Available online at www.sciencedirect.com
SCIENCE DIRECT®

automatica

Automatica 41 (2005) 1479–1489

www.elsevier.com/locate/automatica

Identification of Hammerstein nonlinear ARMAX systems☆

Feng Ding[a,1], Tongwen Chen[b,*]

[a] *Control Science and Engineering Research Center, Southern Yangtze University, Wuxi 214122, China*
[b] *Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada T6G 2V4*

Received 6 July 2004; received in revised form 17 March 2005; accepted 22 March 2005
Available online 13 June 2005

Abstract

Two identification algorithms, an iterative least-squares and a recursive least-squares, are developed for Hammerstein nonlinear systems with memoryless nonlinear blocks and linear dynamical blocks described by ARMAX/CARMA models. The basic idea is to replace unmeasurable noise terms in the information vectors by their estimates, and to compute the noise estimates based on the obtained parameter estimates. Convergence properties of the recursive algorithm in the stochastic framework show that the parameter estimation error consistently converges to zero under the generalized persistent excitation condition. The simulation results validate the algorithms proposed.

Keywords: Recursive identification; Parameter estimation; Convergence properties; Stochastic gradient; Least squares; Hammersteinm models; Wiener models; Martingale convergence theorem

1. Introduction

Many nonlinear systems can be modelled by a Hammerstein model (linear time-invariant (LTI) block following some static nonlinear block), Wiener model (LTI block preceding some static nonlinear block), or Hammerstein–Wiener model (LTI block sandwiched by two static nonlinear blocks). Such models have been widely used in many areas, e.g., nonlinear filtering, actuator saturations, audio-visual processing, signal analysis, biologic systems, chemical processes. There exists a large amount of work on identification of these models, exploring different approaches and frameworks (e.g., Bai, 1998, 2002a,b, 2003, 2004; Haber & Unbehauen, 1990; Greblicki, 1997; Wigren & Nordsjö, 1999). For Hammerstein–Wiener models, Bai reported some interesting results: a two-stage identification algorithm based on the recursive least-squares and on the singular value decomposition (Bai, 1998) and a blind identification approach (Bai, 2002b).

This paper focuses on the identification of Hammerstein models shown in Fig. 1 which consists of a nonlinear memoryless element followed by a linear dynamical system (Narendra & Gallman, 1966; Pawlak, 1991; Ninness & Gibson, 2002; Bai, 2002a, 2004; Vörös, 2003), where the true output (namely, the noise-free output) $x(t)$ and the inner variable $\bar{u}(t)$ (namely, the output of the nonlinear block) are unmeasurable, $u(t)$ is the system input, $y(t)$ is the measurement of $x(t)$ but is corrupted by the disturbance $w(t)$, the output of $N(z)$ driven by an additive white noise $v(t)$ with zero mean, $G(z)$ is the transfer function of the linear part in the model, and $N(z)$ is the transfer function of the noise model. The nonlinear part in the Hammerstein model is a polynomial of a known order in the input as follows:

$$\bar{u}(t) = f(u(t)) = c_1u(t) + c_2u^2(t) + \cdots + c_mu^m(t),$$

☆ This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor Antonio Vicino under the direction of Editor Torsten Soederstroem.

* Corresponding author. Tel.: +1 780 492 3940; fax: +1 780 492 1811.
E-mail addresses: fding@sytu.edu.cn, fding@ece.ualberta.ca (F. Ding), tchen@ece.ualberta.ca (T. Chen).

[1] F. Ding is currently a Research Associate at the University of Alberta, Edmonton, Alberta, Canada T6G 2V4.

0005-1098/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2005.03.026

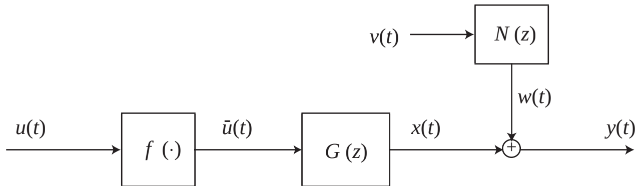

Fig. 1. The discrete-time SISO Hammerstein system.

or, more generally, a nonlinear function of a known basis $(\gamma_1, \gamma_2, \ldots, \gamma_m)$ as follows (Cerone & Regruto, 2003):

$$\bar{u}(t) = f(u(t)) = c_1\gamma_1(u(t)) + c_2\gamma_2(u(t)) + \cdots + c_m\gamma_m(u(t)) = \sum_{j=1}^{m} c_j\gamma_j(u(t)). \tag{1}$$

Existing identification methods for the Hammerstein models can be roughly divided into two categories: the non-parametric and parametric ones. Some non-parametric identification schemes assume that the linear part is an FIR or IIR model (Lang, 1994, 1997; Greblicki, 2002; Bai, 2003, 2004), i.e., the whole system is a simple nonlinear ARX model. Lang (1994, 1997) and Greblicki (2002) studied the convergence of correlation-technique-based identification algorithms of non-parametric Hammerstein models.

Earlier work on the parametric model identification of Hammerstein systems exists: Narendra and Gallman (1966) proposed an iterative algorithm which we refer to as the NG algorithm. Stoica (1981) showed that this algorithm may be divergent; but with normalizing the estimates at each iteration, it is convergent provided that the linear part is FIR and the input is white (Rangan, Wolodkin, & Poolla, 1995). However, the NG algorithm is not suitable for the general case with colored noise, non-FIR linear blocks, and any persistently exciting input. Haist, Chang, and Luus (1973) gave an iterative algorithm of identifying Hammerstein models with correlated noise, but no convergence analysis was carried out. Recently, Cerone and Regruto (2003) derived parameter bounds in the Hammerstein models with $N(z)=1$ in Fig. 1 by assuming that the output measurement error was bounded. To the best of our knowledge, most of the contributions assume that the systems under consideration are the nonlinear ARX models, or equation-error-like models (Narendra & Gallman, 1966; Nešić & Mareels, 1998; Wigren & Nordsjö, 1999; Bai, 1998, 2002a,b; Chang & Luus, 1971), and few address parametric model identification methods and their convergence for the Hammerstein nonlinear ARMAX systems with noises, which are the focus of this work. The main contribution of the paper is to propose iterative and recursive algorithms for parametric identification of general Hammerstein nonlinear ARMAX systems, and to study convergence properties of the recursive algorithm.

The half-substitution approach presented by Vörös (1995, 2003) may be used to study the identification problem of Hammerstein models; but there is no guarantee that the parameter estimates converge to the true parameters.

Briefly, the paper is organized as follows. Section 2 describes the problem formulation related to the Hammerstein nonlinear systems. Section 3 derives an iterative least squares algorithm for Hammerstein ARMAX systems with noise, and Section 4 develops a recursive least squares algorithm and analyzes its performance. Section 5 provides an illustrative example to show the effectiveness of the algorithms proposed. Finally, we offer some concluding remarks in Section 6.

2. Problem description

Assume that the linear dynamical block in Fig. 1 is described by an ARMAX/CARMA model, which has the following input–output relationship:

$$y(t) = x(t) + w(t), \tag{2}$$

$$x(t) = G(z)\,\bar{u}(t) = \frac{B(z)}{A(z)}\,\bar{u}(t), \tag{3}$$

$$w(t) = N(z)\,v(t) = \frac{D(z)}{A(z)}\,v(t). \tag{4}$$

Here $A(z)$, $B(z)$ and $D(z)$ are polynomials in the shift operator z^{-1} $[z^{-1}y(t) = y(t-1)]$ with

$$A(z) = 1 + a_1z^{-1} + a_2z^{-2} + \cdots + a_nz^{-n},$$
$$B(z) = b_1z^{-1} + b_2z^{-2} + b_3z^{-3} + \cdots + b_nz^{-n},$$
$$D(z) = 1 + d_1z^{-1} + d_2z^{-2} + \cdots + d_{n_d}z^{-n_d}.$$

Notice that in the characterization of the Hammerstein model shown in Fig. 1, $f(u)$ and $G(z)$ are actually not unique. Any pair $(\alpha f(u), G(z)/\alpha)$ for some nonzero and finite constant α would produce identical input and output measurements. In other words, any identification scheme cannot distinguish between $(f(u), G(z))$ and $(\alpha f(u), G(z)/\alpha)$. Therefore, to get a unique parameterization, without loss of generality, one of the gains of $f(u)$ and $G(z)$ has to be fixed. There are several ways to normalize the gains (Bai, 1998; Gallman, 1976, Cerone & Regruto, 2003). We adopt the following:

Assumption 1. The first coefficient of the function $f(\cdot)$ equals 1 (Bai, 2002b; Gallman, 1976); i.e., in (1), $c_1 = 1$.

