0 < \int K(u)du \tag{11}$$

and

$$|u|K(u) \to 0 \text{ as } |u| \to \infty. \tag{12}$$

There exists a wide class of kernels satisfying (11) and (12), some of them are listed below,

$$\text{(a)} \qquad K(u) = \begin{cases} 1 & \text{for } |u| \le 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{(b)} \qquad K(u) = \begin{cases} 1 - |u| & \text{for } |u| \le 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{(c)} \qquad K(u) = \begin{cases} 1 - u^2 & \text{for } |u| \le 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{(d)} \qquad K(u) = e^{-u^2},$$

$$\text{(e)} \qquad K(u) = e^{-|u|},$$

$$\text{(f)} \qquad K(u) = \frac{1}{1+u^2}.$$

As concerning the positive number sequence, we assume that

$$h(n) \xrightarrow{n} 0. \tag{13}$$

Examining estimate (6) we assume, moreover, that

$$\frac{1}{n^2} \sum_{i=0}^{n} \frac{1}{h(i)} \xrightarrow{n} 0, \tag{14}$$

whereas studying (7), we restrict our considerations to sequences satisfying

$$\sum_{n=0}^{\infty} h(n) = \infty. \tag{15}$$

As concerning estimate (10), Greblicki and Pawlak (23) showed its consistency assuming that, in addition to (13),

$$nh(n) \xrightarrow{n} \infty.$$

It is clear that classes of sequences satisfying (13), (14) and (13), (15) and (13), (16) are different. Precisely speaking, they are strictly nested, i.e. (16) implies (14) and (14) implies (15). Taking, however, $h(n) = cn^{-\delta}$, which is often the case, we find all of them fulfilled for $0 < \delta < 1$.

The following theorems present the conditions for the estimates' consistency. Some auxiliary results required for the proofs are deferred to the Appendix. For convenience, by $C(f_1, \ldots, f_N)$, we denote a set of all points at which all $f_1, \ldots, f_N$, are continuous.

Theorem I
Let the bounded nonnegative Borel kernel satisfy (11) and (12). Let the nonnegative number sequence satisfy (13) and (14). Then

$$\mu_n(u) \xrightarrow{n} \alpha m(u) \text{ in probability}$$

at all $u \in C(f, m)$ and also for almost all $u \in R$ at which $f(u) > 0$.

Proof: Let $u \in C(f,m)$. It is clear that it suffices to verify that

$$\mu_n'(u) \overset{n}{\to} \alpha m(u) + \beta \text{ in probability.}$$

Denoting

$$g_n(u) = \frac{1}{n}\sum_{i=0}^{n-1} Y_{i+1}\frac{1}{h(i)}K_i \tag{17}$$

and

$$f_n(u) = \frac{1}{n}\sum_{i=0}^{n-1} \frac{1}{h(i)}K_i, \tag{18}$$

we have

$$\mu_n'(u) = \frac{g_n(u)}{f_n(u)}. \tag{19}$$

Here K_i stands for $K[(u-U_i)/h(i)]$.

Using Lemma 1 (see the Appendix) and Toeplitz's lemma, see, e.g. Loéve (**36**), we obtain

$$Eg_n(u) \overset{n}{\to} (\alpha m(u)+\beta)f(u)\int K(v)dv. \tag{20}$$

In turn, applying Lemma 2 (see the Appendix) , we get

$$\begin{aligned}
\mathrm{var} g_n(u) &= \frac{1}{n^2}\sum_{i=0}^{n-1}\frac{1}{h(i)}\mathrm{var}\,(Y_{i+1}K_i) \\
&+ \frac{2}{n^2}\sum_{i=0}^{n-2}\sum_{j=i+1}^{n-1}\frac{1}{h(i)h(j)}\mathrm{cov}\,(Y_{i+1}K_i, Y_{j+1}K_j) \\
&= \frac{1}{n^2}\sum_{i=0}^{n-1}\frac{1}{h^2(i)}\mathrm{var}\,(Y_{i+1}K_i) + \frac{2}{n^2}\sum_{i=0}^{n-2}\sum_{j=i+1}^{n-1}\frac{1}{h(i)h(j)}c^TQ_{i,j}(u)cEK_j \\
&= V_1(u) + V_2(u), \text{ say.}
\end{aligned}$$

