Available online at www.sciencedirect.com
SCIENCE DIRECT®
automatica
Automatica 41 (2005) 315–325
www.elsevier.com/locate/automatica

Brief paper

Hierarchical gradient-based identification of multivariable discrete-time 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, Canada T6G 2V4

Received 21 August 2003; received in revised form 7 June 2004; accepted 22 October 2004

Abstract

In this paper, we use a hierarchical identification principle to study identification problems for multivariable discrete-time systems. We propose a hierarchical gradient iterative algorithm and a hierarchical stochastic gradient algorithm and prove that the parameter estimation errors given by the algorithms converge to zero for any initial values under persistent excitation. The proposed algorithms can be applied to identification of systems involving non-stationary signals and have significant computational advantage over existing identification algorithms. Finally, we test the proposed algorithms by simulation and show their effectiveness.

Keywords: Recursive identification; Estimation; Least squares; Hierarchical identification principle; Multivariable systems; Convergence properties

1. Introduction

For decades, a great deal of work has been published on the identification of multivariable, or multi-input multi-output (MIMO), systems (see, e.g., Gauthier & Landau, 1978; El-Sherief & Sinha, 1979; Sinha & Kwong, 1979; El-Sherief, 1981; Verhaegen & Dewilde, 1992a,b; Verhaegen, 1993, 1994; Overshee & De Moor, 1994, 1996; Chou & Verhaegen, 1997; McKelvey, Akcay, & Ljung, 1996; Pintelon, 2002); however, further research in this area is still required for the following reasons:

- In the area of MIMO system identification based on difference equations, most existing identification algorithms using transfer matrices employ the idea of decomposing a MIMO system into several subsystems, depending on the number of outputs, and then of estimating parameters of the subsystems one by one (Gauthier & Landau, 1978; El-Sherief & Sinha, 1979; Sinha & Kwong, 1979; El-Sherief, 1981). Since such identification algorithms require computing many covariance matrices (one for each subsystem), they have the drawback of having heavy computational load. The simultaneous identification of all parameters of a system can reduce the computational burden, e.g., Sen and Sinha (1976) suggested to use a matrix pseudo-inverse approach; however, the computational load is large due to large number of zero entries in the information matrix in the estimation algorithm. Moreover, the algorithm in Sen and Sinha (1976) handles noise-free data only. Recently, Pintelon (2002) studied the stochastic properties (strong convergence, asymptotic normality, strong consistency) of the frequency-domain subspace algorithms described in McKelvey et al. (1996) and Van Overschee and De Moor (1996), where the true noise covariance matrix was replaced by the sample noise covariance matrix obtained from a small number of independent repeated experiments.

☆ This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor B. Ninness under the direction of Editor T. Soederstrorm. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

* Corresponding author. Tel.: +780 492 3940; fax: +780 492 1811.
E-mail addresses: fding@sytu.edu.cn (F. Ding), tchen@ece.ualberta.ca (T. Chen).
[1] Feng Ding is currently a Research Associate at the University of Alberta, Edmonton, Canada.

0005-1098/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.automatica.2004.10.010

- In the off-line state-space model identification literature, subspace state-space identification (4SID for short) methods based on the RQ factorization and singular value decomposition (SVD) have been developed for MIMO systems (Verhaegen & Dewilde, 1992a,b; Verhaegen, 1993, 1994; Overshee & De Moor, 1994, 1996; Chou & Verhaegen, 1997). The basic idea is to determine the extended observability matrix from the SVD or RQ factorization of an information matrix consisting of given input/output (I/O) data, and then to compute the system parameter matrices. But, as the size of the information matrix grows, the difficulty and complexity in computation increase.
- In the recursive 4SID area, some methods (e.g., Gustafsson, 1998; Oku & Kimura, 2002) are based on the idea of directly updating an estimate of the extended observability matrix by using subspace tracking; other methods (e.g., Verhaegen & Deprettere, 1991; Lovera, Gustafsson, & Verhaegen, 2000) are based on subspace tracking ideas for the recursive update of the RQ factorizations or the SVD by using array signal processing algorithms; see Comon and Golub (1990) for a review of subspace tracking algorithms and Yang (1995, 1996) for projection approximation subspace tracking. Finally, Cho, Xu, and Kailath (1994) presented a recursive identification method of state-space models using the generalized Schur algorithm for updating the noise covariance matrix. However, in order to obtain system parameter estimates, these 4SID methods also require some additional computation, e.g., computing the A-matrix using the shift invariant structure of the extended observability matrix (obtained by use of SVD) and computing B- and D-matrices using least squares methods (Lovera et al., 2000). Our approach in this work is to update *directly* parameter estimates as in the prediction error method based on difference equation descriptions (Ljung, 1999); moreover, we do not assume that the problems are stationary and/or ergodic, which is different from those mentioned above. The algorithms proposed are simple and easy to implement, and have less computational effort than existing algorithms.

Therefore, developing computationally efficient and recursive system identification algorithms is the goal in this paper. We will frame our study in the identification of transfer matrix models. The key idea is the so-called *hierarchical identification*, and is inspired by the hierarchical control based on the decomposition-coordination principle for large-scale systems (Singh, 1980; Tamura & Yoshikawa, 1990; Drouin, Abou-Kandil, & Mariton, 1991). Hierarchical identification uses subsystem decomposition in identification, and is also called *bootstrap identification*.

The principle of hierarchical identification is as follows. A system is decomposed into several subsystems with smaller dimension and fewer variables, and then the parameter vector and/or parameter matrix of each subsystem is identified, respectively. Because of such decomposition, difficulties arise in that there exist common unknown parameters/quantities among subsystems, which normally requires difficult iterative calculation. In order to overcome such difficulties, when recursively computing the parameter estimate of the ith subsystem, the hierarchical identification principle implies that the unknown parameters of other subsystems which appeared in the ith subsystem are replaced with their estimates. Using this idea, we present the hierarchical gradient iterative algorithm and hierarchical stochastic gradient algorithm for MIMO systems. The main advantage of such algorithms is that they require less computational effort than existing identification algorithms, e.g., the Sen and Sinha's algorithm and the 4SID methods mentioned above.

The hierarchical identification methods have important applications in parameter identification of multirate systems (Chen & Qiu, 1994; Qiu & Chen, 1994, 1999; Li, Shah, & Chen, 2001, 2002; Li, Shah, Chen, & Qi, 2003; Tangirala, Li, Patwardhan, Shah, & Chen, 2001; Sheng, Chen, & Shah, 2002), because lifting converts a multirate time-varying system into a time-invariant MIMO system.

The paper is organized as follows. In Section 2, we discuss modeling issues related to MIMO systems. In Sections 3 and 4, we develop the hierarchical gradient iterative algorithm and hierarchical stochastic gradient identification algorithm, and analyze the performance of the proposed algorithms. In Section 5, we compare computational efficiency of our algorithm with several existing ones, establishing a clear advantage. Section 6 presents an illustrative example for the results in this paper. Finally, concluding remarks are given in Section 7.

2. The problem formulation

Consider a linear discrete-time, multivariable system described by the following state-space model:

$$x(t+1) = Ax(t) + Bu(t),$$
$$y(t) = Cx(t) + Du(t),$$

whose input/output relationship can be represented as

$$y(t) = G(z)u(t). \tag{1}$$

Here, $x(t) \in \mathbb{R}^n$ is the state vector, $u(t) \in \mathbb{R}^r$ the system input vector, $y(t) \in \mathbb{R}^m$ the system output vector, (A, B, C, D) the system matrices of appropriate sizes, and $G(z) \in \mathbb{R}^{m\times r}$ the transfer matrix (TM) which relates to the state-space data as follows:

$$G(z) = C(zI - A)^{-1}B + D = \frac{C\,\mathrm{adj}[zI - A]B}{\det[zI - A]} + D$$
$$= \frac{C\,\mathrm{adj}[I - Az^{-1}]B}{z^{-n}\det[zI - A]} + D =: \frac{Q(z)}{\alpha(z)}$$

with $\alpha(z)$ the characteristic polynomial in the unit delay operator z^{-1} [$z^{-1}y(t) = y(t-1)$] of degree n, defined as the least common denominator of $G(z)$, $Q(z)$ a polynomial

matrix in z^{-1}, and both represented as

$$\alpha(z) = 1 + \alpha_1 z^{-1} + \alpha_2 z^{-2} + \cdots + \alpha_n z^{-n}, \quad \alpha_i \in \mathbb{R}^1,$$

$$Q(z) = Q_0 + Q_1 z^{-1} + Q_2 z^{-2} + \cdots + Q_n z^{-n}, \quad Q_i \in \mathbb{R}^{m\times r}.$$

The identification of the characteristic polynomial $\alpha(z)$ is very important for, e.g., pole placement in control design. Also, for systems with unknown parameters, it is unreasonable to assume that some entries of the TM have some common known divisor.