From (1)–(4), we obtain a Hammerstein nonlinear ARMAX (HARMAX) model:

$$A(z)y(t) = B(z)\bar{u}(t) + D(z)v(t),$$
$$\bar{u}(t) = f(u(t)) = c_1\gamma_1(u(t)) + c_2\gamma_2(u(t)) + \cdots + c_m\gamma_m(u(t)). \tag{5}$$

The objective of this paper is to present identification algorithms to estimate the system parameters a_i, b_i, c_i and

d_i of the nonlinear ARMAX model by using the available input–output data $\{u(t), y(t)\}$, and to study the properties of the algorithms involved.

3. The iterative algorithm

Let us introduce some notation first. The symbol I stands for an identity matrix of appropriate sizes; the superscript T denotes the matrix transpose; $|X| = \det[X]$ represents the determinant of the matrix X; the norm of a matrix X is defined by $\|X\|^2 = \mathrm{tr}[XX^{\mathrm{T}}]$; $\lambda_{\max}[X]$ and $\lambda_{\min}[X]$ represent the maximum and minimum eigenvalues of X, respectively; $f(t) = o(g(t))$ represents $f(t)/g(t) \to 0$ as $t \to \infty$; for $g(t) \geqslant 0$, we write $f(t) = O(g(t))$ or $f(t) \sim g(t)$ if there exists a positive constant δ_1 such that $|f(t)| \leqslant \delta_1 g(t)$.

From (5), we easily get the following recursive equation:

$$\begin{aligned} y(t) &= -\sum_{i=1}^{n} a_i y(t-i) + \sum_{i=1}^{n} b_i \bar{u}(t-i) \\ &\quad + \sum_{i=1}^{n_d} d_i v(t-i) + v(t), \\ &= -\sum_{i=1}^{n} a_i y(t-i) + \sum_{i=1}^{n} b_i \sum_{j=1}^{m} c_j \gamma_j(u(t-i)) \\ &\quad + \sum_{i=1}^{n_d} d_i v(t-i) + v(t). \end{aligned}$$

Define the parameter vector θ and information vector $\varphi_0(t)$ as

$$\theta = \begin{bmatrix} \mathbf{a} \\ c_1\mathbf{b} \\ c_2\mathbf{b} \\ \vdots \\ c_m\mathbf{b} \\ \mathbf{d} \end{bmatrix} \in \mathbb{R}^{n_0}, \quad \varphi_0(t) = \begin{bmatrix} \psi(t) \\ v(t-1) \\ v(t-2) \\ \vdots \\ v(t-n_d) \end{bmatrix} \in \mathbb{R}^{n_0},$$
$$n_0 := (m+1)n + n_d, \tag{6}$$

$$\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \in \mathbb{R}^{n}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \in \mathbb{R}^{n}, \quad \mathbf{d} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n_d} \end{bmatrix} \in \mathbb{R}^{n_d},$$
$$\psi(t) = \begin{bmatrix} \psi_0(t) \\ \psi_1(t) \\ \psi_2(t) \\ \vdots \\ \psi_m(t) \end{bmatrix} \in \mathbb{R}^{n_0}, \quad \psi_0(t) = \begin{bmatrix} -y(t-1) \\ -y(t-2) \\ \vdots \\ -y(t-n) \end{bmatrix} \in \mathbb{R}^{n}, \tag{7}$$

$$\psi_j(t) = \begin{bmatrix} \gamma_j(u(t-1)) \\ \gamma_j(u(t-2)) \\ \vdots \\ \gamma_j(u(t-n)) \end{bmatrix} \in \mathbb{R}^{n}, \quad j = 1, 2, \ldots, m. \tag{8}$$

Then we have

$$y(t) = \varphi_0^{\mathrm{T}}(t)\theta + v(t). \tag{9}$$

Note that $\psi(t)$ in $\varphi_0(t)$ is available but $v(t-i)$, $i = 1, 2, \ldots, n_d$, in $\varphi_0(t)$ are unavailable. Let $\hat{\theta}$ denote the estimate of θ. Since $v(t)$ is a "white" noise with zero mean, then

$$\hat{y}(t) = \varphi_0^{\mathrm{T}}(t)\hat{\theta}$$

is the best output prediction. Consider the quadratic output prediction error criterion

$$\begin{aligned} J(\hat{\theta}) &:= \sum_{i=t-p+1}^{t} [y(i) - \hat{y}(i)]^2 \\ &= \sum_{i=t-p+1}^{t} [y(i) - \varphi_0^{\mathrm{T}}(i)\hat{\theta}]^2. \end{aligned} \tag{10}$$

Here, p may be known as the data length ($p \gg n_0$). The quadratic error function in (10) is one of the most common cost functions in the identification literature (Söderström & Stoica, 1989; Ljung, 1999). Many well-known Hammerstein model identification methods (Narendra & Gallman, 1966; Chang & Luus, 1971; Bai, 2002a,b) belong to this class and the differences lie only in the formulation of the information vector $\varphi_0(t)$. Let

$$Y(t) = \begin{bmatrix} y(t) \\ y(t-1) \\ \vdots \\ y(t-p+1) \end{bmatrix}, \quad \Phi_0(t) = \begin{bmatrix} \varphi_0^{\mathrm{T}}(t) \\ \varphi_0^{\mathrm{T}}(t-1) \\ \vdots \\ \varphi_0^{\mathrm{T}}(t-p+1) \end{bmatrix}. \tag{11}$$

Hence

$$\begin{aligned} J(\hat{\theta}) &= [Y(t) - \Phi_0(t)\hat{\theta}]^{\mathrm{T}}[Y(t) - \Phi_0(t)\hat{\theta}] \\ &= \|Y(t) - \Phi_0(t)\hat{\theta}\|^2. \end{aligned}$$

Provided that $\varphi_0(t)$ is persistently exciting, minimizing $J(\hat{\theta})$ gives the least-squares estimate:

$$\hat{\theta} = [\Phi_0^{\mathrm{T}}(t)\Phi_0(t)]^{-1}\Phi_0^{\mathrm{T}}(t)Y(t). \tag{12}$$

However, a difficulty arises because $v(t-i)$ is in $\varphi_0(t)$, thus $\Phi_0(t)$ in the expression on the right-hand side of (12) contains unknown noise terms $v(t-i)$, $i = 1, 2, \ldots, n_d$; so it is impossible to compute the estimate $\hat{\theta}$ by (12). Our approach is based on the iterative identification principle: Let $k = 1, 2, 3, \ldots$, the unknown variables $v(t-i)$ are replaced by their corresponding estimate $\hat{v}_k(t-i)$ at iteration k, and $\varphi_0(t)$ are replaced by $\hat{\varphi}_k(t)$. Let $\hat{\theta}_k$ be the iterative solution of θ. Thus, from (9), the estimate of $v(t)$ is given by

$$\hat{v}_k(t-i) = y(t-i) - \hat{\varphi}_{k-1}^{\mathrm{T}}(t-i)\hat{\theta}_{k-1} \tag{13}$$

with

$$\hat{\varphi}_k(t) = \begin{bmatrix} \psi(t) \\ \hat{v}_k(t-1) \\ \hat{v}_k(t-2) \\ \vdots \\ \hat{v}_k(t-n_d) \end{bmatrix} \in \mathbb{R}^{n_0}. \tag{14}$$

Let

$$\Phi_k(t) = \begin{bmatrix} \hat{\varphi}_k^{\mathrm{T}}(t) \\ \hat{\varphi}_k^{\mathrm{T}}(t-1) \\ \vdots \\ \hat{\varphi}_k^{\mathrm{T}}(t-p+1) \end{bmatrix}. \tag{15}$$

Based on (12) and replacing $\Phi_0(t)$ by $\Phi_k(t)$, the iterative solution $\hat{\theta}_k$ of θ may be also computed by

$$\hat{\theta}_k = [\Phi_k^{\mathrm{T}}(t)\Phi_k(t)]^{-1}\Phi_k^{\mathrm{T}}(t)Y(t), \quad k = 1, 2, 3, \ldots. \tag{16}$$

We refer to Eqs. (13)–(16) as the least-squares iterative identification algorithm for HARMAX systems, or HARMAX-LSI algorithm for short.

To initialize the algorithm in (13)–(16), we take $\hat{\theta}_0 = \mathbf{0}$ or some small real vector, e.g., $\hat{\theta}_0 = 10^{-6}\mathbf{1}_{n_0}$ with $\mathbf{1}_{n_0}$ being an n_0-dimensional column vector whose elements are 1.

The HARMAX-LSI algorithm employs the idea of updating the estimate $\hat{\theta}$ using a fixed data batch with a finite length p. Though this algorithm is important, we think, for *finite data measurement*, the convergence analysis is very challenging and is yet to be developed.