Obviously,

$$\begin{aligned}
V_1(u) &\le \frac{1}{n^2}\sum_{i=0}^{n-1}\frac{1}{h^2(i)}E\{Y_{i+1}^2K_i^2\} \\
&\le k^*\frac{1}{n^2}\sum_{i=0}^{n-1}\frac{1}{h(i)}\left[\frac{1}{h(i)}E\left\{\Phi(U_i)K\left(\frac{u-U_i}{h(i)}\right)\right\}\right],
\end{aligned} \tag{21}$$

where

$$\Phi(u) = E\left\{Y_{n+1}^2|U_n = u\right\} = (\alpha m(u)+\beta)^2 + \mathrm{var}\ \eta_n,$$

and $k^* = \max K(u)$. As, by virtue of Lemma 1 (see the Appendix),

$$\lim_{h\to 0}\frac{1}{h}E\left\{\Phi(U_0)K\left(\frac{u-U_0}{h}\right)\right\} = \Phi(u)f(u)\int K(v)dv,$$

using (21), we get

$$V_1(u) = O\left(\frac{1}{n^2}\sum_{i=0}^{n-1}\frac{1}{h(i)}\right). \tag{22}$$

We shall now consider the term $V_2(u)$. Using Lemma 1 (see the Appendix), the term $c^T Q_{i,j}(u)c$ may be rewritten as follows:

$$\begin{aligned} &EK_i c^T A^{j+1} P (c^T A^{i+1})^T + \operatorname{cov}(K_i, W_i) g_{j-i+1} (c^T A^{i+1} \bar{X}_0)^T \\ &+ E\{\delta_i^2 K_i\} g_1 g_{j-i+1} + \operatorname{var}(W)\, EK_i \sum_{r=0}^{i-1} g_{j-r+1} g_{i-r+1} \\ &+ EW \operatorname{cov}(K_i, W_i) g_{j-i+1} \sum_{r=0}^{i} g_{r+i}. \end{aligned} \tag{23}$$

Let us take into account the fourth term in $V_2(u)$ obtained in this way, i.e., let us consider

$$\frac{1}{n^2}\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1} v_i v_j \sum_{r=0}^{i-1} g_{j-r+1} g_{i-r+1}, \tag{24}$$

where $v_i = h^{-1}(i) EK_i$. The above quantity is certainly not greater than

$$(\sup_n v_n)\frac{1}{n^2}\left(\sum_{i=1}^{\infty}|g_i|\right)^2 \sum_{i=1}^{n-1} v_i \leq \frac{1}{n}\left(\sup_n v_n \sum_{i=1}^{\infty}|g_i|\right)^2.$$

Because of the asymptotic stability of A,

$$\sum_{i=1}^{\infty}|g_i| < \infty.$$

On the other hand, by virtue of Lemma 1 (see the Appendix), $\sup_n v_n$ is also finite. Therefore, the quantity in (24) is of order $O(1/n)$. Using successively Lemma 1 from the Appendix and the stability assumption, the same property can be verified for the remaining terms in (23). Thus, we find

$$V_2(u) = O\left(\frac{1}{n}\right).$$

From this and (22) it follows that

$$\operatorname{var} g_n(u) = O\left(\frac{1}{n^2}\sum_{i=0}^{n-1}\frac{1}{h(i)}\right). \tag{25}$$

In turn, this, along with (20) and (14), implies

$$g_n(u) \xrightarrow{n} (\alpha m(u) + \beta) f(u) \int K(v) dv$$

in probability.

Using similar arguments, we can easily show that

$$f_n(u) \xrightarrow{n} f(u) \int K(v)dv$$

in probability. Finally, recalling (17), (18) and (19), we get the desired convergence of $\mu'_n(u)$.

Since the same property can be shown for almost all $u \in R$, the Theorem has been proved. □

As a simple consequence of Theorem 1 and Lebesgue's dominated convergence theorem on product spaces, see e.g. Glick (**37**), we get a theorem on the integrated absolute error.

Theorem II
Let $|U_0| \le \rho < \infty$ a;most surely. Under conditions of Theorem 1

$$\int |\mu_n(u) - \alpha m(u)| f(u) du \xrightarrow{n} 0$$

in probability.

The next two theorems deal with estimate (7). They are given without proofs since the arguments are similar to those used in the proof of Theorem 1.

Theorem III
Let the bounded nonnegative Borel kernel satisfy (11) and (12). Let the nonnegative number sequence satisfy (13) and (15). Then

$$\bar{\mu}_n(u) \xrightarrow{n} \alpha m(u) \text{ in probability}$$

at all $u \in C(f, m)$ and also for almost all $u \in R$ at which $f(u) > 0$.

As an immediate consequence of Theorem III, we obtain Theorem IV.

Theorem IV
Let $|U_0| \le \rho < \infty$ a;most surely. Under conditions of Theorem 1

$$\int |\bar{\mu}_n(u) - \alpha m(u)| f(u) du \xrightarrow{n} 0$$

in probability.