Eq. (1) can also be expressed as

$$\alpha(z)y(t) = Q(z)u(t) \quad \text{or}$$
$$y(t) + \sum_{i=1}^{n} \alpha_i y(t-i) = \sum_{i=0}^{n} Q_i u(t-i). \tag{2}$$

Define the parameter matrix θ, parameter vector α, information vector $\varphi(t)$ and information matrix $\psi(t)$ as

$$\theta^{\mathrm{T}} = [Q_0 \;\; Q_1 \;\; \cdots \;\; Q_n] \in \mathbb{R}^{m\times n_0}, \quad n_0 := (n+1)r,$$

$$\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix} \in \mathbb{R}^n, \quad \varphi(t) = \begin{bmatrix} u(t) \\ u(t-1) \\ \vdots \\ u(t-n) \end{bmatrix} \in \mathbb{R}^{n_0},$$

$$\psi(t) = [y(t-1) \;\; y(t-2) \;\; \cdots \;\; y(t-n)] \in \mathbb{R}^{m\times n}.$$

Hence, from (2), we obtain the following identification model:

$$y(t) + \psi(t)\alpha = \theta^{\mathrm{T}}\varphi(t). \tag{3}$$

The system parameters to be identified in (3) include two parts: one parameter vector α consisting of the coefficients of the characteristic polynomial of the system, and one parameter matrix θ consisting of the coefficients of the numerator polynomial matrix of the TM. Due to the presence of an unknown parameter vector $\alpha \in \mathbb{R}^n$ and an unknown parameter matrix $\theta^{\mathrm{T}} \in \mathbb{R}^{m\times n_0}$ in (3), we can use the Kronecker product to transform the parameter matrix θ^{T} into a stacked vector $\mathrm{vec}(\theta^{\mathrm{T}})$, then the model in (3) may be rewritten as

$$y(t) = [-\psi(t), (\varphi^{\mathrm{T}}(t) \otimes I_m)] \begin{bmatrix} \alpha \\ \mathrm{vec}(\theta^{\mathrm{T}}) \end{bmatrix} =: H(t)\theta_s, \quad H(t) \in \mathbb{R}^{m\times(mn_0+n)}, \;\; \theta_s \in \mathbb{R}^{mn_0+n}, \tag{4}$$

where I_m is an $m \times m$ identity matrix. Although this model may be identified by the recursive least squares algorithm (in fact, this is the Sen and Sinha's matrix pseudo-inverse approach), as we have pointed out in the introduction that it requires computing large covariance matrices of dimensions $(mn_0 + n) \times (mn_0 + n)$ and an $m \times m$ matrix inversion at each step. A comparison of computational efficiency of several algorithms is given in Section 5.

Therefore, the objectives of this paper are two-fold: first, by means of the hierarchical identification principle, present new algorithms to estimate the unknown parameters $(\alpha_i, \; Q_i)$ in (2) from the given input–output measurement data $\{u(t), y(t) : \; t = 1, 2, \ldots\}$, or equivalently, to estimate (α, θ) in (3) from $\{y(t), \psi(t), \varphi(t) : \; t = 1, 2, \ldots\}$; and second, study convergence performance issues of the new algorithms presented.

3. The hierarchical gradient iterative algorithm

In this section, according to the hierarchical identification principle, we decompose the MIMO system in (3) into two subsystems: one containing the parameter vector α, and the other containing the parameter matrix θ; and then the iterative solutions of the parameter vector and parameter matrix of the two subsystems are established by application of the steepest descent principle. The details are as follows.

Define two vectors

$$b_1(t) := -y(t) + \theta^{\mathrm{T}}\varphi(t),$$
$$b_2(t) := y(t) + \psi(t)\alpha.$$

Then, we can decompose the system in (3) into the following two fictitious subsystems:

$$S_1: \quad \psi(t)\alpha = b_1(t),$$
$$S_2: \quad \theta^{\mathrm{T}}\varphi(t) = b_2(t).$$

Suppose $t \gg mn_0 + n$. Define

$$\Psi(t) := \begin{bmatrix} \psi(1) \\ \psi(2) \\ \vdots \\ \psi(t) \end{bmatrix},$$

$$B_1(t) := \begin{bmatrix} b_1(1) \\ b_1(2) \\ \vdots \\ b_1(t) \end{bmatrix} = \begin{bmatrix} -y(1) + \theta^{\mathrm{T}}\varphi(1) \\ -y(2) + \theta^{\mathrm{T}}\varphi(2) \\ \vdots \\ -y(t) + \theta^{\mathrm{T}}\varphi(t) \end{bmatrix}, \tag{5}$$

$$\Phi(t) := [\varphi(1) \;\; \varphi(2) \;\; \cdots \;\; \varphi(t)],$$

$$B_2(t) := [b_2(1) \;\; b_2(2) \;\; \cdots \;\; b_2(t)] = [y(1) + \psi(1)\alpha \;\; y(2) + \psi(2)\alpha \;\; \cdots \;\; y(t) + \psi(t)\alpha]. \tag{6}$$

So, we have

$$S_1: \quad \Psi(t)\alpha = B_1(t), \tag{7}$$

$$S_2: \quad \theta^{\mathrm{T}}\Phi(t) = B_2(t) \quad \text{or } \Phi^{\mathrm{T}}(t)\theta = B_2^{\mathrm{T}}(t). \tag{8}$$

Using the negative gradient search, we may obtain the iterative solutions α_k and θ_k of α and θ, respectively, as follows:

$$\alpha_k = \alpha_{k-1} + \mu\Psi^{\mathrm{T}}(t)[B_1(t) - \Psi(t)\alpha_{k-1}], \tag{9}$$

$$\theta_k^{\mathrm{T}} = \theta_{k-1}^{\mathrm{T}} + \mu[B_2(t) - \theta_{k-1}^{\mathrm{T}}\Phi(t)]\Phi^{\mathrm{T}}(t), \quad k = 1, 2, \ldots, \tag{10}$$

where $\mu > 0$ is called the iterative step-size or convergence factor to be given later. Substituting $B_1(t)$ in (5) into (9), $B_2(t)$ in (6) into (10) gives

$$\alpha_k = \alpha_{k-1} + \mu\Psi^{\mathrm{T}}(t) \times \left\{ \begin{bmatrix} -y(1) + \theta^{\mathrm{T}}\varphi(1) \\ -y(2) + \theta^{\mathrm{T}}\varphi(2) \\ \vdots \\ -y(t) + \theta^{\mathrm{T}}\varphi(t) \end{bmatrix} - \Psi(t)\alpha_{k-1} \right\}, \tag{11}$$

$$\theta_k^{\mathrm{T}} = \theta_{k-1}^{\mathrm{T}} + \mu\{[y(1) + \psi(1)\alpha \;\; y(2) + \psi(2)\alpha \;\; \cdots \;\; y(t) + \psi(t)\alpha] - \theta_{k-1}^{\mathrm{T}}\Phi(t)\}\Phi^{\mathrm{T}}(t). \tag{12}$$

Here, a difficulty arises in the expressions on the right-hand sides of (11) and (12) contain the unknown parameter matrix θ and unknown parameter vector α, respectively; so it is impossible to realize the iterative algorithm in (11) and (12). Our approach is based on the hierarchical identification principle: the unknown variables θ and α in (11) and (12) are replaced with their corresponding estimates at time $k-1$. Hence, we have the hierarchical gradient iterative (HGI) algorithm

$$\alpha_k = \alpha_{k-1} + \mu\Psi^{\mathrm{T}}(t) \times \left\{ \begin{bmatrix} -y(1) + \theta_{k-1}^{\mathrm{T}}\varphi(1) \\ -y(2) + \theta_{k-1}^{\mathrm{T}}\varphi(2) \\ \vdots \\ -y(t) + \theta_{k-1}^{\mathrm{T}}\varphi(t) \end{bmatrix} - \Psi(t)\alpha_{k-1} \right\},$$

$$\theta_k^{\mathrm{T}} = \theta_{k-1}^{\mathrm{T}} + \mu\{[y(1) + \psi(1)\alpha_{k-1} \;\; y(2) + \psi(2)\alpha_{k-1} \;\; \cdots \;\; y(t) + \psi(t)\alpha_{k-1}] - \theta_{k-1}^{\mathrm{T}}\Phi(t)\}\Phi^{\mathrm{T}}(t).$$

or

$$\alpha_k = \alpha_{k-1} - \mu \sum_{i=1}^{t} \psi^{\mathrm{T}}(i)[y(i) + \psi(i)\alpha_{k-1} - \theta_{k-1}^{\mathrm{T}}\varphi(i)], \tag{13}$$

$$\theta_k^{\mathrm{T}} = \theta_{k-1}^{\mathrm{T}} + \mu \sum_{i=1}^{t} [y(i) + \psi(i)\alpha_{k-1} - \theta_{k-1}^{\mathrm{T}}\varphi(i)]\varphi^{\mathrm{T}}(i). \tag{14}$$