In this paper, in order to distinguish on-line from off-line calculation, we use *iterative* with subscript k, e.g., $\hat{\theta}_k$, for off-line algorithms, and *recursive* with no subscript, e.g., $\hat{\theta}(t)$ to be given later, for on-line ones. We imply that a recursive algorithm can be on-line implemented, but an iterative one cannot.

Under Assumption 1 with $c_1 = 1$, the estimates $\hat{\mathbf{a}} = [\hat{a}_1\ \hat{a}_2\ \cdots\ \hat{a}_n]^{\mathrm{T}}$, $\hat{\mathbf{b}} = [\hat{b}_1\ \hat{b}_2\ \cdots\ \hat{b}_n]^{\mathrm{T}}$ and $\hat{\mathbf{d}} = [\hat{d}_1\ \hat{d}_2\ \cdots\ \hat{d}_{n_d}]^{\mathrm{T}}$ of $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{d}$ can be read from the first, second n entries and last n_d entries of $\hat{\theta}$, respectively. Let $\hat{\theta}_i$ be the ith element of $\hat{\theta}$, referring to the definition of θ, then the estimates of c_j, $j = 2, 3, \ldots, m$, may be computed by

$$\hat{c}_j = \frac{\hat{\theta}_{jn+i}}{\hat{b}_i}, \quad j = 2, 3, \ldots, m; \quad i = 1, 2, \ldots, n.$$

From here, we can see that there is a large amount of redundancy in the establishment of each coefficient $\hat{c}_j$ in the nonlinear part since for each c_j we have n estimates $\hat{c}_j$ for $i = 1, 2, \ldots, n$. Since we do not need such n estimates $\hat{c}_j$, one way is to take their average as the estimate of c_j, i.e.,

$$\hat{c}_j = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{\theta}_{jn+i}}{\hat{b}_i}, \quad j = 2, 3, \ldots, m.$$

4. The recursive algorithm

As is pointed out in the preceding section that the HARMAX-LSI algorithm uses batch data identification and is not suitable for on-line identification. Moreover, the drawback is that it requires computing matrix inversion at each step. In this section, we derive a recursive identification algorithm which can be on-line implemented. For a recursive algorithm, new information (input and/or output data) is always used in the algorithm which recursively computes the parameter estimates every step as t increases.

Define

$$V(t) = \begin{bmatrix} v(t) \\ v(t-1) \\ \vdots \\ v(t-p+1) \end{bmatrix}.$$

From (9) and (11), we have

$$Y(t) = \Phi_0(t)\theta + V(t). \tag{17}$$

Since $V(t)$ is a "white" noise vector, the recursive least-squares algorithm can give the unbiased estimation of θ in (17). As in the preceding section, the unknowns $v(t-i)$, $i = 1, 2, \ldots, n$, in the matrix $\Phi_0(t)$ are replaced by their estimates $\hat{v}(t-i)$. Let $\hat{\theta}(t)$ denote the estimate of θ at time t, and use $\varphi(t)$ as $\varphi_0(t)$ and $\Phi(t)$ as $\Phi_0(t)$, then according to the least squares principle (Ljung, 1999), it is not difficult to get the following recursive least squares algorithm of estimating θ based on the noise estimation:

$$\hat{\theta}(t) = \hat{\theta}(t-1) + P(t)\Phi^{\mathrm{T}}(t)[Y(t) - \Phi(t)\hat{\theta}(t-1)],$$
$$P^{-1}(t) = P^{-1}(t-1) + \Phi^{\mathrm{T}}(t)\Phi(t),$$

$$\Phi(t) = \begin{bmatrix} \varphi^{\mathrm{T}}(t) \\ \varphi^{\mathrm{T}}(t-1) \\ \vdots \\ \varphi^{\mathrm{T}}(t-p+1) \end{bmatrix}, \quad \varphi(t) = \begin{bmatrix} \psi(t) \\ \hat{v}(t-1) \\ \hat{v}(t-2) \\ \vdots \\ \hat{v}(t-n_d) \end{bmatrix} \in \mathbb{R}^{n_0},$$

$$\hat{v}(t) = y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t).$$

As $p = 1$, $e(t) := y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t-1) \in \mathbb{R}^1$ is called the innovation (Ljung, 1999), and $E(t) := Y(t) - \Phi(t)\hat{\theta}(t-1) \in \mathbb{R}^p$ may be referred to as the innovation vector. Thus, p here may be also known as the innovation length. When $p = 1$, we obtain a simple recursive least squares algorithm based on the noise estimation:

$$\hat{\theta}(t) = \hat{\theta}(t-1) + P(t)\varphi(t)[y(t) - \varphi(t)\hat{\theta}(t-1)], \tag{18}$$

$$P^{-1}(t) = P^{-1}(t-1) + \varphi(t)\varphi^{\mathrm{T}}(t), \quad P(0) = p_0 I, \tag{19}$$

$$\varphi(t) = \begin{bmatrix} \psi(t) \\ \hat{v}(t-1) \\ \hat{v}(t-2) \\ \vdots \\ \hat{v}(t-n_d) \end{bmatrix} \in \mathbb{R}^{n_0}, \tag{20}$$

$$\psi(t) = \begin{bmatrix} \psi_0(t) \\ \psi_1(t) \\ \psi_2(t) \\ \vdots \\ \psi_m(t) \end{bmatrix} \in \mathbb{R}^{n_0}, \quad \psi_0(t) = \begin{bmatrix} -y(t-1) \\ -y(t-2) \\ \vdots \\ -y(t-n) \end{bmatrix} \in \mathbb{R}^n, \tag{21}$$

$$\psi_j(t) = \begin{bmatrix} \gamma_j(u(t-1)) \\ \gamma_j(u(t-2)) \\ \vdots \\ \gamma_j(u(t-n)) \end{bmatrix} \in \mathbb{R}^n, \quad j = 1, 2, \ldots, m, \tag{22}$$

$$\hat{v}(t) = y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t). \tag{23}$$

To initialize the algorithm, we take p_0 to be a large positive real number, e.g., $p_0 = 10^6$, and $\hat{\theta}(0) = 10^{-6}\mathbf{1}_{n_0 \times 1}$.

We assume that $\{v(t), \mathscr{F}_t\}$ is a martingale difference vector sequence defined on a probability space $\{\Omega, \mathscr{F}, P\}$, where $\{\mathscr{F}_t\}$ is the σ algebra sequence generated by $\{v(t)\}$, i.e., $\mathscr{F}_t = \sigma(v(t), v(t-1), v(t-2), \ldots,)$ or $\mathscr{F}_t = \sigma(y(t), y(t-1), y(t-2), \ldots,)$ for the deterministic input sequence $\{u(t)\}$. We make the following assumptions on the noise sequence $\{v(t)\}$ (Goodwin & Sin, 1984):

(~~A1~~) $\mathrm{E}[v(t)|\mathscr{F}_{t-1}] = 0, \quad$ a.s.;

(~~A2~~) $\mathrm{E}[v^2(t)|\mathscr{F}_{t-1}] = \sigma_v^2(t) \leqslant \bar{\sigma}_v^2 < \infty, \quad$ a.s.

That is, $\{v(t)\}$ is a stochastic noise with zero mean and bounded time-varying variances. [Thus the system in (9) may be non-stationary.] The following lemmas are required to establish the main convergence results.

Lemma 1. *For the algorithm in* (18)–(23), *for any* $\beta > 1$, *the covariance matrix* $P(t)$ *in* (19) *satisfies the following inequality*:

$$\sum_{i=1}^{\infty} \frac{\varphi^{\mathrm{T}}(i)P(i)\varphi(i)}{[\ln|P^{-1}(t)|]^{\beta}} < \infty, \quad \text{a.s.}$$

Proof. From the definition of $P(t)$ in (19), we have

$$\begin{aligned} P^{-1}(t-1) &= P^{-1}(t) - \varphi(t)\varphi^{\mathrm{T}}(t) \\ &= P^{-1}(t)[I - P(t)\varphi(t)\varphi^{\mathrm{T}}(t)]. \end{aligned}$$

Taking determinants on both sides and using the equality, $|I + EF| = |I + FE|$, give

$$\begin{aligned} |P^{-1}(t-1)| &= |P^{-1}(t)|\,|I - P(t)\varphi(t)\varphi^{\mathrm{T}}(t)| \\ &= |P^{-1}(t)|\,[1 - \varphi^{\mathrm{T}}(t)P(t)\varphi(t)]. \end{aligned}$$

Thus

$$\varphi^{\mathrm{T}}(t)P(t)\varphi(t) = \frac{|P^{-1}(t)| - |P^{-1}(t-1)|}{|P^{-1}(t)|}.$$

