2009 American Control Conference
Hyatt Regency Riverfront, St. Louis, MO, USA
June 10-12, 2009

Hierarchical Stochastic Gradient Parameter Estimation Algorithms for Multivariable Systems with Colored Noises

Feng Ding, Yanjun Liu

Abstract— **This paper develops a hierarchical extended stochastic gradient identification algorithms for MIMO ARMAX-like systems to deal with colored noises based on the hierarchical identification principle. The convergence performance of such algorithms is studied in detail; in particular, conditions for parameter estimation errors to converge to zero are established, which include persistent excitation of the extended information vectors and strict positive realness of the noise models. Finally, the proposed algorithms are tested on an example to show their advantages and effectiveness.**

Index terms: Recursive identification, estimation, stochastic gradient, convergence properties, hierarchical identification principle,multivariable systems, ARMAX models

I. Introduction

Consider MIMO ARMAX-like systems of the form

$$\alpha(z)y(t) = Q(z)u(t) + D(z)v(t), \tag{1}$$

where $u(t) \in \mathbb{R}^r$ is the system input vector, $y(t) \in \mathbb{R}^m$ the system output vector, $v(t) \in \mathbb{R}^m$ an uncorrelated random noise vector with zero mean, $G(z) := \frac{Q(z)}{\alpha(z)} \in \mathbb{R}^{m\times r}$ the transfer matrix (TM) of the system, with z^{-1} representing the unit delay operator: $z^{-1}y(t) = y(t-1)$; $\alpha(z)$ is the characteristic polynomial in z^{-1} of the system (of degree n) defined as the monic least common denominator of $G(z)$, $Q(z)$ and $D(z)$ are polynomials (matrices) in z^{-1}, and they can be represented as

$$\begin{aligned}
\alpha(z) &= 1 + \alpha_1 z^{-1} + \alpha_2 z^{-2} + \cdots + \alpha_n z^{-n},\ \alpha_i \in \mathbb{R}^1,\\
Q(z) &= Q_0 + Q_1 z^{-1} + Q_2 z^{-2} + \cdots + Q_n z^{-n} \in \mathbb{R}^{m\times r},\\
&\qquad Q_i \in \mathbb{R}^{m\times r},\\
D(z) &= 1 + d_1 z^{-1} + d_2 z^{-2} + \cdots + d_n z^{-n},\ d_i \in \mathbb{R}^1.
\end{aligned}$$

Notice that the case with interactive noises in the outputs, namely, the case with $D(z)$ being a polynomial matrix, will be treated in a similar way. Assume that $u(t) = \mathbf{0}$, $y(t) = \mathbf{0}$ and $v(t) = \mathbf{0}$ as $t \leq 0$. The disturbance into the systems in (1), $w(t) := D(z)v(t)$, is colored (correlated) even if $v(t)$ is a white noise vector. The model in (1) forms an extension from the ARMAX model for scalar systems [1] to the multi-input, multi-output (MIMO) case; and is thus referred to as an ARMAX-like model.

The hierarchical identification (HI) principle [2], [3] is a very useful tool for studying parameter estimation of MIMO ARMAX-like models, in which there exist both a parameter vector α and a parameter matrix θ [see Equation (2) later] to be identified. The HI algorithms have advantages in computational efficiency over existing algorithms. Based on the HI principle, several identification algorithms, e.g., the hierarchical gradient algorithms [2] and the hierarchical least squares algorithms [3], were proposed recently for MIMO systems described by transfer matrices; however, a serious limitation is the white noise assumption with $D(z) = 1$. The main purpose of this paper is to propose more general identification algorithms, based on the HI principle, which are suitable for MIMO ARMAX-like models with *colored* noises [$D(z)$ being general polynomials instead of $D(z) = 1$], and to study the related performance issues. We consider this to be a significant step in making the new algorithms practical.

The identification problem for the ARMAX-like model in (1) with $D(z) \neq 1$ is much more difficult and challenging in that the information vector/matrix in the identification model contains unmeasurable noise vectors $v(t-1)$, $v(t-2)$, $\cdots$, $v(t-n)$ – see $\psi_0(t)$ in (2). Thus, the reported hierarchical algorithms in [2], [3] are not suitable in this case – this is main motivation for our work here.