The convergence factor can be taken as

$$\mu = \left(\sum_{i=1}^{t} [\|\psi(i)\|^2 + \|\varphi(i)\|^2] \right)^{-1} =: \mu_0. \tag{15}$$

Here, the norm of the matrix X is defined by $\|X\|^2 = \mathrm{tr}[XX^{\mathrm{T}}]$. To initialize the algorithm in (13)–(15), we take $\alpha_0 = \mathbf{0}$ or some small real vector, e.g., $\alpha_0 = 10^{-6}\mathbf{1}_{n\times 1}$, and $\theta_0^{\mathrm{T}} = \mathbf{0}$ or some small real matrix, e.g., $\theta_0^{\mathrm{T}} = 10^{-6}\mathbf{1}_{m\times n_0}$ with $\mathbf{1}_{m\times n_0}$ being an $m \times n_0$ matrix whose elements are all 1.

The HGI algorithm employs the iterative update of the estimates $\hat{\alpha}$ and $\hat{\theta}$ using a fixed data batch with a finite length t. In this paper, in order to distinguish on-line from off-line calculation, we use *iterative* with subscript k, e.g., $\hat{\alpha}_k$, for off-line algorithms, and *recursive* with no subscript, e.g., $\hat{\alpha}(t)$ in the next section, for on-line ones. We imply that a recursive algorithm can be on-line implemented, but an iterative one cannot.

If two different convergence factors (μ_1 and μ_2) are used for iterations (9) and (10), then the algorithm obtained will be more general, but the convergence proof may be more difficult.

The dimensions of the HGI algorithm variables are listed in Table 1 for convenience. For the convergence analysis, we need a preliminary result.

Table 1
The dimensions of the HGI algorithm variables

Item	Variables	Dimensions
1	Output variable	$y(t) \in \mathbb{R}^m$
2	Parameter vector	$\alpha \in \mathbb{R}^n$
3	Information matrix	$\psi(t) \in \mathbb{R}^{m\times n}$
4	Stacked information matrix	$\Psi(t) \in \mathbb{R}^{(mt)\times n}$
5	Parameter matrix	$\theta^{\mathrm{T}} \in \mathbb{R}^{m\times n_0}$
6	Information vector	$\varphi(t) \in \mathbb{R}^{n_0}$
7	Stacked information matrix	$\Phi(t) \in \mathbb{R}^{n_0\times t}$

Lemma 1. *Assume that there exist vector sequences $x(k) \in \mathbb{R}^n$ and $\phi(i) \in \mathbb{R}^n$ satisfying*

$$\phi^{\mathrm{T}}(i)x(k) = 0, \quad \text{as} \;\; k \to \infty,$$

for each $i \in [1, t]$ ($t \gg n$), and that the vector $\phi(i)$ is sufficiently rich, i.e., there exist a positive constant c and an integer $N \geqslant n$ such that, for any $i \geqslant N$, the following inequality holds:

(A1) $\frac{1}{N}\sum_{l=1}^{N} \phi(i+l)\phi^{\mathrm{T}}(i+l) \geqslant cI$.

Then $\lim_{k\to\infty} x(k) = \mathbf{0}$.

Proof. Letting $\varepsilon(k, i) = \phi^{\mathrm{T}}(i)x(k)$, we have $\lim_{k\to\infty} \varepsilon(k, i) = 0$. Since $\phi^{\mathrm{T}}(i + l)x(k) = \varepsilon(k, i + l)$, after taking the norm $\|\cdot\|^2$ of both sides of the above equation, the summation from $l = 1$ to N is

$$x^{\mathrm{T}}(k)\left[\sum_{l=1}^{N} \phi(i+l)\phi^{\mathrm{T}}(i+l)\right]x(k) = \sum_{l=1}^{N} \varepsilon^2(k, i+l).$$

By using condition (A1), it follows that

$$0 \leqslant cN\|x(k)\|^2 \leqslant \sum_{i=1}^{N} \varepsilon^2(k, i+l).$$

Taking the limit of both sides of the above inequality will obtain the conclusion of Lemma 1 based on properties of the limiting process. □

Theorem 1. *For the system in* (3) *and the HGI algorithm in* (13)–(15)*, for any given initial values* α_0 *and* θ_0*, then the parameter estimation error given by the HGI algorithm is bounded, i.e.,*

$$\|\alpha_k-\alpha\|^2+\|\theta_k-\theta\|^2\leqslant\delta_0,\quad \textit{for any } k\geqslant 1.$$

Here, $\delta_0=\|\alpha_0-\alpha\|^2+\|\theta_0-\theta\|^2<\infty$.

Proof. Define the parameter estimation error vector $\tilde{\alpha}_k$ and the parameter estimation error matrix $\tilde{\theta}_k$ as

$$\tilde{\alpha}_k=\alpha_k-\alpha \quad\text{and}\quad \tilde{\theta}_k=\theta_k-\theta.$$

Using (13), (14) and (3), we have

$$\tilde{\alpha}_k=\tilde{\alpha}_{k-1}-\mu\sum_{i=1}^{t}\psi^{\mathrm{T}}(i)[\psi(i)\tilde{\alpha}_{k-1}-\tilde{\theta}_{k-1}^{\mathrm{T}}\varphi(i)],$$
$$\tilde{\theta}_k^{\mathrm{T}}=\tilde{\theta}_{k-1}^{\mathrm{T}}+\mu\sum_{i=1}^{t}[\psi(i)\tilde{\alpha}_{k-1}-\tilde{\theta}_{k-1}^{\mathrm{T}}\varphi(i)]\varphi^{\mathrm{T}}(i).$$

Using the formulae $\|x+y\|^2=\|x\|^2+2\mathrm{tr}[x^{\mathrm{T}}y]+\|y\|^2$ and $\|x^{\mathrm{T}}y\|^2\leqslant\|x\|^2\|y\|^2$, it is easy to obtain

$$\begin{aligned}\|\tilde{\alpha}_k\|^2&=\tilde{\alpha}_k^{\mathrm{T}}\tilde{\alpha}_k=\|\tilde{\alpha}_{k-1}\|^2\\&\quad-2\mu\sum_{i=1}^{t}\tilde{\alpha}_{k-1}^{\mathrm{T}}\psi^{\mathrm{T}}(i)[\psi(i)\tilde{\alpha}_{k-1}-\tilde{\theta}_{k-1}^{\mathrm{T}}\varphi(i)]\\&\quad+\mu^2\left\|\sum_{i=1}^{t}\psi^{\mathrm{T}}(i)[\psi(i)\tilde{\alpha}_{k-1}-\tilde{\theta}_{k-1}^{\mathrm{T}}\varphi(i)]\right\|^2\\&\leqslant\|\tilde{\alpha}_{k-1}\|^2-2\mu\sum_{i=1}^{t}\\&\quad[\psi(i)\tilde{\alpha}_{k-1}]^{\mathrm{T}}[\psi(i)\tilde{\alpha}_{k-1}-\tilde{\theta}_{k-1}^{\mathrm{T}}\varphi(i)]\\&\quad+\mu^2\sum_{i=1}^{t}\|\psi(i)\|^2\\&\quad\times\sum_{i=1}^{t}\|\psi(i)\tilde{\alpha}_{k-1}-\tilde{\theta}_{k-1}^{\mathrm{T}}\varphi(i)\|^2,\end{aligned}\tag{16}$$

$$\begin{aligned}\|\tilde{\theta}_k\|^2&=\mathrm{tr}[\tilde{\theta}_k^{\mathrm{T}}\tilde{\theta}_k]=\|\tilde{\theta}_{k-1}\|^2+2\mu\mathrm{tr}\\&\quad\left\{\sum_{i=1}^{t}[\psi(i)\tilde{\alpha}_{k-1}-\tilde{\theta}_{k-1}^{\mathrm{T}}\varphi(i)]\varphi^{\mathrm{T}}(i)\tilde{\theta}_{k-1}\right\}\\&\quad+\mu^2\left\|\sum_{i=1}^{t}[\psi(i)\tilde{\alpha}_{k-1}-\tilde{\theta}_{k-1}^{\mathrm{T}}\varphi(i)]\varphi^{\mathrm{T}}(i)\right\|^2\\&\leqslant\|\tilde{\theta}_{k-1}\|^2+2\mu\sum_{i=1}^{t}\\&\quad[\tilde{\theta}_{k-1}^{\mathrm{T}}\varphi(i)]^{\mathrm{T}}[\psi(i)\tilde{\alpha}_{k-1}-\tilde{\theta}_{k-1}^{\mathrm{T}}\varphi(i)]\\&\quad+\mu^2\sum_{i=1}^{t}\|\varphi(i)\|^2\\&\quad\times\sum_{i=1}^{t}\|\psi(i)\tilde{\alpha}_{k-1}-\tilde{\theta}_{k-1}^{\mathrm{T}}\varphi(i)\|^2.\end{aligned}\tag{17}$$