Dividing $[\ln|P^{-1}(t)|]^{\beta}$ and summing for t from 1 to ∞ yield (noting that $|P^{-1}(t)|$ is a non-decreasing function of t)

$$\begin{aligned} &\sum_{t=1}^{\infty} \frac{\varphi^{\mathrm{T}}(t)P(t)\varphi(t)}{[\ln|P^{-1}(t)|]^{\beta}} \\ &= \sum_{t=1}^{\infty} \frac{|P^{-1}(t)| - |P^{-1}(t-1)|}{|P^{-1}(t)|[\ln|P^{-1}(t)|]^{\beta}} \\ &= \sum_{t=1}^{\infty} \int_{|P^{-1}(t-1)|}^{|P^{-1}(t)|} \frac{\mathrm{d}x}{|P^{-1}(t)|[\ln|P^{-1}(t)|]^{\beta}} \\ &\leqslant \sum_{t=1}^{\infty} \int_{|P^{-1}(t-1)|}^{|P^{-1}(t)|} \frac{\mathrm{d}x}{x[\ln x]^{\beta}} = \int_{|P^{-1}(0)|}^{|P^{-1}(\infty)|} \frac{\mathrm{d}x}{x[\ln x]^{\beta}} \\ &= \frac{-1}{\beta - 1} \frac{1}{[\ln x]^{\beta-1}} \Bigg|_{|P^{-1}(0)|}^{|P^{-1}(\infty)|} \\ &= \frac{1}{\beta - 1}\left[\frac{1}{[\ln|P^{-1}(0)|]^{\beta-1}} - \frac{1}{[\ln|P^{-1}(\infty)|]^{\beta-1}}\right] \\ &< \infty, \quad \text{a.s.,} \quad \text{for any} \quad \beta > 1. \end{aligned}$$

This proves Lemma 1. $\square$

Next, we study the properties of this algorithm. Define

$$\tilde{\theta}(t) := \hat{\theta}(t) - \theta, \tag{24}$$

$$\begin{aligned} &e(t) := y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t-1), \\ &P_0^{-1}(t) := P_0^{-1}(t-1) + \varphi_0(t)\varphi_0^{\mathrm{T}}(t), \quad P(0) = p_0 I, \\ &r(t) := \mathrm{tr}[P^{-1}(t)], \quad r_0(t) := \mathrm{tr}[P_0^{-1}(t)], \\ &W(t) := \tilde{\theta}^{\mathrm{T}}(t)P^{-1}(t)\tilde{\theta}(t). \end{aligned}$$

It follows easily that

$$|P^{-1}(t)| \leqslant r^n(t), \quad r(t) \leqslant n_0 \lambda_{\max}[P^{-1}(t)], \tag{25}$$

$$\begin{aligned} &\ln|P^{-1}(t)| = O(\ln r(t)) = O(\ln \lambda_{\max}[P^{-1}(t)]), \\ &\ln|P_0^{-1}(t)| = O(\ln r_0(t)) = O(\ln \lambda_{\max}[P_0^{-1}(t)]), \end{aligned}$$

$$\begin{aligned} \hat{v}(t) &= [1 - \varphi^{\mathrm{T}}(t)P(t)\varphi(t)]e(t) \\ &= \frac{e(t)}{1 + \varphi^{\mathrm{T}}(t)P(t-1)\varphi(t)}, \end{aligned} \tag{26}$$

$$\|\tilde{\theta}(t)\|^2 \leqslant \frac{\mathrm{tr}[\tilde{\theta}^{\mathrm{T}}(t)P^{-1}(t)\tilde{\theta}(t)]}{\lambda_{\min}[P^{-1}(t)]} = \frac{W(t)}{\lambda_{\min}[P^{-1}(t)]}. \tag{27}$$

Lemma 2. *For the system in* (9) *and the algorithm in* (18)–(23), *assume that* (A1) *and* (A2) *hold, and*

(~~A3~~) $H(z) := D^{-1}(z) - \frac{1}{2}$ *is strictly positive real.*

Then

$$\begin{aligned} \mathrm{E}[W(t) + S(t)|\mathscr{F}_{t-1}] \leqslant\ & W(t-1) + S(t-1) \\ &+ 2\varphi^{\mathrm{T}}(t)P(t)\varphi(t)\bar{\sigma}_v^2, \text{ a.s.} \end{aligned} \tag{28}$$

where

$$S(t) = 2\sum_{i=1}^{t} \tilde{u}(i)\tilde{y}(i),$$

$$\tilde{y}(t) = \tfrac{1}{2}\tilde{\theta}^{\mathrm{T}}(t)\varphi(t) + [y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t) - v(t)], \tag{29}$$

$$\tilde{u}(t) = -\tilde{\theta}^{\mathrm{T}}(t)\varphi(t). \tag{30}$$

Here, (A3) *guarantees that* $S(t) \geqslant 0$.

Proof. Substituting (18) into (24) and using (26) give

$$\begin{aligned}\tilde{\theta}(t) &= \tilde{\theta}(t-1) + P(t)\varphi(t)e(t)\\ &= \tilde{\theta}(t-1) + P(t-1)\varphi(t)\hat{v}(t).\end{aligned} \tag{31}$$

Or

$$P^{-1}(t-1)\tilde{\theta}(t) = P^{-1}(t-1)\tilde{\theta}(t-1) + \varphi(t)\hat{v}(t).$$

Pre-multiplying $\tilde{\theta}^{\mathrm{T}}(t)$ and using (31) yield

$$\begin{aligned}&\tilde{\theta}^{\mathrm{T}}(t)P^{-1}(t-1)\tilde{\theta}(t)\\ &\quad= \tilde{\theta}^{\mathrm{T}}(t)P^{-1}(t-1)\tilde{\theta}(t-1) + \tilde{\theta}^{\mathrm{T}}(t)\varphi(t)\hat{v}(t)\\ &\quad= [\tilde{\theta}(t-1) + P(t-1)\varphi(t)\hat{v}(t)]^{\mathrm{T}}\\ &\qquad\times P^{-1}(t-1)\tilde{\theta}(t-1) + \varphi^{\mathrm{T}}(t)\tilde{\theta}(t)\hat{v}(t)\\ &\quad= \tilde{\theta}^{\mathrm{T}}(t-1)P^{-1}(t-1)\tilde{\theta}(t-1)\\ &\qquad+ \varphi^{\mathrm{T}}(t)\tilde{\theta}(t-1)\hat{v}(t) + \varphi^{\mathrm{T}}(t)\tilde{\theta}(t)\hat{v}(t).\end{aligned}$$

By using (19), (26) and (31), it follows that

$$\begin{aligned}W(t) &= W(t-1) + [\varphi^{\mathrm{T}}(t)\tilde{\theta}(t)]^2\\ &\qquad+ \varphi^{\mathrm{T}}(t)\tilde{\theta}(t-1)\hat{v}(t) + \varphi^{\mathrm{T}}(t)\tilde{\theta}(t)\hat{v}(t)\\ &= W(t-1) + [\varphi^{\mathrm{T}}(t)\tilde{\theta}(t)]^2 + \varphi^{\mathrm{T}}(t)\\ &\qquad\times [\tilde{\theta}(t) - P(t)\varphi(t)e(t)]\hat{v}(t) + \varphi^{\mathrm{T}}(t)\tilde{\theta}(t)\hat{v}(t)\\ &= W(t-1) + [\varphi^{\mathrm{T}}(t)\tilde{\theta}(t)]^2\\ &\qquad+ 2\varphi^{\mathrm{T}}(t)\tilde{\theta}(t)\hat{v}(t) - \varphi^{\mathrm{T}}(t)P(t)\varphi(t)\hat{v}(t)e(t)\\ &= W(t-1) + [\varphi^{\mathrm{T}}(t)\tilde{\theta}(t)]^2 + 2\varphi^{\mathrm{T}}(t)\tilde{\theta}(t)\hat{v}(t)\\ &\qquad- \varphi^{\mathrm{T}}(t)P(t)\varphi(t)[1 - \varphi^{\mathrm{T}}(t)P(t)\varphi(t)]e^2(t)\\ &\leqslant W(t-1) + [\varphi^{\mathrm{T}}(t)\tilde{\theta}(t)]^2 + 2\varphi^{\mathrm{T}}(t)\tilde{\theta}(t)\hat{v}(t)\\ &= W(t-1) + 2\varphi^{\mathrm{T}}(t)\tilde{\theta}(t)\left[\tfrac{1}{2}\tilde{\theta}^{\mathrm{T}}(t)\varphi(t)\right.\\ &\qquad\left.+[\hat{v}(t) - v(t)]\right] + 2\varphi^{\mathrm{T}}(t)\tilde{\theta}(t)v(t)\\ &= W(t-1) - 2\tilde{u}^{\mathrm{T}}(t)\tilde{y}(t) + 2\varphi^{\mathrm{T}}(t)[\tilde{\theta}(t-1)\\ &\qquad+ P(t)\varphi(t)e(t)]v(t)\\ &= W(t-1) - 2\tilde{u}^{\mathrm{T}}(t)\tilde{y}(t) + 2\varphi^{\mathrm{T}}(t)\tilde{\theta}(t-1)v(t)\\ &\qquad+ 2\varphi^{\mathrm{T}}(t)P(t)\varphi(t)\{[e(t) - v(t)]v(t) + v^2(t)\}.\end{aligned}$$