The main contributions of this paper lie in the following:

- The hierarchical stochastic algorithm in [2] is extended to give a hierarchical extended stochastic gradient (HESG) algorithm for MIMO ARMAX-like systems by using the HI principle. The basic idea is to replace unmeasurable noise terms in the information vector/matrix by the estimation residuals like in the scalar ARMAX case [4].
- The convergence properties of the HESG algorithm using the stochastic martingale theory are studied in detail; in particular, conditions for the parameter estimation error to converge to zero are given, which include that the extended information vector is persistently exciting (i.e., the data product moment matrices have bounded condition numbers) and that the process noises $\{v(t)\}$ are zero mean and uncorrelated.
- The convergence conditions obtained in this paper are the weakest conditions for stochastic gradient algorithms known. The main convergence results in the paper do not assume that the noise variances and high-order moments exist and are finite. These results remove the strict assumptions, made in existing references, that the noise variances and high-order moments exist [5]–

This work was supported by the National Natural Science Foundation of China and the Natural Science Foundation of Jiangsu Province (China, BK2007017) and by Program for Innovative Research Team of Jiangnan University.

The authors are with the School of Communication and Control Engineering, Jiangnan University, Wuxi, P. R. China 214122. E-mail addresses: fding@jiangnan.edu.cn; yanjunliu_1983@126.com.

978-1-4244-4524-0/09/$25.00 ©2009 AACC

[10] and that processes are stationary and ergodic [4], [11].

The convergence analysis of the hierarchical algorithms in [2], [3] was based on the assumptions that both the noise variances are time-varying but bounded and that the strong persistent excitation conditions hold. This paper relaxes these conditions, and assumes only that the weak persistent excitation conditions hold, and does not require that the noise variances are bounded.

The paper is organized as follows. Sections II develops a hierarchical extended stochastic gradient (HESG) algorithm. Section III is the main results in this paper and studies the convergence properties in detail by using the martingale convergence theorem. Section IV presents an illustrative example for the results in this paper. Finally, in Section V, we offer some concluding remarks.

II. The hierarchical extended stochastic gradient algorithm

In this section, we derive the HESG parameter estimation algorithm for the ARMAX-like model.

Let us introduce some notation first. The symbol I (I_m) stands for an identity matrix of appropriate sizes ($m \times m$); the superscript T denotes the matrix transpose; the norm of a matrix X is defined by $\|X\|^2 = \mathrm{tr}[XX^{\mathrm{T}}]$; $\mathbf{1}_{m\times n}$ represents an $m \times n$ matrix whose elements are all 1, and $\mathbf{1}_n := \mathbf{1}_{n\times 1}$; $\lambda_{\max}[X]$ and $\lambda_{\min}[X]$ represent the maximum and minimum eigenvalues of the symmetric matrix X, respectively; for $g(t) \geq 0$, we write $f(t) = O(g(t))$ if there exists positive constants δ_1 and t_0 such that $|f(t)| \leq \delta_1 g(t)$ for $t \geq t_0$.

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

$$\begin{aligned}\theta^{\mathrm{T}} &:= [Q_0,\ Q_1,\ \cdots,\ Q_n] \in \mathbb{R}^{m\times n_0},\ \ n_0 := (n+1)r,\\ \alpha &:= [\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_n,\ d_1,\ d_2,\ \cdots,\ d_n]^{\mathrm{T}} \in \mathbb{R}^{2n},\\ \varphi(t) &:= [u^{\mathrm{T}}(t),\ u^{\mathrm{T}}(t-1),\ \cdots,\ u^{\mathrm{T}}(t-n)]^{\mathrm{T}} \in \mathbb{R}^{n_0},\\ \psi_0(t) &:= [y(t-1),\ y(t-2),\ \cdots,\ y(t-n),\\ &\quad -v(t-1),\ -v(t-2),\ \cdots,\ -v(t-n)] \in \mathbb{R}^{m\times(2n)}.\end{aligned}$$

In vector forms, Equation (1) can be written as

$$y(t) + \psi_0(t)\alpha = \theta^{\mathrm{T}}\varphi(t) + v(t). \tag{2}$$

Equation (2) is called the (hierarchical) identification model for MIMO ARMAX-like systems in (1).