IV. The Rate of Convergence and Efficiency

In this section, we examine the rate of convergence of both the algorithms. In order to do it, we need the following easily verified fact. If

$$|\xi_n - \xi| = O(a_n) \text{ and } |\eta_n - \eta| = O(b_n)$$

in probability, then

$$\left|\frac{\xi_n}{\eta_n} - \frac{\xi}{\eta}\right| = O(c_n),$$

in probability, where c_n =max(a_n, b_n), provided that $\eta > 0$. Writing $\xi_n = O(a_n)$ in probability we mean that $\alpha_n \xi_n / a_n \to 0$ as $n \to \infty$ in probability for any number sequence $\{\alpha_n\}$ convergent to zero.

In this section, both m and f have two bounded derivatives in the neighborhood of $u \in R$. The Borel kernel satisfies not only (11) and (12), but also

$$\int vK(v)dv = 0 \tag{27}$$

and

$$\int v^2 K(v)dv \neq 0. \tag{28}$$

Let us observe that (27) is fulfilled by all even kernels having bounded support, e.g. those given in Section III in examples (a)–(c). The restriction is also satisfied by kernels (d), (e), but not by (f). We also assume that

$$h(n) \sim n^{-1/5}, \tag{29}$$

i.e. that $n^{1/5}h(n)$ has a positive limit as n tends to infinity. Expanding both m and f the Taylor series at $u \in R$ and using (27) and (28), we find

$$\left|\frac{1}{h(i)} E\{Y_{i+1}K_i\} - (\alpha m(u) + \beta) f(u) \int K(v)dv\right| = O(h^2(i)) = O(i^{-2/5})$$

and

$$\left|\frac{1}{h(i)} EK_i - f(u) \int K(v)dv\right| = O(h^2(i)) = O(i^{-2/5}).$$

From (25) and (28), it follows that

$$\left|g_n(u) - (\alpha m(u) + \beta) f(u) \int K(v)dv\right| = O(n^{-2/5})$$

in probability. In the same way, we can verify that

$$\left|f_n(u) - f(u) \int K(v)dv\right| = O(n^{-2/5})$$

in probability. From this and (26), it follows that

$$\left|\mu_n(u) - \alpha m(u) \int K(v)dv\right| = O(n^{-2/5}) \text{ in probability.}$$

Using similar arguments, one can show that

$$\left|\bar{\mu}_n(u) - \alpha m(u) \int K(v)dv\right| = O(n^{-2/5})$$

in probability.

Therefore, the rates obtained for both the estimates are asymptotically equivalent. They are also equal that obtained by Greblicki and Pawlak (**23**) for estimate (10).

It would be interesting. however, to know which estimate, if any, is the most efficient one. To do this, let us try to get the asymptotic variance of the estimate (6), (7) and (10). Quick inspection of the proof of Theorem 2 in Greblicki and Pawlak (**24**) follows

$$nh(n)\mathrm{var}(\hat{\mu}'_n) \overset{n}{\to} \theta(u), \tag{30}$$

where

$$\theta(u) = \left[(\alpha m(u) + \beta)^2 + \mathrm{var}(\eta) + \mathrm{var}(W)\sum_{i=2}^{\infty} g_i^2\right] \int K^2(z)dz/f(u)$$

at all $u \in C(m, f)$.

Proceeding similarly and taking the proof of Theorem I and Theorem III into account, we obtain

$$\frac{n^2}{\sum_{i=1}^{n} \frac{1}{h(i)}} \mathrm{var}(\mu'_n) \overset{n}{\to} \theta(u)$$

and

$$\sum_{i=1}^{n} h(i)\mathrm{var}(\bar{\mu}'_n) \overset{n}{\to} \theta(u),$$

at all $u \in C(m, f)$.

Let, now $h(n) = cn^{-\delta}$, $c > 0$, $0 < \delta < 1$, what is a common description for the sequence. For that $\{h(n)\}$, the expression in (30) comes to

$$n^{1-\delta}\mathrm{var}(\hat{\mu}'_n) \overset{n}{\to} \theta(u)/c,$$

at all $u \in C(m, f)$.

In turn, using the following easily verified facts

$$\frac{1}{n^{1+\delta}} \sum_{i=1}^{n} i^{\delta} \overset{n}{\to} \frac{1}{1+\delta}$$

and

$$\frac{1}{n^{1-\delta}} \sum_{i=1}^{n} i^{-\delta} \overset{n}{\to} \frac{1}{1-\delta},$$

we have

$$n^{1-\delta}\mathrm{var}(\mu'_n) \overset{n}{\to} \frac{1}{1+\delta}\theta(u)/c$$

and

$$n^{1-\delta}\mathrm{var}(\bar{\mu}'_n) \overset{n}{\to} (1-\delta)\theta(u)/c$$

at all $u \in C(m, f)$.