Since $S(t-1)$, $W(t-1)$, $\varphi^{\mathrm{T}}(t)\tilde{\theta}(t-1)$ and $\varphi^{\mathrm{T}}(t)P(t)\varphi(t)[e(t) - v(t)]$ are uncorrelated with $v(t)$ and are $\mathscr{F}_{t-1}$ measurable, adding $S(t)$, taking the conditional expectation with respect to $\mathscr{F}_{t-1}$ and using (A1) and (A2) lead to (28). Next, we show $S(t) \geqslant 0$. Since

$$\begin{aligned}D(z)[\hat{v}(t) - v(t)] &= D(z)\hat{v}(t) - A(z)y(t) + B(z)u(t)\\ &= \hat{v}(t) - y(t) + \varphi^{\mathrm{T}}(t)\theta\\ &= -\varphi^{\mathrm{T}}(t)\hat{\theta}(t) + \varphi^{\mathrm{T}}(t)\theta\\ &= -\varphi^{\mathrm{T}}(t)[\hat{\theta}(t) - \theta]\\ &= -\varphi^{\mathrm{T}}(t)\tilde{\theta}(t) = \tilde{u}(t),\end{aligned} \tag{32}$$

from (29), (30) and (32), we have

$$\begin{aligned}\tilde{y}(t) &= \frac{1}{2}\tilde{\theta}^{\mathrm{T}}(t)\varphi(t) + [y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t) - v(t)]\\ &= \frac{1}{2}\tilde{\theta}^{\mathrm{T}}(t)\varphi(t) + [\hat{v}(t) - v(t)]\\ &= -\frac{1}{2}\tilde{u}(t) + D^{-1}(z)\tilde{u}(t)\\ &= \left[D^{-1}(z) - \frac{1}{2}\right]\tilde{u}(t) = H(z)\tilde{u}(t)\\ &= \left[D^{-1}(z) - \frac{1+\rho}{2}\right]\tilde{u}(t) + \frac{\rho}{2}\tilde{u}(t)\\ &=: \tilde{y}_1(t) + \frac{\rho}{2}\tilde{u}(t),\end{aligned}$$

where

$$\tilde{y}_1(t) = H_1(z)\tilde{u}(t), \quad H_1(z) = D^{-1}(z) - \frac{1+\rho}{2}.$$

Here, $\tilde{y}_1(t)$ may be regarded as the output of the linear system $H_1(z)$ driven by $\hat{v}(t) - v(t)$. Since $H(z)$ is strictly positive real, there exists a small constant $\rho > 0$ such that $H_1(z)$ is (also strictly) positive real. Referring to Appendix C in (Goodwin & Sin, 1984), we have

$$2\sum_{i=1}^{t}\tilde{u}(i)\tilde{y}_1(i) \geqslant 0, \quad \text{a.s.},$$

$$S(t) = 2\sum_{i=1}^{t}\tilde{u}^{\mathrm{T}}(i)\tilde{y}_1(i) + \rho\sum_{i=1}^{t}\tilde{u}^2(i) \geqslant 0, \ \text{a.s.} \tag{33}$$

This proves Lemma 2. □

The positive real Assumption (A3) depends on the noise model parameters which are unknown—a difficulty in validation. However, such conditions are sufficient technical assumptions often found in convergence analysis of linear or nonlinear identification problems. How to find some weaker and more practical condition for convergence of estimation algorithms has still been an open topic. An alternative way in the linear case is to filter the input–output data by choosing a known polynomial $F(z)$ so that (A3) becomes

$$\cancel{(\mathrm{A3}')}\ \frac{F(z)}{D(z)} - \frac{1}{2} \text{ is strictly positive real}$$

(Goodwin & Sin, 1984).

However, how to validate (A3′) is still a problem due to unknown $D(z)$. A possible method is to use the algorithm to obtain an estimated noise model $\hat{D}(z)$, and validate Assumption (A3) based on $\hat{D}(z)$ instead of $D(z)$. In our simulations, the algorithm works quite well for stable $D(z)$.

Theorem 1. *For the system in* (9) *and the algorithm in* (18)–(23), *assume that the conditions in Lemma* 2 *hold. Then for any* $\beta > 1$, *we have*

$$\|\hat{\theta}(t) - \theta\|^2 = O\left(\frac{\{\ln\lambda_{\max}[P_0^{-1}(t)]\}^{\beta}}{\lambda_{\min}[P_0^{-1}(t)]}\right), \quad \text{a.s.}$$

Proof. Let

$$W_1(t)=\frac{W(t)+S(t)}{[\ln|P^{-1}(t)|]^{\beta}},\quad \beta>1.$$

Since $\ln|P^{-1}(t)|$ is nondecreasing, using Lemma 2 gives

$$\begin{aligned}\mathrm{E}[W_1(t)|\mathscr{F}_{t-1}]&\leqslant\frac{W(t-1)+S(t-1)}{[\ln|P^{-1}(t)|]^{\beta}}\\&\quad+\frac{2\varphi^{\mathrm{T}}(t)P(t)\varphi(t)}{[\ln|P^{-1}(t)|]^{\beta}}\bar{\sigma}_v^2\\&\leqslant W_1(t-1)+\frac{2\varphi^{\mathrm{T}}(t)P(t)\varphi(t)}{[\ln|P^{-1}(t)|]^{\beta}}\bar{\sigma}_v^2,\ \text{a.s.}\end{aligned}$$

Using Lemma 1, we can see that the sum of the right-hand last term for t from 1 to ∞ is finite. Now applying the martingale convergence theorem (Lemma D.5.3 in Goodwin & Sin, 1984) to the above inequality, we conclude that $W_1(t)$ converges a.s. to a finite random variable, say, W_1; i.e.,

$$W_1(t)=\frac{W(t)+S(t)}{[\ln|P^{-1}(t)|]^{\beta}}\to W_1<\infty,\ \text{a.s.}$$

This means

$$\begin{aligned}&W(t)=O([\ln|P^{-1}(t)|]^{\beta}),\ \text{a.s.},\\&S(t)=O([\ln|P^{-1}(t)|]^{\beta}),\ \text{a.s.}\end{aligned}\tag{34}$$

Due to the assumption that $H(z)$ is strictly positive real, using (33) gives

$$\sum_{i=1}^{t}\|\tilde{u}(i)\|^2=O([\ln|P^{-1}(t)|]^{\beta}),\ \text{a.s.}$$

From (27) and (34), we have

$$\begin{aligned}\|\tilde{\theta}(t)-\theta\|^2&=O\left(\frac{[\ln|P^{-1}(t)|]^{\beta}}{\lambda_{\min}[P^{-1}(t)]}\right)\\&=O\left(\frac{[\ln r(t)]^{\beta}}{\lambda_{\min}[P^{-1}(t)]}\right),\ \text{a.s.},\ \beta>1.\end{aligned}\tag{35}$$

Since $D(z)$ is stable, applying Lemma B.3.3 in (Goodwin & Sin, 1984) to (32) gets that there exist positive constants k_1 and k_2 such that

$$\begin{aligned}\sum_{i=1}^{t}\|\hat{v}(i)-v(i)\|^2&\leqslant k_1\sum_{i=1}^{t}\|\tilde{u}(i)\|^2+k_2\\&=O([\ln|P^{-1}(t)|]^{\beta})\\&=O([\ln r(t)]^{\beta}),\ \text{a.s.}\end{aligned}$$

The following is to show that $r(t)=O(r_0(t))$, $\lambda_{\min}[P^{-1}(t)]=O(\lambda_{\min}[P_0^{-1}(t)])$. Define $\tilde{\varphi}(t)$ (see the definitions of $\varphi_0(t)$ in (6) and $\varphi(t)$ in (20)):

$$\begin{aligned}\tilde{\varphi}(t)&:=\varphi_0(t)-\varphi(t)\\&=[0\cdots0\ v(t-1)-\hat{v}(t-1)\cdots v(t-n_d)\\&\quad-\hat{v}(t-n_d)]^{\mathrm{T}}.\end{aligned}$$