The objective of this paper is, by means of the hierarchical identification principle, to present identification algorithms to estimate the unknown parameters (α, θ) in (2) from the given input-output measurement data $\{u(t),\ y(t):\ t = 1,\ 2,\ \cdots\}$, and to study convergence performance issues of the algorithms proposed in the stochastic framework.

Here, a difficulty arises because the information matrix $\psi_0(t)$ in (2) contains the unmeasurable noise terms $v(t-1)$, $v(t-2)$, $\cdots$, $v(t-n)$; so the hierarchical stochastic gradient algorithm in [2] cannot be applied directly to (2). The approach here is to use the estimation residuals $\hat{v}(t)$ to replace these noise terms $v(t)$. After doing such a replacement, $\psi_0(t)$ is denoted by $\psi(t)$. Let $\hat{\alpha}(t)$ and $\hat{\theta}(t)$ be the estimates of α and θ at time t, respectively; then the estimation residuals $\hat{v}(t)$ may be computed by

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

with

$$\begin{aligned}\psi(t) = [&y(t-1),\ y(t-2),\ \cdots,\ y(t-n),\\ &-\hat{v}(t-1),\ -\hat{v}(t-2),\ \cdots,\ -\hat{v}(t-n)].\end{aligned}$$

As in [2], by introducing two intermediate vectors and defining and minimizing two error criteria, it is easy to get the HESG algorithm of estimating α and θ as follows:

$$\begin{aligned}\hat{\alpha}(t) &= \hat{\alpha}(t-1) - \frac{\psi^{\mathrm{T}}(t)}{r(t)}[y(t) + \psi(t)\hat{\alpha}(t-1)\\ &\quad -\hat{\theta}^{\mathrm{T}}(t-1)\varphi(t)], && (3)\\ \hat{\theta}(t) &= \hat{\theta}(t-1) + \frac{\varphi(t)}{r(t)}[y(t) + \psi(t)\hat{\alpha}(t-1) -\\ &\quad \hat{\theta}^{\mathrm{T}}(t-1)\varphi(t)]^{\mathrm{T}}, && (4)\\ r(t) &= r(t-1) + \|\psi(t)\|^2 + \|\varphi(t)\|^2,\ r(0) = 1, && (5)\\ \varphi(t) &= [u^{\mathrm{T}}(t),\ u^{\mathrm{T}}(t-1),\ \cdots,\ u^{\mathrm{T}}(t-n)]^{\mathrm{T}}, && (6)\\ \psi(t) &= [y(t-1),\ y(t-2),\ \cdots,\ y(t-n),\\ &\quad -\hat{v}(t-1),\ -\hat{v}(t-2),\ \cdots,\ -\hat{v}(t-n)], && (7)\\ \hat{v}(t) &= y(t) + \psi(t)\hat{\alpha}(t) - \hat{\theta}^{\mathrm{T}}(t)\varphi(t). && (8)\end{aligned}$$

The initial values may be chosen to be some small real vector/matrix as in [2], e.g., $\hat{\alpha}(0) = 10^{-6}\mathbf{1}_{2n}$ and $\hat{\theta}(0) = 10^{-6}\mathbf{1}_{m\times n_0}$.

Because the algorithm in (3)-(8) is obtained by expanding the parameter vector α by adding the noise model parameters d_i, it is called the hierarchical *extended* stochastic gradient algorithm.

Although the HESG algorithm is simple, its convergence analysis is very challenging under the weak assumptions on the statistical properties of the noises. Next, without assuming that the noise variances and high-order moments exist and are finite, we establish the convergence properties of the HESG algorithm.

III. Main convergence results

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 the observation data up to and including time t [12]. The sequence $\{v(t)\}$ satisfies:

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

$(A2)$ $\mathrm{E}[\|v(t)\|^2|\mathscr{F}_{t-1}] = \sigma_v^2 r^{\varepsilon}(t),\ \sigma_v^2 < \infty,\ \varepsilon < 1,$ a.s.

Here, we do not assume that the noise vector $\{v(t)\}$ has finite variances and high-order moments.