Define a non-negative definite function

$$V(k)=\|\tilde{\alpha}_k\|^2+\|\tilde{\theta}_k\|^2.$$

Using (16) and (17) gives

$$\begin{aligned}V(k)&\leqslant V(k-1)-2\mu\sum_{i=1}^{t}\\&\quad\|\psi(i)\tilde{\alpha}_{k-1}-\tilde{\theta}_{k-1}^{\mathrm{T}}\varphi(i)\|^2\\&\quad+\mu^2\sum_{i=1}^{t}[\|\psi(i)\|^2+\|\varphi(i)\|^2]\\&\quad\times\sum_{i=1}^{t}\|\psi(i)\tilde{\alpha}_{k-1}-\tilde{\theta}_{k-1}^{\mathrm{T}}\varphi(i)\|^2\\&=V(k-1)-\mu\left\{2-\mu\sum_{i=1}^{t}[\|\psi(i)\|^2+\|\varphi(i)\|^2]\right\}\\&\quad\times\sum_{i=1}^{t}\|\psi(i)\tilde{\alpha}_{k-1}-\tilde{\theta}_{k-1}^{\mathrm{T}}\varphi(i)\|^2\\&=V(0)-\mu\left\{2-\mu\sum_{i=1}^{t}[\|\psi(i)\|^2+\|\varphi(i)\|^2]\right\}\\&\quad\times\sum_{j=0}^{k-1}\sum_{i=1}^{t}\|\psi(i)\tilde{\alpha}_j-\tilde{\theta}_j^{\mathrm{T}}\varphi(i)\|^2.\end{aligned}$$

If the convergence factor μ is chosen to satisfy

$$0<\mu<2\mu_0,$$

then $V(k)\leqslant V(0)=\delta_0$. This proves Theorem 1. □

Furthermore, from the above proof we also have

$$\sum_{k=0}^{\infty}\sum_{i=1}^{t}\|\psi(i)\tilde{\alpha}_k-\tilde{\theta}_k^{\mathrm{T}}\varphi(i)\|^2<\infty.$$

It follows that as $k\to\infty$,

$$\sum_{i=1}^{t}\|\psi(i)\tilde{\alpha}_k-\tilde{\theta}_k^{\mathrm{T}}\varphi(i)\|^2=0\quad\text{or}$$
$$\psi(i)\tilde{\alpha}_k-\tilde{\theta}_k^{\mathrm{T}}\varphi(i)=0\quad\text{for any } i\in[1,t].\tag{18}$$

According to this equation, we can obtain the consistent parameter estimates from the following theorem.

Theorem 2. *For the system in* (3) *and the HGI algorithm in* (13)–(15)*, if the input–output data vectors*

$$\phi_j(i):=\begin{bmatrix}\psi_j^{\mathrm{T}}(i)\\ \varphi(i)\end{bmatrix},\quad j=1,2,\ldots,m$$

are sufficiently rich, where $\psi_j(i)$ is the jth row of $\psi(i)$, then the parameter estimation error given by the HGI algorithm converges to zero for any finite initial values, i.e.,

$$\lim_{k\to\infty} \|\alpha_k - \alpha\|^2 + \|\theta_k - \theta\|^2 = 0 \quad or$$
$$\lim_{k\to\infty} \alpha_k = \alpha \; and \; \lim_{k\to\infty} \theta_k = \theta.$$

Proof. Let $(\tilde{\theta}_k^{\mathrm{T}})_j$ represent the jth row of $\tilde{\theta}_k^{\mathrm{T}}$, and

$$x_j(k) := \begin{bmatrix} \tilde{\alpha}_k \\ -(\tilde{\theta}_k^{\mathrm{T}})_j^{\mathrm{T}} \end{bmatrix}.$$

Then (18) may be decomposed into the following m equations:

$$\phi_j^{\mathrm{T}}(i)x_j(k) = 0, \quad j = 1, 2, \ldots, m \text{ as } k \to \infty.$$

Since $\phi_j(i)$ $(j = 1, 2, \ldots, m)$ is sufficiently rich, according to Lemma 1, we have

$$\lim_{k\to\infty} x_j(k) = 0, \quad j = 1, 2, \ldots, m.$$

This proves Theorem 2. □

The off-line HGI algorithm here is developed for deterministic MIMO systems; so the convergence results of these two theorems are obvious. The HGI algorithm can be extended to stochastic cases.

4. The hierarchical stochastic gradient algorithm

In this section, we derive a hierarchical gradient parameter estimation algorithm based on the model discussed in (3) in the stochastic framework, and establish the convergence properties of the algorithm.

Based on the model in (3) and introducing a noise term $w(t)$, we have

$$y(t) + \psi(t)\alpha = \theta^{\mathrm{T}}\varphi(t) + w(t). \tag{19}$$

We assume that $\{w(t), \mathscr{F}_t\}$ is a martingale difference vector sequence defined on a probability space $\{\Omega, \mathscr{F}, \mathscr{P}\}$, where $\{\mathscr{F}_t\}$ is the σ algebra sequence generated by $\{w(t)\}$, i.e.,

$$\mathscr{F}_t = \sigma(w(t), w(t-1), w(t-2), \ldots) \quad \text{or}$$
$$\mathscr{F}_t = \sigma(y(t), y(t-1), y(t-2), \ldots)$$

for the deterministic sequence $\{u(t)\}$. The sequence $\{w(t)\}$ satisfies (Goodwin & Sin, 1984):

(A2) $\mathrm{E}[w(t)|\mathscr{F}_{t-1}] = 0$ a.s.;
(A3) $\mathrm{E}[\|w(t)\|^2|\mathscr{F}_{t-1}] = \sigma_w^2(t) \leqslant \sigma_w^2 < \infty$ a.s.;
(A4) $\lim\sup_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} \|w(i)\|^2 \leqslant \sigma_w^2 < \infty$ a.s.

That is, $w(t)$ is a noise vector with zero mean and time-varying variances. Thus, the system in (19) may involve non-stationary signals.

In order to derive the hierarchical gradient algorithm, which can be on-line implemented, we need to introduce two intermediate vectors $Y(t)$ and $Z(t)$ as follows:

$$Y(t) := y(t) - \theta^{\mathrm{T}}\varphi(t), \tag{20}$$

$$Z(t) := y(t) + \psi(t)\alpha. \tag{21}$$

According to the hierarchical identification principle, we can decompose the system in (19) into the following two fictitious subsystems:

$$S_3: \quad Y(t) = -\psi(t)\alpha + w(t), \tag{22}$$

$$S_4: \quad Z(t) = \theta^{\mathrm{T}}\varphi(t) + w(t). \tag{23}$$

Here, $Y(t) \in \mathbb{R}^m$, $\psi(t) \in \mathbb{R}^{m\times n}$ and $\alpha \in \mathbb{R}^n$ in (22) can be regarded as the output vector, information matrix and parameter vector of subsystem S_3, respectively. Similarly, $Z(t) \in \mathbb{R}^m$, $\varphi(t) \in \mathbb{R}^{n_0}$ and $\theta^{\mathrm{T}} \in \mathbb{R}^{m\times n_0}$ in (23) as the output vector, information vector and parameter matrix of subsystem S_4, respectively. Then, for the subsystems S_3 in (22) and S_4 in (23), we can form two prediction error criteria (Ljung, 1999; Söderström & Stoica, 1988)

$$J_1(\alpha) = \|Y(t) + \psi(t)\alpha\|^2 \quad \text{and}$$
$$J_2(\theta) = \|Z(t) - \theta^{\mathrm{T}}\varphi(t)\|^2.$$