In particular, the value h picked up in (29), yields

$$n^{4/5}\mathrm{var}(\hat{\mu}'_n) \xrightarrow{n} \theta(u)/c,$$

$$n^{4/5}\mathrm{var}(\mu'_n) \xrightarrow{n} (5/6)\theta(u)/c$$

and

$$n^{4/5}\mathrm{var}(\bar{\mu}'_n) \xrightarrow{n} (4/5)\theta(u)/c.$$

It reveals that the recursive estimates are more efficient that their nonrecursive counterpart (10). Moreover, the estimate (7) posseses a lower asymptotic variance than estimate (6).

Nevertheless, more study is needed since little is known about the behavior of the algorithms for small and moderate number of observations even if pairs (Y_{i+1}, U_i) are mutually independent. Some small sample experiments will be carried out in the next section.

V. Numerical Examples

The results given in previous sections are asymptotic in nature. To get some insight into a small sample size behavior of the algorithms we make a few simulation experiments.

In the first experiment, we use the following dynamical system:

$$\left.\begin{aligned} X_{n+1} &= \begin{bmatrix} 0 & 1 \\ -0.03 & -0.4 \end{bmatrix} X_n + \begin{bmatrix} 1 \\ 1 \end{bmatrix} m(U_n) \\ Y_n &= \begin{bmatrix} 0,5 & 0.5 \end{bmatrix} X_n + \eta_n, \end{aligned}\right\} \tag{31}$$

where the transfer matrix is asymptotically stable since its eigenvalues are -0.1 and -0.3. The noise η_n and the input U_n are Gaussian $N(0; 0.1)$ and $N(0; 1)$, respectively.

Two models of the nonlinear characteristic have been chosen. The first one has a polynomial form

$$m_1(u) = u - 0.08u^3 + 0.1u^5.$$

The five level quantizer was selected as the second model, i.e.

$$m_2(u) = \begin{cases} 0, & \text{for } 0 < u \leq 0.2 \\ 0.75, & \text{for } 0.2 < u \leq 0.5 \\ 1, & \text{for } 0.5 < u \end{cases}$$

and

$$m_2(u) = -m_2(-u) \text{ for } u \leq 0.$$

We note that under these conditions $\alpha = 1$ and $\beta = 0$.

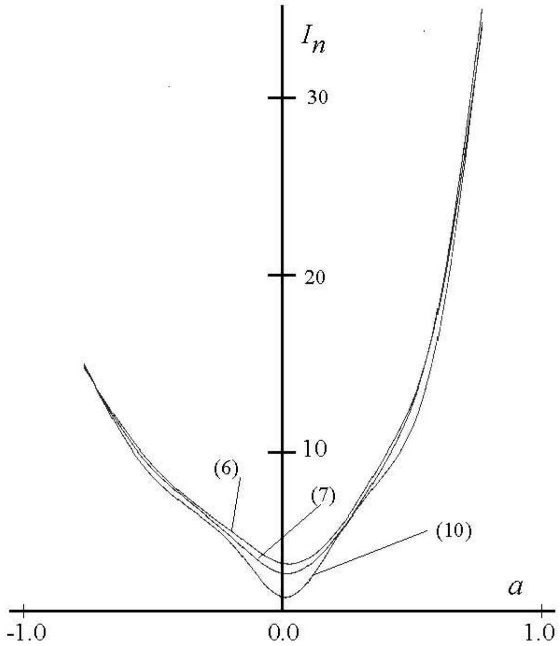

Figure 2: The mean integrated absolute error vs the parameter a for estimates (6), (7) and (10), $m = m_2$, $n = 30$.

The kernel from example (c) has been employed. Moreover, the smoothing parameter was selected experimentally as $h(n) = cn^{-\delta}$, where c and δ have been picked up to minimize the mean integrated absolute error

$$I_n = \int_{-1}^{1} |\hat{m}(u) - m(u)| \, f(u) du,$$

where $\hat{m}(u)$ is some estimate of the characteristic. The range of values of (c, δ) is the rectangular $[0.8, 2.2] \times [1/30, 1/3]$.

To evaluate the error, the integral was approximated numerically using a mesh over $(-1, 1)$ with increments of 0.01. Averaging the error obtained for the 50 repetitions gives an empirical value of I_n. Table I shows the values of I_n for m_1 when n varies from 10 to 80. The same is presented in Table II for the second nonlinear characteristic.