Hence

$$\begin{aligned}\sum_{i=1}^{t}\|\tilde{\varphi}(i)\|^2&=\sum_{i=1}^{t}\sum_{j=1}^{n_d}[v(i-j)-\hat{v}(i-j)]^2\\&=O\left(\sum_{i=1}^{t}[v(i)-\hat{v}(i)]^2\right)\\&=O([\ln r(t)]^{\beta}),\ \text{a.s.},\\\sum_{i=1}^{t}\|\varphi(i)\|^2&=\sum_{i=1}^{t}\|\varphi_0(i)-\tilde{\varphi}(i)\|^2\\&\leqslant2\sum_{i=1}^{t}\|\varphi_0(i)\|^2+2\sum_{i=1}^{t}\|\tilde{\varphi}(i)\|^2\\&=2\sum_{i=1}^{t}\|\varphi_0(i)\|^2+O([\ln r(t)]^{\beta}),\ \text{a.s.}\end{aligned}$$

Thus, according to the definitions of $r_0(t)$ and $r(t)$, we have

$$r(t)=2r_0(t)+O([\ln r(t)]^{\beta})=O(r_0(t)),\ \text{a.s.}\tag{36}$$

For any vector $\omega\in\mathbb{R}^{n_0}$ with $\|\omega\|=1$, we have

$$\begin{aligned}\sum_{i=1}^{t}[\omega^{\mathrm{T}}\varphi(i)]^2&=\sum_{i=1}^{t}[\omega^{\mathrm{T}}\varphi_0(i)-\omega^{\mathrm{T}}\tilde{\varphi}(i)]^2\\&\leqslant2\sum_{i=1}^{t}[\omega^{\mathrm{T}}\varphi_0(i)]^2+2\sum_{i=1}^{t}\|\tilde{\varphi}(i)\|^2\\&=2\sum_{i=1}^{t}[\omega^{\mathrm{T}}\varphi_0(i)]^2+O([\ln r(t)]^{\beta}),\ \text{a.s.}\end{aligned}$$

It follows that

$$\begin{aligned}&\lambda_{\min}[P^{-1}(t)]\leqslant2\lambda_{\min}[P_0^{-1}(t)]+O(\lambda_{\min}[P^{-1}(t)]),\ \text{a.s.},\\&\lambda_{\min}[P^{-1}(t)]=O(\lambda_{\min}[P_0^{-1}(t)]),\ \text{a.s.}\end{aligned}\tag{37}$$

Combining (35) with (36) and (37) gives

$$\begin{aligned}\|\hat{\theta}(t)-\theta\|^2&=O\left(\frac{[\ln r_0(t)]^{\beta}}{\lambda_{\min}[P_0^{-1}(t)]}\right)\\&=O\left(\frac{\{\ln\lambda_{\max}[P_0^{-1}(t)]\}^{\beta}}{\lambda_{\min}[P_0^{-1}(t)]}\right),\ \text{a.s.},\ \beta>1.\end{aligned}$$

This completes the proof of Theorem 1. □

The assumptions in (A1) and (A2) imply that the noise $v(t)$ is a non-stationary white noise sequence with zero mean, time-varying but bounded variance. Theorem 1 shows that for the noise sequence $\{v(t)\}$ with a bounded variance, the rate of convergence of the parameter estimation to their true values is the ratio of the logarithm of the maximum eigenvalue to the minimum eigenvalue of the covariance matrix, $P_0^{-1}(t)$. Moreover, we easily get the following corollary from Theorem 1.

Table 1
The parameter estimates (θ_i) ($\sigma_v^2 = 0.50^2$, $\delta_{ns} = 11.653\%$)

t	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	δ (%)
100	−1.59597	0.80123	0.75070	0.69354	0.45732	0.31299	0.26216	0.14621	−0.36960	13.18011
200	−1.59923	0.79958	0.78164	0.77377	0.40964	0.33924	0.26842	0.09929	−0.58852	7.68636
300	−1.59684	0.79672	0.79524	0.76342	0.41883	0.32851	0.25575	0.11465	−0.60019	6.51204
500	−1.59244	0.79295	0.77672	0.82254	0.42924	0.31283	0.25961	0.06693	−0.57727	9.95851
1000	−1.59887	0.79782	0.80607	0.72766	0.43724	0.30104	0.23026	0.12530	−0.56187	5.67999
1500	−1.60207	0.80036	0.82259	0.68292	0.43186	0.30921	0.22889	0.13800	−0.58364	3.47695
2000	−1.60256	0.80138	0.82528	0.67519	0.42312	0.32007	0.22526	0.14281	−0.60951	2.32431
2500	−1.60282	0.80114	0.82879	0.67596	0.42478	0.31635	0.22336	0.14394	−0.62808	1.88293
3000	−1.60291	0.80156	0.84478	0.67441	0.43202	0.31340	0.21612	0.14622	−0.64690	1.49570
True values	−1.60000	0.80000	0.85000	0.65000	0.42500	0.32500	0.21250	0.16250	−0.64000	

Table 2
The parameter estimates (θ_i) ($\sigma_v^2 = 2.00^2$, $\delta_{ns} = 46.611\%$)

t	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	δ (%)
100	−1.54508	0.76421	0.56011	0.86926	0.53976	0.29065	0.37292	0.10681	−0.32423	23.37386
200	−1.57785	0.77943	0.60637	1.12032	0.34540	0.40403	0.42654	-0.07202	−0.57475	27.94078
300	−1.58668	0.78434	0.64597	1.08046	0.38563	0.34884	0.38080	-0.02243	−0.59590	23.97798
500	−1.57976	0.78454	0.56133	1.31829	0.44127	0.28573	0.39912	-0.21849	−0.56429	37.47846
1000	−1.59187	0.79061	0.67639	0.95419	0.47500	0.23726	0.28259	0.01695	−0.55421	18.04558
1500	−1.60397	0.80027	0.74428	0.78297	0.45507	0.26942	0.27671	0.06543	−0.58491	9.83487
2000	−1.60696	0.80480	0.75587	0.75304	0.42046	0.31142	0.26159	0.08366	−0.61362	7.54129
2500	−1.60791	0.80325	0.76915	0.75519	0.42702	0.29587	0.25454	0.08912	−0.63243	7.09530
3000	−1.60866	0.80516	0.83140	0.74788	0.45593	0.28331	0.22632	0.09851	−0.65213	5.79282
True values	−1.60000	0.80000	0.85000	0.65000	0.42500	0.32500	0.21250	0.16250	−0.64000	

Table 3
The parameter estimates (a_i, b_i, c_i, d_i) ($\sigma_v^2 = 0.50^2$, $\delta_{ns} = 11.653\%$)

t	a_1	a_2	b_1	b_2	c_2	c_3	d_1	δ_s (%)
100	−1.59597	0.80123	0.75070	0.69354	0.53024	0.28002	−0.36960	13.08107
200	−1.59923	0.79958	0.78164	0.77377	0.48125	0.23586	−0.58852	6.76563
300	−1.59684	0.79672	0.79524	0.76342	0.47849	0.23589	−0.60019	5.98176
500	−1.59244	0.79295	0.77672	0.82254	0.46648	0.20781	−0.57727	9.11354
1000	−1.59887	0.79782	0.80607	0.72766	0.47807	0.22893	−0.56187	5.44013
1500	−1.60207	0.80036	0.82259	0.68292	0.48889	0.24017	−0.58364	3.21442
2000	−1.60256	0.80138	0.82528	0.67519	0.49337	0.24223	−0.60951	2.12493
2500	−1.60282	0.80114	0.82879	0.67596	0.49026	0.24122	−0.62808	1.68998
3000	−1.60291	0.80156	0.84478	0.67441	0.48805	0.23632	−0.64690	1.41302
True values	−1.60000	0.80000	0.85000	0.65000	0.50000	0.25000	−0.64000	

Corollary 1. *Assume that there exist positive constants β_0, β_1, β_2 and t_0 such that, for $t \geqslant t_0$, the following generalized persistent excitation condition (unbounded condition number) holds*:

$$\text{(A4)} \quad \beta_1 I \leqslant \frac{1}{t}\sum_{i=1}^{t} \varphi_0(i)\varphi_0^{\mathrm{T}}(i) \leqslant \beta_2 t^{\beta_0} I, \text{ a.s.}$$

Then

$$\|\hat{\theta}(t) - \theta\|^2 = O\left(\frac{[\ln t]^{\beta}}{t}\right) \to 0, \text{ a.s.}, \ \beta > 1.$$

For an arbitrary small positive real ε, we have $[\ln t]^{\beta} = o(t^{\varepsilon})$. Hence

$$\|\hat{\theta}(t) - \theta\|^2 = O\left(\frac{1}{t^{1-\varepsilon}}\right) \to 0, \text{ a.s.}$$

Condition (A4) is also termed as the generalized persistent excitation condition, because setting $\beta_0 = 0$ in Condition (A4), we get the weak persistent excitation condition (bounded condition number) (Ljung, 1999).