Let $\hat{\alpha}(t)$ and $\hat{\theta}(t)$ be the estimates of α and θ at time t. Using the steepest descent gradient method, we obtain the estimates of α in subsystem S_3 and θ in subsystem S_4 by minimizing $J_1(\alpha)$ and $J_2(\theta)$, respectively, as follows:

$$\hat{\alpha}(t) = \hat{\alpha}(t-1) - \frac{\psi^{\mathrm{T}}(t)}{r(t)}[Y(t) + \psi(t)\hat{\alpha}(t-1)], \tag{24}$$

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \frac{\varphi(t)}{r(t)}[Z(t) - \hat{\theta}^{\mathrm{T}}(t-1)\varphi(t)]^{\mathrm{T}}. \tag{25}$$

Here, $1/r(t)$ represents a time-varying convergence factor to be given later. Substituting (20) into (24), (21) into (25) gives

$$\hat{\alpha}(t) = \hat{\alpha}(t-1) - \frac{\psi^{\mathrm{T}}(t)}{r(t)} \times [y(t) - \theta^{\mathrm{T}}\varphi(t) + \psi(t)\hat{\alpha}(t-1)], \tag{26}$$

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \frac{\varphi(t)}{r(t)} \times [y(t) + \psi(t)\alpha - \hat{\theta}^{\mathrm{T}}(t-1)\varphi(t)]^{\mathrm{T}}. \tag{27}$$

Here, we can see that the expressions on the right-hand sides of (26) and (27) contain the unknown θ and unknown α, respectively. As in the preceding section, by using hierarchical identification principle, we replace these unknown variables θ in (26) and α in (27) with their corresponding estimates $\hat{\theta}$ and $\hat{\alpha}$ at time $(t-1)$. Hence, we have

$$\hat{\alpha}(t) = \hat{\alpha}(t-1) - \frac{\psi^{\mathrm{T}}(t)}{r(t)} \times [y(t) + \psi(t)\hat{\alpha}(t-1) - \hat{\theta}^{\mathrm{T}}(t-1)\varphi(t)], \tag{28}$$

$$\hat{\theta}(t)=\hat{\theta}(t-1)+\frac{\varphi(t)}{r(t)} \times[y(t)+\psi(t)\hat{\alpha}(t-1)-\hat{\theta}^{\mathrm{T}}(t-1)\varphi(t)]^{\mathrm{T}}. \quad (29)$$

As in Goodwin and Sin (1984), we take $r(t)$ to be

$$r(t)=r(t-1)+\|\psi(t)\|^2+\|\varphi(t)\|^2, \quad r(0)=1. \quad (30)$$

Then we obtain the hierarchical stochastic gradient (HSG) algorithm in (28)–(30). The choice of the initial values of the HSG algorithm is as in the HGI algorithm.

The standard least squares (LS) algorithm may be applied to generate the parameter estimate of models of form: $y(t)=\Theta^{\mathrm{T}}\Phi(t)+w(t)$, Θ representing a parameter vector or matrix, $\Phi(t)$ the information vector. In general, the estimate can be expressed as (Ljung, 1999)

$$\begin{aligned}\hat{\Theta}(t)&=\hat{\Theta}(t-1)+L(t)[y^{\mathrm{T}}(t)-\Phi^{\mathrm{T}}(t)\hat{\Theta}(t-1)]\\&=\hat{\Theta}(t-1)+E(t),\end{aligned}$$

where $E(t)=L(t)[y^{\mathrm{T}}(t)-\Phi^{\mathrm{T}}(t)\hat{\Theta}(t-1)]$ is the innovation vector, $L(t)$ denotes the gain vector. The LS algorithm employs the idea of the innovation modification, i.e., the estimate $\hat{\Theta}(t)$ at time t equals the estimate $\hat{\Theta}(t-1)$ at time $t-1$ plus the innovation $E(t)$. But the hierarchical identification produce two estimates: a vector $\hat{\alpha}(t)$ and a matrix $\hat{\theta}(t)$. The estimate $\hat{\alpha}(t)$ at time t depends not only on $\hat{\alpha}(t-1)$ but also on $\hat{\theta}(t-1)$; similarly, the estimate $\hat{\theta}(t)$ at time t depends not only on $\hat{\theta}(t-1)$ but also on $\hat{\alpha}(t-1)$.

The following is to prove the convergence of the HSG algorithm by formulating a martingale process and by using the martingale convergence theorem in (Goodwin & Sin, 1984, Lemma D.5.3) rather than the martingale hyperconvergence theorem (Ding, Yang, & Xu, 2000).

Lemma 2. *Assume that the vector sequences $x(t)\in\mathbb{R}^n$ and $\phi(t)\in\mathbb{R}^n$ satisfy the following equations*:

$$\phi^{\mathrm{T}}(t)x(t)=0 \quad \text{as} \quad t\to\infty;$$

$$\lim_{t\to\infty}[x(t)-x(t-j)]=0, \quad \textit{for any } 0<j<\infty \text{ a.s.};$$

and the vector $\phi(t)$ is persistently exciting, i.e., there exist a positive constant c, c_1 and an integer $N\geqslant n$ such that the following persistent excitation condition holds:

$$\text{(A5)} \quad cI\leqslant\frac{1}{N}\sum_{i=1}^{N}\phi(t+i)\phi^{\mathrm{T}}(t+i)\leqslant c_1 I \quad \text{a.s.} \quad \textit{for any } t\geqslant 0.$$

Then $\lim_{t\to\infty}x(t)=0$.

Proof. Letting $\varepsilon(t+j)=x(t+j)-x(t)$ or $x(t+j)=x(t)+\varepsilon(t+j)$, we have $\lim_{t\to\infty}\varepsilon(t)=0$. In the same way, let $\varepsilon_1(t)=\phi^{\mathrm{T}}(t)x(t)$, we have $\lim_{t\to\infty}\varepsilon_1(t)=0$. So

$$\begin{aligned}&\phi^{\mathrm{T}}(t+i)x(t+i)=\varepsilon_1(t+i), \quad \text{or}\\&\phi^{\mathrm{T}}(t+i)x(t)=-\phi^{\mathrm{T}}(t+i)\varepsilon(t+i)+\varepsilon_1(t+i).\end{aligned}$$

After taking the norm $\|\cdot\|^2$ of both sides of the above equation, the summation from $i=1$ to N is

$$\begin{aligned}&x^{\mathrm{T}}(t)\left[\sum_{i=1}^{N}\phi(t+i)\phi^{\mathrm{T}}(t+i)\right]x(t)\\&=\sum_{i=1}^{N}\|-\phi^{\mathrm{T}}(t+i)\varepsilon(t+i)+\varepsilon_1(t+i)\|^2\\&\leqslant 2\sum_{i=1}^{N}[\|\phi(t+i)\|^2\varepsilon^2(t+i)+\varepsilon_1^2(t+i)].\end{aligned}$$

Taking the trace of condition (A5) will lead to

$$\|\phi(t)\|^2\leqslant nNc_1=:\delta_m<\infty.$$

Using condition (A5), we have

$$0\leqslant cN\|x(t)\|^2\leqslant 2\sum_{i=1}^{N}[\delta_m\varepsilon^2(t+i)+\varepsilon_1^2(t+i)].$$

Taking the limit of both sides of the above inequality will give the conclusion of Lemma 2 according to limit existence criterion. □

Lemma 3. *For the HSG algorithm in* (28)–(30), *the following inequality holds*:

$$s:=\sum_{t=1}^{\infty}\frac{\|\psi(t)\|^2+\|\varphi(t)\|^2}{r^2(t)}<\infty \quad \text{a.s.}$$

Proof. According to the definition of $r(t)$, we have

$$\begin{aligned}s&\leqslant\sum_{t=1}^{\infty}\frac{\|\psi(t)\|^2+\|\varphi(t)\|^2}{r(t-1)r(t)}\\&=\sum_{t=1}^{\infty}\left[\frac{1}{r(t-1)}-\frac{1}{r(t)}\right]\\&=\frac{1}{r(0)}-\frac{1}{r(\infty)}<\infty, \quad \text{a.s.}\end{aligned}$$

This completes the proof of Lemma 3. □

Theorem 3. *For the system in* (19) *and the HSG algorithm in* (28)–(30), *if Assumptions* (A2)–(A4) *hold, then the parameter estimation error given by the HSG algorithm is consistently bounded, i.e.,*

$$\|\hat{\alpha}(t)-\alpha\|^2+\|\hat{\theta}(t)-\theta\|^2\to W_0<\infty \quad \text{a.s. as } t\to\infty.$$

Here, $\mathrm{E}[W_0]\leqslant\|\hat{\alpha}(0)-\alpha\|^2+\|\hat{\theta}(0)-\theta\|^2+\frac{\sigma_w^2}{r(0)}<\infty$.