The rate of convergence for all estimates seems to be similar. However, it has been observed that the recursive procedures are less sensitive for the non-optimal selection of $h(n)$. When n varies from 10 to 100 the $h(n)$ minimizing I_n lies in the (c, δ) rectangular $[1.3; 1.4] \times [1/15; 1/6]$ for both recursive estimates. The corresponding rectangular for estimate (10) is $[1.3; 1.9] \times [1/15; 1/6]$. The analogical rectangulars in the case of m_2 are $[0.9; 1.0] \times [1/30; 4/15]$ and $[0.9; 1.4] \times [1/30; 3/10]$,

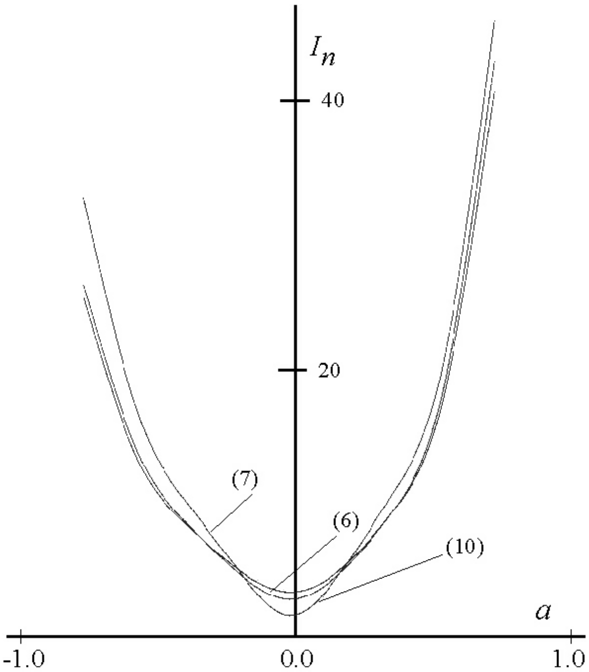

Figure 3: The mean integrated absolute error vs the parameter a for estimates (6), (7) and (10) with $h = 0.9n^{-1/30}$,, $m = m_2$, $n = 30$.

respectively. It is worthy to note that the rate resulting from Table I (polynomial model) is not better than that from Table II (m_2 characteristic).

This paradox (m_1 is much smoother than m_2) can be explained as follows. The rate of convergence for all estimates is not only limited by the proper selection of $h(n)$ but also by the applied kernel function. The kernel in (c) satisfies conditions (27) and (28). That is, it is well fitted to the characteristics which possess two bounded derivatives. Thus, kernel (c) is not able to distinguish m_2 from m_1. For characteristic m_1, which has five nonzero derivatives, instead of (27) and (28) one should take the kernel of order five, i.e.

$$\int v^i K(v)dv = 0, \ i = 1, \dots, 4$$

and

$$\int v^5 K(v)dv \neq 0.$$

The rate could be of order $O(n^{-5/11})$ which is clearly better than $O(n^{-2/5})$ proved in Section IV.

Generally, however, it is difficult to match the smoothness of the characteristic with the kernel order. Another solution of this problem will be suggested in

Table I: *The mean integrated absolute error* $(\times 100)$ *vs sample size for the* $m = m_1$

n	10	20	30	50	80
Estimate (6)	30.7	26.7	24.8	22.8	21.7
Estimate (7)	30.4	26.7	24.8	22.8	21.7
Estimate (10)	30.0	26.6	24.6	22.8	21.9

SectionVI.

The second experiment illustrates the dependence of the error on the system dynamics. That is, the system as in (31) is taken into account with $m = m_2$ and with the transition matrix of the form

$$A = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}.$$

Figure 2 plots the error versus a. The values of a are restricted to the stability region $(-1, 1)$. The number of the input–output data has been fixed to $n = 30$.

Surprisingly, the error exhibits asymmetrical behavior, i.e. the error increases faster for positive values of a, than for the negative ones. Moreover, the recursive estimates seem to have better performance for larger values of $|a|$, while the nonrecursive one does better for $|a| < 0.25$.

In the last experiment the values of (c, δ) have been selected optimally. It is interesting how much we lost selecting h sub-optimally. To do this the optimal h for $a = 0.0$ (memoryless system) has been fixed. This value for all estimates has been found equal to $0.9n^{-1/30}$. Other conditions are the same as in Experiment 2. Figure 3 plots the error versus a for that h.

VI. Concluding Remarks

In the paper, the recursive identification algorithms for the single input Hammerstein system have been studied. All results, however, remain in force if the multidimensional input signal $u = (u^{(1)}, \ldots, u^{(d)})$ is considered.