For linear ARMAX systems, i.e., $\bar{u}(t) = f(u(t)) = u(t)$, the convergence results of the parameter estimation in

Table 4
The parameter estimates (a_i, b_i, c_i, d_i) ($\sigma_v^2 = 2.00^2$, $\delta_{\rm ns} = 46.611\%$)

t	a_1	a_2	b_1	b_2	c_2	c_3	d_1	δ_s (%)
100	−1.54508	0.76421	0.56011	0.86926	0.64902	0.39434	−0.32423	23.47209
200	−1.57785	0.77943	0.60637	1.12032	0.46513	0.31957	−0.57475	23.99653
300	−1.58668	0.78434	0.64597	1.08046	0.45992	0.28437	−0.59590	21.40112
500	−1.57976	0.78454	0.56133	1.31829	0.50143	0.27265	−0.56429	32.54924
1000	−1.59187	0.79061	0.67639	0.95419	0.47545	0.21778	−0.55421	16.12954
1500	−1.60397	0.80027	0.74428	0.78297	0.47776	0.22768	−0.58491	8.05827
2000	−1.60696	0.80480	0.75587	0.75304	0.48490	0.22859	−0.61362	6.42708
2500	−1.60791	0.80325	0.76915	0.75519	0.47348	0.22448	−0.63243	6.13756
3000	−1.60866	0.80516	0.83140	0.74788	0.46360	0.20197	−0.65213	5.22011
True values	−1.60000	0.80000	0.85000	0.65000	0.50000	0.25000	−0.64000	

Theorem 1 and Corollary 1 still hold. Even in this special case, we believe the proof and result have some degree of novelty. Here, unlike in Guo and Chen (1991), Lai and Wei (1986), Ren and Kumar (1994), there is no assumption that the high-order moments of the noise $\{v(t)\}$ exist, i.e., we do not assume that $\mathrm{E}[|v(t)|^{\beta_3}|\mathscr{F}_{t-1}] < \infty$, a.s. for some $\beta_3 > 2$.

If we take $N(z) = 1/A(z)$, we obtain the nonlinear ARX model, whose identification is very simple because it is a special case of the HARMAX model just studied.

The strictly positive condition (A3) and the persistent excitation condition (A4) are standard assumptions, and are not due to the nonlinear block. They are inherently related to the convergence of some identification algorithms for linear systems, see the work in, e.g., Guo and Chen (1991), Lai and Wei (1986), Ren and Kumar (1994) and Solo (1979).

If the input is taken as a pseudo-random binary sequence or uncorrelated white noise sequence, then (A4) is automatically satisfied because $u(t)$ is a persistent excitation signal, and so is $\bar{u}(t)$ (Ljung, 1999; Pearson & Pottmann, 2000). Since the white noise is a best persistent excitation signal (Ljung, 1999), the generalized persistent excitation condition naturally holds as long as the persistent excitation input signal is uncorrelated with the noise.

5. Example

An example is given to demonstrate the effectiveness of the proposed algorithms. Consider the following system:

$$
\begin{aligned}
&A(z)y(t) = B(z)\bar{u}(t) + D(z)v(t),\\
&A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} = 1 - 1.60z^{-1} + 0.80z^{-2},\\
&B(z) = b_1 z^{-1} + b_2 z^{-2} = 0.85z^{-1} + 0.65z^{-2},\\
&D(z) = 1 + d_1 z^{-1} = 1 - 0.64z^{-1},\\
&\bar{u}(t) = f(u(t)) = c_1 u(t) + c_2 u^2(t) + c_3 u^3(t)\\
&\qquad = u(t) + 0.5u^2(t) + 0.25u^3(t),\\
&\theta_s = [a_1\ \ a_2\ \ b_1\ \ b_2\ \ c_2\ \ c_3\ \ d_1]^{\mathrm{T}}.
\end{aligned}
$$

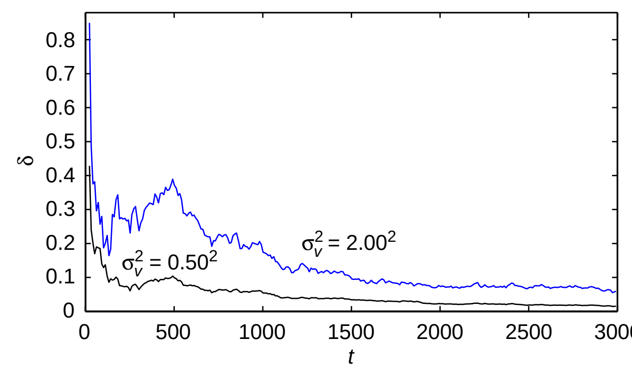

Fig. 2. The parameter estimation errors δ vs. t.

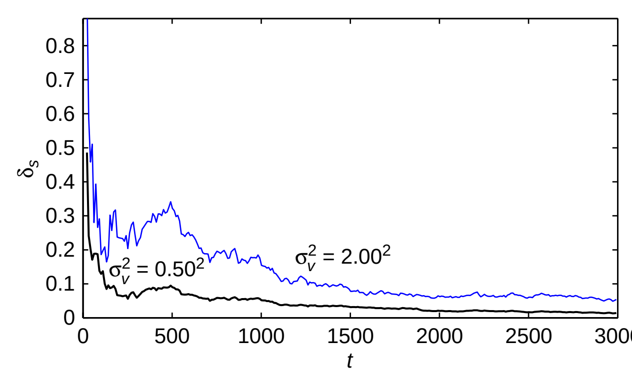

Fig. 3. The parameter estimation errors δ_s vs. t.

$\{u(t)\}$ is taken as a persistent excitation signal sequence with zero mean and unit variance $\sigma_u^2 = 1.00^2$, and $\{v(t)\}$ as a white noise sequence with zero mean and constant variance σ_v^2. Apply the proposed algorithm in (18)–(23) to estimate the parameters of this system, the parameter estimates θ_i and θ_s and their errors with different noise variances are shown in Tables 1–4, and the parameter estimation errors δ and δ_s versus t are shown in Figs. 2 and 3, where $\delta_{\rm ns}$ is the noise-to-signal ratio defined by the square root of the ratio of the

Table 5
The parameter estimates (θ_i) ($\sigma_v^2 = 0.50^2$, $\delta_{ns} = 11.653\%$)

k	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	δ (%)
1	−1.58627	0.78725	0.85442	0.62485	0.41056	0.32202	0.21080	0.19035	0.18700	36.62919
2	−1.59130	0.79259	0.85531	0.63659	0.41814	0.32905	0.20997	0.18561	−0.44101	8.90498
3	−1.59259	0.79416	0.85582	0.63906	0.41639	0.33369	0.20964	0.18467	−0.56286	3.65865
4	−1.59299	0.79465	0.85624	0.64411	0.41588	0.33530	0.20933	0.18225	−0.60813	1.85369
5	−1.59347	0.79516	0.85615	0.64117	0.41591	0.33539	0.20935	0.18315	−0.62812	1.36329
6	−1.59354	0.79524	0.85620	0.64168	0.41571	0.33586	0.20933	0.18288	−0.63654	1.26149
True values	−1.60000	0.80000	0.85000	0.65000	0.42500	0.32500	0.21250	0.16250	−0.64000	

Table 6
The parameter estimates (a_i, b_i, c_i, d_i) ($\sigma_v^2 = 0.50^2$, $\delta_{ns} = 11.653\%$)

k	a_1	a_2	b_1	b_2	c_2	c_3	d_1	δ_s (%)
1	−1.58627	0.78725	0.85442	0.62485	0.49793	0.27568	0.18700	36.78368
2	−1.59130	0.79259	0.85531	0.63659	0.50289	0.26853	−0.44101	8.91685
3	−1.59259	0.79416	0.85582	0.63906	0.50435	0.26696	−0.56286	3.58177
4	−1.59299	0.79465	0.85624	0.64411	0.50313	0.26371	−0.60813	1.64142
5	−1.59347	0.79516	0.85615	0.64117	0.50444	0.26509	−0.62812	1.06110
6	−1.59354	0.79524	0.85620	0.64168	0.50446	0.26474	−0.63654	0.91171
True values	−1.60000	0.80000	0.85000	0.65000	0.50000	0.25000	−0.64000	

variances of $w(t)$ and $x(t)$ in Fig. 1, i.e.,

$$\delta_{ns} = \sqrt{\frac{\mathrm{var}[w(t)]}{\mathrm{var}[x(t)]}} \times 100\% = \frac{\sigma_w}{\sigma_u} \times 100\%,$$

$\delta = \|\hat{\theta}(t) - \theta\|/\|\theta\|$ and $\boxed{\delta_s = \|\hat{\theta}_s(t) - \theta_s\|/\|\theta\|}$ are the relative parameter estimation errors, $\hat{\theta}_s(t)$ being the estimate of θ_s. When $\sigma_v^2 = 0.50^2$ and $\sigma_v^2 = 2.00^2$, the corresponding noise-to-signal ratios are $\delta_{ns} = 11.653\%$ and $\delta_{ns} = 46.611\%$, respectively.