Proof. Define the parameter estimation error vector $\tilde{\alpha}(t)$ and the parameter estimation error matrix $\tilde{\theta}(t)$ as

$$\tilde{\alpha}(t) = \hat{\alpha}(t) - \alpha, \tag{31}$$

$$\tilde{\theta}(t) = \hat{\theta}(t) - \theta. \tag{32}$$

Substituting (19) and (28) into (31) gives

$$\begin{aligned}\tilde{\alpha}(t) &= \tilde{\alpha}(t-1) - \frac{\psi^{\mathrm{T}}(t)}{r(t)}[-\psi(t)\alpha + \theta^{\mathrm{T}}\varphi(t) \\ &\quad + w(t) + \psi(t)\hat{\alpha}(t-1) - \hat{\theta}^{\mathrm{T}}(t-1)\varphi(t)] \\ &= \tilde{\alpha}(t-1) - \frac{\psi^{\mathrm{T}}(t)}{r(t)}[\xi(t) - \eta(t) + w(t)],\end{aligned} \tag{33}$$

where

$$\xi(t) = \psi(t)\hat{\alpha}(t-1) - \psi(t)\alpha = \psi(t)\tilde{\alpha}(t-1),$$
$$\eta(t) = \hat{\theta}^{\mathrm{T}}(t-1)\varphi(t) - \theta^{\mathrm{T}}\varphi(t) = \tilde{\theta}^{\mathrm{T}}(t-1)\varphi(t).$$

Substituting (19) and (29) into (32) gives

$$\tilde{\theta}(t) = \tilde{\theta}(t-1) + \frac{\varphi(t)}{r(t)}[\xi(t) - \eta(t) + w(t)]^{\mathrm{T}}. \tag{34}$$

Define the stochastic Lyapunov function

$$W(t) = \|\tilde{\alpha}(t)\|^2 + \|\tilde{\theta}(t)\|^2.$$

Using (33) and (34) gives

$$\begin{aligned}W(t) &= W(t-1) - \frac{2}{r(t)} \\ &\quad \times [\|\xi(t) - \eta(t)\|^2 + (\xi(t) - \eta(t))^{\mathrm{T}} w(t)] \\ &\quad + [\xi(t) - \eta(t) + w(t)]^{\mathrm{T}} \frac{\psi(t)\psi^{\mathrm{T}}(t) + \|\varphi(t)\|^2 I}{r^2(t)} \\ &\quad \times [\xi(t) - \eta(t) + w(t)] \\ &\leqslant W(t-1) - \frac{2}{r(t)}\|\xi(t) - \eta(t)\|^2 \\ &\quad - \frac{2}{r(t)}(\xi(t) - \eta(t))^{\mathrm{T}} w(t) \\ &\quad + \frac{\|\psi(t)\|^2 + \|\varphi(t)\|^2}{r^2(t)}[\|\xi(t) - \eta(t)\|^2 + \|w(t)\|^2] \\ &\quad + 2[\xi(t) - \eta(t)]^{\mathrm{T}} \frac{\psi(t)\psi^{\mathrm{T}}(t) + \|\varphi(t)\|^2 I}{r^2(t)} w(t) \\ &\leqslant W(t-1) - \frac{1}{r(t)}\|\xi(t) - \eta(t)\|^2 \\ &\quad + \frac{\|\psi(t)\|^2 + \|\varphi(t)\|^2}{r^2(t)}\|w(t)\|^2 - 2[\xi(t) - \eta(t)]^{\mathrm{T}} \\ &\quad \times \frac{[r(t) - \|\varphi(t)\|^2]I - \psi(t)\psi^{\mathrm{T}}(t)}{r^2(t)} w(t).\end{aligned} \tag{35}$$

Since $\xi(t) - \eta(t)$, $r(t)$, $\psi(t)$, $\varphi(t)$ are uncorrelated with $w(t)$, and are $\mathscr{F}_{t-1}$ measurable, taking the conditional expectation of both sides of (35) with respect to $\mathscr{F}_{t-1}$ and using assumptions (A2)–(A4) give

$$\begin{aligned}\mathrm{E}[W(t)|\mathscr{F}_{t-1}] &\leqslant W(t-1) - \frac{1}{r(t)}\|\xi(t) - \eta(t)\|^2 \\ &\quad + \frac{\|\psi(t)\|^2 + \|\varphi(t)\|^2}{r^2(t)}\sigma_w^2.\end{aligned} \tag{36}$$

Since the sum of the last term on the right-hand side from $t=1$ to ∞ is finite (from Lemma 3), applying the martingale convergence theorem to (36) shows that $W(t)$ converges a.s. to a finite random variable W_0. This proves Theorem 3. □

Furthermore, suppose that $r(t) = O(t) \to \infty$ and $\|\psi(t)\|^2 + \|\varphi(t)\|^2 < \infty$, by referring to Lai and Ying (1991), it can be proved that $W(t) - W(t-1)$ converges a.s. to zero at the rate of $(1/t^2)$, and $\|\xi(t) - \eta(t)\|^2$ converges a.s. to zero at the rate of $(1/t)$, i.e.,

$$\|\xi(t) - \eta(t)\|^2 = 0, \text{ a.s. as } t \to \infty. \tag{37}$$

Theorem 4. *For the system in* (19) *and the HSG algorithm in* (28)–(30)*, if the conditions of Theorem* 3 *hold, and the input–output data vectors*

$$\phi_i(t) := \begin{bmatrix}\psi_i^{\mathrm{T}}(t) \\ \varphi(t)\end{bmatrix}, \quad i = 1, 2, \ldots, m$$

are persistently exciting, where $\psi_i(t)$ is the ith row of $\psi(t)$, then the parameter estimation error given by the HSG algorithm consistently converges to zero, i.e.,

$$\lim_{t\to\infty} \|\hat{\alpha}(t) - \alpha\|^2 + \|\hat{\theta}(t) - \theta\|^2 = 0 \text{ a.s.},$$
$$\textit{or } \lim_{t\to\infty} \hat{\alpha}(t) = \alpha \text{ a.s., } \textit{and } \lim_{t\to\infty} \hat{\theta}(t) = \theta \text{ a.s.}$$

Proof. Since $\phi_i(t)$, $i = 1, 2, \ldots, m$, are persistently exciting, then

$$r(t) = O(t) \to \infty \quad \text{as } t \to \infty. \tag{38}$$

From (37), we have

$$\xi(t) = \eta(t) \quad \text{as } t \to \infty,$$

or

$$\psi(t)\tilde{\alpha}(t-1) = \tilde{\theta}^{\mathrm{T}}(t-1)\varphi(t), \quad \text{as } t \to \infty. \tag{39}$$

Let $\tilde{\theta}_i^{\mathrm{T}}(t-1)$ represent the ith row of $\tilde{\theta}^{\mathrm{T}}(t-1)$, and

$$x_i(t) := \begin{bmatrix}\tilde{\alpha}(t-1) \\ -\tilde{\theta}_i(t-1)\end{bmatrix}.$$

Then (39) can be decomposed into the following m equations:

$$\phi_i^{\mathrm{T}}(t)x_i(t) = 0, \quad i = 1, 2, \ldots, m, \quad \text{as } t \to \infty. \tag{40}$$

From (A4), (37), (38), (33) and (34), it is very easy to obtain

$$\lim_{t\to\infty}[x_i(t) - x_i(t-j)] = 0, \quad \text{for any } 0 < j < \infty. \tag{41}$$

According to Lemma 2, (40) and (41) give the conclusions of Theorem 4. □

The HSG algorithm has low computational effort, but its convergence is slow, just like the stochastic gradient algorithm of scalar systems in Goodwin and Sin (1984). In order to improve the convergence rate and tracking performance of the HSG algorithms, we introduce a forgetting factor λ in (30) to get

$$r(t)=\lambda r(t-1)+\|\psi(t)\|^2+\|\varphi(t)\|^2,\quad 0\leqslant\lambda\leqslant 1,\quad r(0)=1, \tag{42}$$

and obtain the HSG algorithm with forgetting factor (FFHSG algorithm for short) in (28), (29) and (42). When $\lambda=1$, the FFHSG algorithm reduces to the HSG algorithm; when $\lambda=0$, the FFHSG algorithm is the hierarchical projection algorithm.