Table II: *The mean integrated absolute error* $(\times 100)$ *vs sample size for the* $m = m_2$

n	10	20	30	50	80
Estimate (6)	16.2	12.7	11.4	9.8	8.8
Estimate (7)	16.8	13.2	11.4	10.3	9.2
Estimate (10)	16.1	12.3	11.0	9.7	8.4

Then one can verify that Theorem I and Theorem II remain valid if (14) and (15) are replaced by

$$\frac{1}{n^2}\sum_{i=0}^{n}\frac{1}{h^d(i)} \xrightarrow{n} 0$$

and

$$\sum_{i=0}^{\infty}\frac{1}{h^d(i)} = \infty,$$

respectively.

However, the rate of convergence obtained in Section IV is altered, i.e., becomes $O(n^{-2/(4+d)})$. It is seen that the slower rate of convergence is achieved when the dimension is bigger. This "curse of dimensionality" can be partially overcome if some prior information about characteristic is available. In many applications $m(u^{(1)}, \ldots, u^{(d)})$ has the following additive form

$$\sum_{j=1}^{d} m_j(u^{(j)}),$$

where $m_j, j = 1, \ldots, d$ are real variable functions.

Even if m is not genuinely additive, an additive approximation of m may be sufficiently accurate. In any such case it is natural to conjecture that the algorithms achieve the rate for $d = 1$. Such a dimensionality reduction principle has been deeply examined, in the case of independent observations (i.e. when the dynamic subsystem degenerates to memoryless linear amplifier), by Friedman and Stuetzle (**38**), Stone (**30**), (**39**) and Hastie and Tibshirani (**40**). That problem, in the context of the Hammerstein system, is left for future work.

Furthermore, the consistency results were obtained assuming that the input probability density function exists. Using a technique worked out in Greblicki and Pawlak (**35**) one can extend the results for the case of any input distribution function. The latter may be important in some applications where, e.g., only random impulse inputs are admitted, see Krausz (**41**).

The simulation experiments presented in Section V reveal that it would be reasonable to consider an estimate of $m(u)$ being a mixture of the parametric and nonparametric identification methods. That is, for an adopted parametric model $m(u, \theta)$ of $m(u)$ one can estimate θ by the least-squares analysis, say, and then estimate $m(u)$ by $m(u, \hat{\theta})$. It is natural to assume that information on $m(u)$ may be incomplete and correction of $m(u, \hat{\theta})$ with the help of a. nonparametric estimate $\hat{m}(u)$ is necessary. Thus, the overall estimate of $m(u)$ is of the form $\alpha m(u, \hat{\theta}) + (1-\alpha)\hat{m}(u)$. Here, the number α reflects the importance of the proposed estimates, e.g., the parameter α is expected to be close to unity when the parametric model does better. The selection of α may be dome subjectively or it may be estimated from the observed data. One can conjecture that such an approach provides a better understanding of nonlinear system identification.

At last, the inherent recursive form of the examined algorithms may be helpful in the case of wandering characteristics, i.e. when we deal with a nonstationary environment.

References

(1) E. DeBoer, "Cross-correlation function of a Bandpass Nonlinear Network", *Proc. IEEE*, Vol. 64, pp. 1443-1446, 1976.

(2) R. R. Pfeiffer, "A model for two-tone inhibition of single Cochlear-nerve fibers", *Acoust. Soc. Am.*, Vol. 48, pp. 1373-1378, 1970.

(3) C. F. Hall and E. L. Hall, "A nonlinear model for the spatial characteristics of the human visual system", *IEEE Trans. System Man Cyber.*, Vol SMC-7, pp. 161-170, 1977.

(4) M. Korenberg, "Cross-correladon analysis of neutral cascades", in Proc. 10th Ann. Rocky Mountains Bioeng. Symp. (edited by D. Dick), pp. 47-52, 1973.

(5) A. A. Sawchuk and T. C. Strand. "Fourier optics in nonlinear image processing", in *Applications of Optical Fourier Transforms* (edited by H. Sark). pp. 371-429, Academic Press, New York, 1982.

(6) A. B. Martinez, P. F. Swaszek and J. B. Thomas, "Locally optimal detection in multivariate non-Gaussian noise", *IEEE Trans. Info. Theory*, Vol. IT-30, pp. 815- 822, 1984.

(7) M. C. King and B. F. Womack, "Discrete time adaptive control of linear dynamic systems with two-segment piecewise-linear asymmetric nonlinearity", *IEEE Trans. Aut. Control*, Vol. AC-29, pp. 170-173, 1984.

(8) T. C. Stapleton and S. C. Bass. "Adaptive noise cancellation for a class. of nonlinear dynamic reference channels", *IEEE Trans. Circuits Syst.*, Vol. CAS-32, pp. 143-150, 1985.