We use the HARMAX-LSI algorithm in (13)–(16) to iteratively compute the parameter estimates of this example, as shown in Tables 5 and 6, where $\sigma_v^2 = 0.50^2$ and the data length is 3000.

From Tables 1–6 and Figs. 2–3, we can draw the following conclusions:

- Increasing data length generally leads to smaller parameter estimation errors.
- A high noise level results in a slow rate of convergence of the parameter estimates to the true parameters.
- It is clear that the errors δ and δ_s are becoming smaller (in general) as t increases. This confirms the proposed theorem.
- For the same data length, the iterative algorithm gives better parameter estimates than the recursive algorithm because the former repeatedly uses the available data (compare Table 1 with Table 5, and Table 2 with Table 6).

6. Conclusions

An iterative and a recursive algorithms based on replacing unmeasurable noise variables by their estimates are derived for Hammerstein nonlinear models. The analysis using the martingale convergence theorem indicates that the proposed recursive least-squares algorithm of Hammerstein nonlinear ARMAX models can give consistent parameter estimation. Although the algorithms are developed for the Hammerstein models, the basic idea can be extended to identify Hammerstein–Wiener models in (Bai, 1998). The least-squares iterative algorithm presented is quite interesting, but its convergence analysis is more difficult and is worth further research.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada.

References

Bai, E. W. (1998). An optimal two-stage identification algorithm for Hammerstein–Wiener nonlinear systems. *Automatica*, *34*(3), 333–338.
Bai, E. W. (2002a). Identification of linear systems with hard input nonlinearities of known structure. *Automatica*, *38*(5), 853–860.
Bai, E. W. (2002b). A blind approach to the Hammerstein–Wiener model identification. *Automatica*, *38*(6), 967–979.
Bai, E. W. (2003). Frequency domain identification of Wiener models. *Automatica*, *39*(9), 1521–1530.

Bai, E. W. (2004). Decoupling the linear and nonlinear parts in Hammerstein model identification. *Automatica*, *40*(4), 671–676.

Cerone, V., & Regruto, D. (2003). Parameter bounds for discrete-time Hammerstein models with bounded output errors. *IEEE Transactions on Automatic Control*, *48*(10), 1855–1860.

Chang, F., & Luus, R. (1971). A noniterative method for identification using Hammerstein model. *IEEE Transactions on Automatic Control*, *16*(5), 464–468.

Gallman, P. G. (1976). A comparison of two Hammerstein model identification algorithms. *IEEE Transactions on Automatic Control*, *21*(1), 124–126.

Goodwin, G. C., & Sin, K. S. (1984). *Adaptive filtering, prediction and control*. Englewood Cliffs, New Jersey: Prentice-Hall.

Greblicki, W. (1997). Non parametric approach to Wiener system identification. *IEEE Transactions on Circuits System I*, *44*(6), 538–545.

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

Guo, L., & Chen, H. F. (1991). The Åström–Wittenmark self-tuning regulator revisited and ELS-based adaptive trackers. *IEEE Transactions on Automatic Control*, *36*(7), 802–812.

Haber, R., & Unbehauen, H. (1990). Structure identification of nonlinear dynamic systems—A survey of input-output approaches. *Automatica*, *26*(4), 651–677.

Haist, N. D., Chang, F., & Luus, R. (1973). Nonlinear identification in the presence of correlated noise using a Hammerstein model. *IEEE Transactions on Automatic Control*, *18*(5), 553–555.

Lai, T. L., & Wei, C. Z. (1986). Extended least squares and their applications to adaptive control and prediction in linear systems. *IEEE Transactions on Automatic Control*, *31*(10), 898–906.

Lang, Z. (1994). On identification of the controlled plants described by the Hammerstein system. *IEEE Transactions on Automatic Control*, *39*(3), 569–573.

Lang, Z. (1997). A nonparametric polynomial identification algorithm for the Hammerstein system. *IEEE Transactions on Automatic Control*, *42*(10), 1435–1441.

Ljung, L. (1999). *System identification*: *theory for the user*. (2nd ed). Englewood Cliffs, New Jersey: Prentice-Hall.

Narendra, K. S., & Gallman, P. G. (1966). An iterative method for the identification of nonlinear systems using a Hammerstein model. *IEEE Transactions on Automatic Control*, *11*(3), 546–550.

Nesic, D., & Mareels, I. M. Y. (1998). Dead-beat control of simple Hammerstein models. *IEEE Transactions on Automatic Control*, *43*(8), 1184–1188.

Ninness, B., & Gibson, S. (2002). Quantifying the accuracy of Hammerstein model estimation. *Automatica*, *38*(12), 2037–2051.

Pawlak, M. (1991). On the series expansion approach to the identification of Hammerstein system. *IEEE Transactions on Automatic Control*, *36*(6), 763–767.

Pearson, R. K., & Pottmann, M. (2000). Gray-box identification of block-oriented nonlinear models. *Journal of Process Control*, *10*(4), 301–315.

Rangan, S., Wolodkin, G., & Poolla, K. (1995). Identification methods for Hammerstein systems. *Proceedings of Control and Decision Conference*, New Orleans, 697–702.

Ren, W., & Kumar, P. K. (1994). Stochastic adaptive prediction and model reference control. *IEEE Transactions on Automatic Control*, *39*(10), 2047–2060.

Söderström, T., & Stoica, P. (1989). *System identification*. New York: Prentice-Hall.

Solo, V. (1979). The Convergence of AML. *IEEE Transactions on Automatic Control*, *24*(6), 958–962.

Stoica, P. (1981). On the convergence of an iterative algorithm used for Hammerstein system identification. *IEEE Transactions on Automatic Control*, *26*(4), 967–969.

Vörös, J. (1995). Identification of nonlinear dynamic systems using extended Hammerstein and Wiener models. *Control-Theory and Advanced Technology*, *10*(4), Part 2, 1203–1212.

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

Wigren, T., & Nordsjö, A. E. (1999). Compensation of the RLS algorithm for output nonlinearities. *IEEE Transactions on Automatic Control*, *44*(10), 1913–1918.

Feng Ding was born in Guangshui, Hubei Province. He received the B.Sc. degree in Electrical Engineering from Hubei University of Tchnology (Wuhan, P.R. China) in 1984, and the M.Sc. and Ph.D. degrees in automatic control both from the Department of Automation, Tsinghua University in 1991 and 1994, respectively. From 1984 to 1988, he was an Electrical Engineer at the Hubei Pharmaceutical Factory. Since 1994 he was with Department of Automation, Tsinghua University. He is now a Professor in the Control Science and Engineering Research Center at the Southern Yangtze University, Wuxi 214122 China, and has been a Research Associate at the University of Alberta, Edmonton, Canada since 2002. His current research interests include model identification, adaptive control, process control, stochastic systems and multirate systems. He co-authored the book *Adaptive Control Systems* (Tsinghua University Press, Beijing, 2002), and published over eighty papers on modelling and identification as the first author.

Tongwen Chen received the B.Sc. degree from Tsinghua University (Beijing) in 1984, and the M.A.Sc. and Ph.D. degrees from the University of Toronto in 1988 and 1991, respectively, all in Electrical Engineering. From 1991 to 1997, he was an Assistant/Associate Professor in the Department of Electrical and Computer Engineering at the University of Calgary, Canada. Since 1997, he has been with the Department of Electrical and Computer Engineering at the University of Alberta, Edmonton, Canada, and is presently a Professor. He held visiting positions at the Hong Kong University of Science and Technology, Tsinghua University, and Kumamoto University.

His current research interests include process control, multirate systems, robust control, network based control, digital signal processing, and their applications to industrial problems. He co-authored with B.A. Francis the book *Optimal Sampled-Data Control Systems* (Springer, 1995). Dr. Chen received a University of Alberta McCalla Professorship for 2000/2001, and a Fellowship from the Japan Society for the Promotion of Science for 2004. He was an Associate Editor for IEEE Transactions on Automatic Control during 1998-2000. Currently he is an Associate Editor for Automatica, Systems and Control Letters, and DCDIS Series B. He is a registered Professional Engineer in Alberta, Canada.