5. Comparing computational efficiency

In this section, we compare in detail the computational efficiency of our HSG algorithm with several existing ones: the recursive LS and stochastic gradient algorithms based on the model in (4), the stochastic gradient algorithm based on m subsystems.

The LS algorithm of identifying θ_s in (4), in the stochastic framework, namely, the Sen and Sinha's algorithm, can be expressed as

$$\text{I:}\begin{cases}\hat{\theta}_s(t) &= \hat{\theta}_s(t-1)+L(t)[y(t)-H(t)\hat{\theta}_s(t-1)],\\ &\quad \hat{\theta}_s(t)\in\mathbb{R}^{mn_0+n},\\ L(t) &= P(t)H^{\mathrm{T}}(t)\\ &= P(t-1)H^{\mathrm{T}}(t)[I_m+H(t)P(t-1)\\ &\quad \times H^{\mathrm{T}}(t)]^{-1},\\ P(t) &= [I_{mn_0+n}-L(t)H(t)]P(t-1).\end{cases}$$

The stochastic gradient algorithm of identifying θ_s in (4) is given by

$$\text{II}:\begin{cases}\hat{\theta}_s(t) &= \hat{\theta}_s(t-1)+\frac{H^{\mathrm{T}}(t)}{R(t)}\\ &\quad \times[y(t)-H(t)\hat{\theta}_s(t-1)],\\ R(t) &= R(t-1)+\|H(t)\|^2,\ R(0)=1.\end{cases}$$

Introducing a noise vector $w(t)$, the model in (2) may be decomposed into m subsystems,

$$\alpha(z)y_j(t)=Q_j(z)u(t)+w_j(t),\quad j=1,2,\ldots,m, \tag{43}$$

where $Q_j(z)$ represents the jth row of $Q(z)$, $y_j(t)$ the jth element of $y(t)$; then each subsystem contains the same parameters for $\alpha(z)$. Any identification method applied to each subsystem would generate m different estimates of $\alpha(z)$, which is certainly undesirable, because it leads to increased computation, although one may take their average as the estimate of α. This is also a motivation for us to develop the hierarchical identification algorithm.

Eq. (43) can be written as a vector form

$$y_j(t)=[-\psi_j(t)\ \ \varphi^{\mathrm{T}}(t)]\begin{bmatrix}\alpha\\ \theta_j\end{bmatrix}+w_j(t),\quad \begin{bmatrix}\alpha\\ \theta_j\end{bmatrix}\in\mathbb{R}^{n+n_0}, \tag{44}$$

where $\psi_j(t)$ and θ_j^{T} represent the jth row of $\psi(t)$ and θ^{T}, respectively.

Based on the model in (44), a comparable stochastic gradient algorithm with the HSG algorithm is as follows:

$$\text{III}:\begin{cases}\begin{bmatrix}\hat{\alpha}(t)\\ \hat{\theta}_j(t)\end{bmatrix}\\ \quad=\begin{bmatrix}\hat{\alpha}(t-1)\\ \hat{\theta}_j(t-1)\end{bmatrix}+\frac{1}{r_j(t)}\begin{bmatrix}-\psi_j^{\mathrm{T}}(t)\\ \varphi(t)\end{bmatrix}\\ \quad\times\left(y_j(t)-[-\psi_j(t)\ \ \varphi^{\mathrm{T}}(t)]\begin{bmatrix}\hat{\alpha}(t-1)\\ \hat{\theta}_j(t-1)\end{bmatrix}\right),\\ r_j(t)=r_j(t-1)+\|\psi_j(t)\|^2+\|\varphi(t)\|^2,\\ r_j(0)=1,\ \ j=1,2,\ldots,m.\end{cases} \tag{45}$$

From here, we can see that for a MISO system, i.e., $m=1$, the hierarchical algorithm in (28)–(30) reduces to a non-hierarchical one in (45). This is also the reason why we take the same step-size μ in the HGI algorithm or $1/r(t)$ in the HSG algorithm.

The computation loads of the four algorithms are listed in Table 2, where numbers of multiplications and additions are for each iteration step, and the numbers in the brackets in Table 2 denote the recorded numbers for a 10-input, 10-output and 10th-order system at each step. From Table 2, it is clear that the HSG algorithm is computationally more efficient than other algorithms.

6. Example

In this section, we present an example to illustrate the performance of the proposed algorithms.

Consider the following simulated plant:

$$\alpha(z)y(t)=Q(z)u(t)+w(t),$$

where

$$\alpha(z)=1+\alpha_1z^{-1}+\alpha_2z^{-2}+\alpha_3z^{-3},$$
$$Q(z)=Q_1z^{-1}+Q_2z^{-2}+Q_3z^{-3},$$
$$\alpha=[\alpha_1\ \ \alpha_2\ \ \alpha_3]^{\mathrm{T}}=[-1.15\ \ 0.425\ \ -0.05]^{\mathrm{T}},$$
$$\theta^{\mathrm{T}}=[Q_1\ \ Q_2\ \ Q_3]=\begin{bmatrix}1 & 1 & -0.9 & -0.75 & 0.2 & 0.125\\ 1.2 & 1.2 & -1.08 & -0.78 & 0.24 & 0.12\end{bmatrix}.$$

Here $u(t)=[u_1(t),\ u_2(t)]^{\mathrm{T}}$ is taken as a persistent excitation vector sequence with zero mean and unit variances, and $w(t)=[w_1(t),\ w_2(t)]^{\mathrm{T}}$ as a white noise vector sequence with zero mean and variances $[\sigma_w^2(1),\ \sigma_w^2(2)]$, and

Table 2
Comparing computational efficiency

Algorithms	Number of multiplications	Number of additions
I[a]	$2m(mn_0+n)^2+2m(m+1)(mn_0+n)$ [24886200]	$2m(mn_0+n+m)(mn_0+n)-m^2+m$ [24863910]
II	$3m(mn_0+n)+m$ [33310]	$3m(mn_0+n)$ [33300]
III	$3m(n+n_0)+m$ [3610]	$3m(n+n_0)$ [3600]
HSG	$2mn_0+3mn+n_0$ [2610]	$2mn_0+3mn+n_0$ [2610]

[a]This does not contain computing the inverse of the $m\times m$ matrix in the gain matrix $L(t)$.

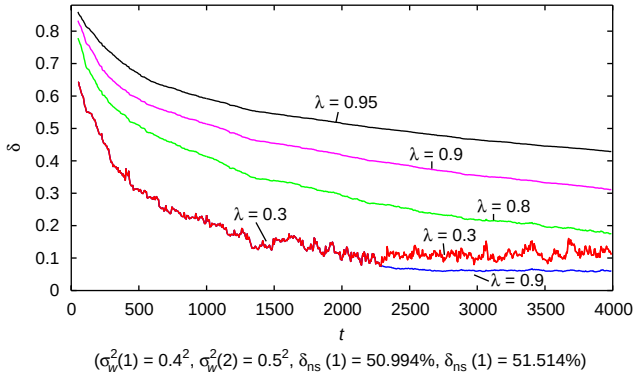

Fig. 1. The estimation errors δ vs. t with different forgetting factors.

$w_1(t)$ is uncorrelated with $w_2(t)$. Taking the initial values, $\hat{\alpha}(0)=10^{-6}\mathbf{1}_{3\times 1}$ and $\hat{\theta}(0)=10^{-6}\mathbf{1}_{2\times 6}$, we apply the HSG algorithm with forgetting factor to estimate the parameters of this system. The estimation errors δ with different forgetting factors versus t are shown in Fig. 1, where

$$\delta=\sqrt{\frac{\|\hat{\alpha}(t)-\alpha\|^2+\|\hat{\theta}(t)-\theta\|^2}{\|\alpha\|^2+\|\theta\|^2}}$$

is the relative parameter estimation error, $\delta_{\mathrm{ns}}(1)$ and $\delta_{\mathrm{ns}}(2)$ are the noise-to-signal ratios of two output channels, respectively.