(9) K. S. Narendra and P. G. Gallman, "An iterative method for the identification of nonlinear systems using the Hammerstein model", *IEEE Trans. Aut. Control*. Vol. AC-11, pp. 546-550, 1966.

(10) S. A. Billings and S. Y. Fakhouri, "Non-linear system identification using the Hammerstein model", *Int. J. Systems Sci.*, Vol. 10, pp. 567-578, 1979.

(11) S. A. Billings and S. Y. Fakhouri, "Identification of a class of nonlinear systems using correlation analysis", *Proc. IEE*, Vol. 125, pp. 691-697, 1978.

(12) S. A. Billings and S. Y. Fakhouri, "Identification of systems containing linear dynamic and static nonlinear elements", *Automatica*, Vol. 18, pp. 15-26,1982.

(13) F. H. I. Chang and R. Luus, "A noniterative method for identification using the Hammerstein model", *IEEE Trans. Aut. Control*, Vol. AC-16, pp. 464-468, 1971.

(14) P. G. Gallman. "An iterative method for the identification of nonlinear systems using a Uryson model", *IEEE Trans. Aut. Control*, Vol. AC-20, pp. 771-175, 1975.

(15) P. G. Gallman. "A comparison of two Hammerstein model identification algorithms", *IEEE Trans. Aut. Control*, Vol. AC-21, pp. 124-126, 1976.

(16) M. A. L. Thatachar and S. Ramaswamy, "Identification of a class of nonlinear systems", *Int. J. Control*, Vol. 18, pp. 741-752, 1973.

(17) V. A. Kaminskas, "Parameter estimation of discrete systems of the Hammerstein class", *Automation Remote Control,*Vol. 36, pp. 1107-1113.1975.

(18) T. C. Hsia. "A multi-stage least squares method for identifying Hammerstein model nonlinear systems", in Proc. 1976 IEEE Cont. on Decision Control, pp. 934-938. 1976.

(19) P. Stoke. "On the convergence of an iterative algorithm used for Hammerstein system identification". *IEEE Trans. Aut. Control*, Vol. AC-26, pp. 967-969, 1981.

(20) P. Stoica and T. Soderstrem, "Instrumental-variable for identification of Hammerstein systems", *Int. J. Control*, Vol. 35, pp. 458-476, 1982.

(21) C. Kung and D. H. Shih. "Analysis and identification of Hammerstein and non-linear delay systems wing block-pulse function expansions", *Int. J. Control*, Vol. 43, pp. 139-147,1986.

(22) W. Sandberg, "Series expansions for nonlinear systems", *Circ. Syst. Signal Process*, Vol. 2, p. 77-87, 1983.

(23) W. Greblicki and M. Pawlak. "Identification of discrete Hammerstein systems using kernel regression estimates", *IEEE Trans. Aut. Control*, Vol. AC-31, pp. 74-77, 1986.

(24) W. Greblicki and M. Pawlak, "Hammerstein system identification by nonparametric regression estimation", *Int. J. Control*, Vol. 45, pp. 343-354, 1987.

(25) W. Greblicki and M. Pawlak, "Nonparametric identification of two-channel nonlinear systems", Proc. 25th IEEE Conf. on Decision and Control, Athens, Vol. 3, pp. 2012- 2015, 1986.

(26) T. W. Anderson, "The Statistical Analysis of Time Series", Wiley, New York, 1971.

(26) T. L. Doob, "Stochastic Processes", Wiley, New York, 1953.

(28) G. S. Watson, "Smooth regression analysis", Sankhyā, S. A., Vol. 26, pp. 335-372, 1964.

(29) E. A. Nadaraya. "On estimating regression", *Theor. Probability Appl.*, Vol. 9, pp. 141-142, 1964.

(30) C. J. Stone. "The dimensionality reduction principle for generalized additive models", *Ann. Statist.*, Vol. 14, p. 590-606, 1986.

(31) L. Devroye and T. J. Wagner. "On the L_1 convergence of kernel estimators of regression function with application in discrimination", *Z. Wahrsch. Verv. Gebiete*, Vol. 51. pp. 15-25, 1980.

(32) L. Devroye, "On the almost everywhere convergence of non-parametric regression function estimates", *Ann. Statist.*, Vol. 9. pp. 1310-1319, 1980.

(33) W. Greblicki, A. Krzyzak and M. Pawlak, "Distribution-free pointwise consistency of kernel regression estimate", *Ann. Statist.*, Vol. 12, pp. 1570-1575, 1984.