Lemma 1: For the HESG algorithm in (3)-(8), define the innovation vector,

$$e(t) := y(t) + \psi(t)\hat{\alpha}(t-1) - \hat{\theta}^{\mathrm{T}}(t-1)\varphi(t). \tag{9}$$

Then the residual $\hat{v}(t)$ and innovation $e(t)$ are related by

$$\hat{v}(t)=W(t)e(t),$$

where

$$W(t):=I_m-\frac{\psi(t)\psi^{\mathrm{T}}(t)+\|\varphi(t)\|^2 I_m}{r(t)}.$$

Proof Substituting (3) and (4) into (8) yields

$$\begin{aligned}\hat{v}(t)&=y(t)+\psi(t)\left[\hat{\alpha}(t-1)-\frac{\psi^{\mathrm{T}}(t)}{r(t)}e(t)\right]\\&\quad-\left[\hat{\theta}(t-1)+\frac{\varphi(t)}{r(t)}e^{\mathrm{T}}(t)\right]^{\mathrm{T}}\varphi(t)\\&=e(t)-\frac{\psi(t)\psi^{\mathrm{T}}(t)}{r(t)}e(t)-\frac{\|\varphi(t)\|^2}{r(t)}e(t)\\&=\left[I_m-\frac{\psi(t)\psi^{\mathrm{T}}(t)+\|\varphi(t)\|^2 I_m}{r(t)}\right]e(t)=W(t)e(t).\end{aligned}$$

This proves Lemma 1.

Lemma 2: For the HESG algorithm in (3)-(8), the following inequality holds:

$$\sum_{i=1}^{t}\frac{\|\psi(i)\|^2+\|\varphi(i)\|^2}{r(i)}\leq \ln r(t),\ \text{a.s.}$$

Lemma 3: For the HESG algorithm in (3)-(8), the following inequality holds:

$$\begin{aligned}S_m:=\sum_{t=1}^{\infty}&\mathrm{tr}\left\{\frac{\psi(t)\psi^{\mathrm{T}}(t)+\|\varphi(t)\|^2 I_m}{r^2(t)}W^{-1}(t)\right\}\sigma_v^2 r^{\varepsilon}(t)\\&<\infty,\ \text{a.s.},\ \varepsilon<1.\end{aligned}\tag{10}$$

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

$$\begin{aligned}\tilde{\alpha}(t)&:=\hat{\alpha}(t)-\alpha,\\\tilde{\theta}(t)&:=\hat{\theta}(t)-\theta,\end{aligned}$$

and the incremental changes of $\tilde{\alpha}(t)$ and $\tilde{\theta}(t)$:

$$\begin{aligned}\Delta\tilde{\alpha}(t)&:=\hat{\alpha}(t)-\hat{\alpha}(t-1),\\\Delta\tilde{\theta}(t)&:=\hat{\theta}(t)-\hat{\theta}(t-1).\end{aligned}$$

Using (3)-(4) and (9), it follows that

$$\begin{aligned}\Delta\tilde{\alpha}(t)&=[\hat{\alpha}(t)-\alpha]-[\hat{\alpha}(t-1)-\alpha]\\&=\tilde{\alpha}(t)-\tilde{\alpha}(t-1)=-\frac{\psi^{\mathrm{T}}(t)}{r(t)}e(t),\qquad(11)\\\Delta\tilde{\theta}(t)&=[\hat{\theta}(t)-\theta]-[\hat{\theta}(t-1)-\theta]\\&=\tilde{\theta}(t)-\tilde{\theta}(t-1)=\frac{\varphi(t)}{r(t)}e^{\mathrm{T}}(t).\qquad(12)\end{aligned}$$

Or

$$\begin{aligned}\tilde{\alpha}(t)&=\tilde{\alpha}(t-1)-\Delta\tilde{\alpha}(t),\qquad(13)\\\tilde{\theta}(t)&=\tilde{\theta}(t-1)+\Delta\tilde{\theta}(t).\qquad(14)\end{aligned}$$

We state the main convergence results and show the proof by formulating a martingale process [13], [15], [16] and by using the martingale convergence theorem (Lemma D.5.3 in [12]).