From Fig. 1, we can see that as the forgetting factor λ is increased, the rate of change of the parameter estimates (or the estimation error) becomes more stationary, but the estimation error gets larger. In other words, if we decrease the forgetting factor λ, the convergence rate of the parameter estimation is faster initially, but the variance of the estimation error becomes larger. Therefore, a compromise/tradeoff is to choose a smaller forgetting factor at the initial period of the operation, and then let the forgetting factor gradually increase with t, and finally approach 1 so that more accurate parameter estimates are obtained. For example, in the bottom curve in Fig. 1, if we take

$$\lambda=\begin{cases}0.3 & t<2300,\\ 0.9, & t\geqslant 2300,\end{cases}$$

we can obtain good convergence rate as well as acceptable stationarity in the estimation error. From Fig. 1, it is clear that as long as we choose appropriate forgetting factors, δ is becoming smaller (in general) as t increases, this verifies the theorems proposed. Moreover, from simulation (not included in this paper), we find that the correlation noises will degrade the estimation accuracy.

7. Conclusions

According to a hierarchical identification principle, a HGI algorithm and HSG are developed for MIMO systems. The analysis indicates that the algorithms proposed can achieve good performance properties (i.e., the parameter estimation errors are uniformly bounded, and consistently converges to zero under persistent excitation), and require less computational efforts than the existing algorithms.

Although the algorithms are proposed for MIMO stochastic systems with an additive white noise disturbance, the methods developed can be easily extended to study stochastic systems with colored noises. The estimation error bound analysis of the HGI algorithm with noise and the HSG algorithm with forgetting factors are currently being studied in the stochastic framework. Finally, the simulated results verify the theoretical findings.

Acknowledgements

The authors are grateful to the Associate Editor and anonymous reviewers for their helpful comments and suggestions.

References

Chen, T., & Qiu, L. (1994). $\mathscr{H}_\infty$ design of general multirate sampled-data control systems. *Automatica*, *30*(7), 1139–1152.

Cho, Y. M., Xu, G., & Kailath, T. (1994). Fast recursive identification of state space models via exploitation of displacement structure. *Automatica*, *30*(1), 45–49.

Chou, C. T., & Verhaegen, M. (1997). Subspace algorithms for the identification of multivariable dynamic error-in-variable models. *Automatica*, *33*(10), 1857–1869.

Comon, P., & Golub, G. (1990). Tracking a few extreme singular values and vectors in signal processing. *Proceedings of the IEEE*, *78*(8), 1327–1343.

Ding, F., Yang, J. B., & Xu, Y. M. (2000). Hierarchical identification and its convergence for the transfer function matrix. *Proceedings of IFAC Symposium on System Identification* (*SYSID 2000*), 21–23 June 2000, Santa Barbara, CA, USA.

Drouin, M., Abou-Kandil, H., & Mariton, M. (1991). *Control of complex systems: methods and technology*. New York, London: Plenum Press, (Applied Information Technology. Series Editor: M. G. Singh).

El-Sherief, H. (1981). Parametric identification of a state-space model of multivariable systems using the extended least-squares method. *IEEE Transactions on Systems Man, and Cybernetics*, *11*(3), 223–227.

El-Sherief, H., & Sinha, N. K. (1979). Choices of models for the identification of linear multivariable discrete-time systems. *IEEE Proceedings, Part D, Control Theory and Applications*, *126*(12), 1326–1330.

Gauthier, A., & Landau, I. D. (1978). On the recursive identification of multi-input multi-output systems. *Automatica*, *14*(6), 609–614.

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

Gustafsson, T. (1998). Instrumental variable subspace tracking using projection approximation. *IEEE Transactions on Signal Processing*, *46*(3), 669–681.

Lai, T. L., & Ying, Z. L. (1991). Recursive identification and adaptive prediction in linear stochastic systems. *SIAM Journal on Control and Optimization*, *29*(5), 1061–1090.

Li, D., Shah, S. L., & Chen, T. (2001). Identification of fast-rate models from multirate data. *International Journal of Control*, *74*(7), 680–689.

Li, D., Shah, S. L., & Chen, T. (2002). Analysis of dual-rate inferential control systems. *Automatica*, *38*(6), 1053–1059.

Li, D., Shah, S. L., Chen, T., & Qi, K. Z. (2003). Application of dual-rate modeling to CCR octane quality inferential control. *IEEE Transactions on Control Systems Technology*, *11*(1), 43–51.

Ljung, L. (1999). *System identification: theory for the user*. Englewood Cliffs, NJ: Prentice-Hall.

Lovera, M., Gustafsson, T., & Verhaegen, M. (2000). Recursive subspace identification of linear and non-linear Wiener state-space models. *Automatica*, *36*(11), 1651–1658.

McKelvey, T., Akcay, H., & Ljung, L. (1996). Subspace-based multivariable system identification from frequency response data. *IEEE Transactions on Automatic Control*, *41*(7), 960–979.

Oku, H., & Kimura, H. (2002). Recursive 4SID algorithms using gradient type subspace tracking. *Automatica*, *38*(6), 1035–1043.

Overshee, P. V., & De Moor, B. (1994). N4SID: Subspace algorithm for the identification of combined determine-stochastic systems. *Automatica*, *30*(1), 75–93.

Overshee, P. V., & De Moor, B. (1996). *Subspace identification for linear systems: theory, implementation and applications*. Boston: Kluwer Academic Publisher.

Pintelon, R. (2002). Frequency-domain subspace system identification using non-parametric noise models. *Automatica*, *38*(8), 1295–1311.

Qiu, L., & Chen, T. (1994). $\mathscr{H}_\infty$-optimal design of multirate sampled-data systems. *IEEE Transactions on Automatic Control*, *39*(12), 2506–2511.

Qiu, L., & Chen, T. (1999). Multirate sampled-data systems: All $\mathscr{H}_\infty$ suboptimal controllers and the minimum entropy controllers. *IEEE Transactions on Automatic Control*, *44*(3), 537–550.

Sen, A., & Sinha, N. K. (1976). On-line estimation of the parameters of a multivariable system using matrix pseudoinverse. *International Journal of Systems Science*, *7*(4), 461–471.

Sheng, J., Chen, T., & Shah, S. L. (2002). Generalized predictive control for non-uniformly sampled systems. *Journal of Process Control*, *12*, 875–888.

Singh, M. G. (1980). *Dynamical hierarchical control*. New York: North-Holland Publishing Company.

Sinha, N. K., & Kwong, Y. H. (1979). Recursive estimation of the parameters of linear multivariable systems. *Automatica*, *15*, 471–475.

Söderström, T., & Stoica, P. (1988). *System identification*. Englewood-Cliffs, NJ: Prentice-Hall.

Tamura, H., & Yoshikawa, T. (1990). *Large-scale systems control and decision making*. New York and Basel: Marcel Dekker, Inc.

Tangirala, A. K., Li, D., Patwardhan, R. S., Shah, S. L., & Chen, T. (2001). Ripple-free conditions for lifted multirate control systems. *Automatica*, *37*(10), 1637–1645.

Van Overschee, P., & De Moor, B. (1996). Continuous time-frequency domain subspace system identification. *Signal Processing*, *52*(2), 179–194.

Verhaegen, M. (1993). Subspace model identification Part 3: analysis of the ordinary output-error state space model identification algorithm. *International Journal of Control*, *58*(3), 555–586.

Verhaegen, M. (1994). Identification of the deterministic part of MIMO state space models given in innovations form from input–output data. *Automatica*, *30*(1), 61–71.

Verhaegen, M., & Deprettere, E. (1991). A fast recursive MIMO state space model identification algorithm. *Proceedings of the 1991 conference on decision and control* (*30th CDC*), Brighton, USA (pp. 1349–1354).

Verhaegen, M., & Dewilde, P. (1992a). Subspace model identification Part 1: the output-error state space model identification class of algorithms. *International Journal of Control*, *56*(5), 1187–1210.

Verhaegen, M., & Dewilde, P. (1992b). Subspace model identification Part 2: analysis of the elementary output-error state space model identification algorithm. *International Journal of Control*, *56*(5), 1211–1241.

Yang, B. (1995). Projection approximation subspace tracking. *IEEE Transactions on Signal Processing*, *43*(1), 95–107.

Yang, B. (1996). Asymptotic convergence analysis of the projection approximation subspace tracking algorithm. *Signal Processing (EURASIP)*, *50*(1), 123–136.

Feng Ding was born in Guangshui, Hubei Province. He received the B.Sc. degree in Electrical Engineering from the Information Engineering College, Hubei University of Technology (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 1990 and 1994, respectively. From 1984 to 1988, he was an Electrical Engineer at the Hubei Pharmaceutical Factory. Since 1994 he is 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, China, and has been a Research Associate at the University of Alberta, Edmonton, Canada since 2002. His current research interests include model identification and adaptive control. 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 MASc and PhD 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 of Electrical Engineering. 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.