(34) W. Greblicki and M. Pawlak, "Necessary and sufficient conditions for Bars risk consistency of recursive kernel classification rule", *IEEE Trans. Info. Theory*, Vol. 33, pp. 408-412, 1987.
(35) W. Greblicki and M. Pawlak, "Necessary and sufficient conditions for a recursive kernel regression estimate", *J. Multivariate Analysis*, Vol. 23, pp. 67-76, 1987.
(36) M. Loéve, "Probability Theory", I, Springer, Berlin, 1977.
(37) N. Glick, "Consistency conditions for probability estimators and integrals for density-estimators", *Utilitas Math.*, Vol. 5, pp. 61-74, 1974.
(38) T. FL Friedman and W. Stuetzle, "Projection pursuit regression", *J. Am. Stat. Assoc.*, Vol. 76, pp. 817-823, 1981.
(39) C. J. Stone, "Additive regression and other nonparametric models", *Ann. Statist.*, Vol. 13, pp. 689-705, 1985.
(40) T. Hastie and R. Tibshirani, "Generalized Additive Models", *Statistical Sci.*, Vol. 1, pp. 297-318, 1986.
(41) H. I. Krausz, "Identification of nonlinear systems using random impulse train inputs", *Biol. Cybernetics*, Vol. 19, pp. 217-230, 1975.
(42) R. L. Wheeden and A. Zygmund, "Measure and Integral", Dekker, New York, 1977.

Appendix

Lemma 1
Let the bounded Borel kernel satisfy (11) and (12). Then

$$\lim_{h\to 0}\frac{1}{h}E\left\{Y_1K\left(\frac{u-U_0}{h}\right)\right\}=f(u)(\alpha m(u)+\beta)\int K(v)dv$$

and

$$\sup_{h>0}\frac{1}{h}E\left\{Y_1K\left(\frac{u-U_0}{h}\right)\right\}<\infty$$

for all $u\in C(f,m)$ and also for almost all $u\in R$.

The lemma is a simple consequence of Theorem 9.9 and Theorem 9.13 in Wheeden and Zygmund (**42**).

Lemma 2
Let K be a Borel integrable function. Then for $0\le i<j$,

$$\mathrm{cov}\,(X_{i+1}K_i,X_{j+1}K_j)=Q_{i,j}(u)EK_j,$$

where

$$\begin{aligned}Q_{i,j}(u)&=EK_iA^{j+1}P(A^{i+1})^T\\&\quad+\mathrm{cov}(K_i,W_i)A^{j-1}b(A^{i+1}\bar{X}_0)^T+E\left\{(W_i-EW_i)^2K_i\right\}A^{j-i}bb^T\\&\quad+\mathrm{var}(W)EK_i\sum_{r=0}^{i-1}A^{j-r}b(A^{i-r}b)^T+EW\mathrm{cov}(K_i,W_i)A^{j-i}b\sum_{r=0}^{i}(A^rb)^T,\end{aligned}$$

$$P = \mathrm{cov}(X_0, X_0) \text{ and } \bar{X}_0 = EX_0.$$

Proof: Let $\delta_n = W_n - EW_n$. Using (13), we have

$$\begin{aligned} X_{j+1} &= A^{j+1}X_0 + \sum_{s=0}^{j} A^s b \delta_{j-s} + EW \sum_{s=0}^{j} A^s b \\ &= G_{j1} + G_{j2} + G_{j3}, \text{ say.} \end{aligned}$$

Clearly $\mathrm{cov}(K_iG_{i1}, K_jG_{j3})$, $\mathrm{cov}(K_iG_{i2}, K_jG_{j1})$, $\mathrm{cov}(K_iG_{i2}, K_jG_{j3})$, $\mathrm{cov}(K_iG_{i3}, K_jG_{j1})$ and $\mathrm{cov}(K_iG_{i3}, K_jG_{j3})$ are equal to zero.

Consider $\mathrm{cov}(K_iG_{i2}, K_jG_{j3})$. Obviously

$$\begin{aligned} \mathrm{cov}\,(K_iG_{i2}, K_jG_{j3}) &= \sum_{s=0}^{j}\sum_{r=0}^{j} A^{j-s}b(A^{i-r}b)^T \mathrm{cov}(\delta_r K_i, \delta_s K_j) \\ &= \sum_{s=0}^{i}\sum_{r=0}^{i} A^{j-s}b(A^{i-r}b)^T \mathrm{cov}(\delta_r K_i, \delta_s K_j) \\ &= A^{j-i}bb^T EK_j \mathrm{cov}(\delta_i K_i, \delta_i) \\ &+ \sum_{r=0}^{i} A^{i-r}b(A^{i-r}b)^T EK_j EK_i E\delta_0^2. \end{aligned}$$

Since the remaining terms can be evaluated in the same way, the proof is completed. □