Theorem 1: For the system in (2) and the HESG algorithm in (3)-(8), if Assumptions (A1) and (A2) hold, and

$$(A3)\quad D(z)\ \text{ is a strictly positive real function,}$$

then the parameter estimation errors of the HESG algorithm are bounded, i.e.,

$$\|\hat{\alpha}(t)-\alpha\|^2+\|\hat{\theta}(t)-\theta\|^2\leq V_0<\infty,\ \ \text{a.s.},$$

and further the parameter estimation differences converge, i.e.,

$$\sum_{t=i}^{\infty}\|\hat{\alpha}(t)-\hat{\alpha}(t-i)\|^2+\|\hat{\theta}(t)-\hat{\theta}(t-i)\|^2<\infty,\ \text{a.s.},\ i>0.$$

Proof Let

$$\begin{aligned}\xi(t)&:=\psi(t)\tilde{\alpha}(t),\qquad(15)\\\eta(t)&:=\tilde{\theta}^{\mathrm{T}}(t)\varphi(t).\qquad(16)\end{aligned}$$

Taking the norms of both sides of (13) and (14), respectively, gives

$$\begin{aligned}\|\tilde{\alpha}(t)\|^2&=\|\tilde{\alpha}(t-1)\|^2-\frac{2}{r(t)}\xi^{\mathrm{T}}(t)e(t)\\&\quad-e^{\mathrm{T}}(t)\frac{\psi(t)\psi^{\mathrm{T}}(t)}{r^2(t)}e(t),\qquad(17)\\\|\tilde{\theta}(t)\|^2&=\|\tilde{\theta}(t-1)\|^2+\frac{2}{r(t)}\eta^{\mathrm{T}}(t)e(t)\\&\quad-\frac{\|\varphi(t)\|^2}{r^2(t)}\|e(t)\|^2.\qquad(18)\end{aligned}$$

Define a stochastic Lyapunov function:

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

Using (17), (18), (13) and (14) gives

$$\begin{aligned}V(t)&=V(t-1)\\&-\frac{2}{r(t)}[\xi(t)-\eta(t)]^{\mathrm{T}}W^{-1}(t)[\hat{v}(t)-v(t)]\\&-\frac{2}{r(t)}[\psi(t)\tilde{\alpha}(t-1)-\tilde{\theta}^{\mathrm{T}}(t-1)\varphi(t)]^{\mathrm{T}}W^{-1}(t)v(t)\\&+2[e(t)-v(t)]^{\mathrm{T}}\frac{\psi(t)\psi^{\mathrm{T}}(t)+\|\varphi(t)\|^2 I_m}{r^2(t)}W^{-1}(t)v(t)\\&+2\mathrm{tr}\left\{\frac{\psi(t)\psi^{\mathrm{T}}(t)+\|\varphi(t)\|^2 I_m}{r^2(t)}W^{-1}(t)v(t)v^{\mathrm{T}}(t)\right\}\\&-e^{\mathrm{T}}(t)\frac{\psi(t)\psi^{\mathrm{T}}(t)+\|\varphi(t)\|^2 I_m}{r^2(t)}e(t).\end{aligned}\tag{19}$$

According to the definition of $\hat{v}(t)$, we have

$$D(z)[\hat{v}(t)-v(t)]=\xi(t)-\eta(t).\tag{20}$$

Hence

$$\begin{aligned}\xi(t)-\eta(t)&=[D(z)-\rho][\hat{v}(t)-v(t)]+\rho[\hat{v}(t)-v(t)]\\&=:\tilde{y}_1(t)+\rho[\hat{v}(t)-v(t)],\end{aligned}$$

where

$$\tilde{y}_1(t):=D_1(z)[\hat{v}(t)-v(t)],\quad D_1(z):=D(z)-\rho.$$

Since $D(z)$ is strictly positive real, there exists a small constant $\rho > 0$ such that $D_1(z)$ is also strictly positive real. Referring to Appendix C in [12], the following inequalities hold:

$$S_0(t) := \sum_{i=1}^{t} \frac{2\tilde{y}_1^{\mathrm{T}}(i)W^{-1}(i)[\hat{v}(t)-v(t)]}{r(i)} \geq 0, \text{ a.s.,}$$

$$\begin{aligned} S(t) &:= S_0(t)+\rho\sum_{i=1}^{t}\frac{2[\hat{v}(i)-v(i)]^{\mathrm{T}}W^{-1}(i)[\hat{v}(i)-v(i)]}{r(i)} \\ &= \sum_{i=1}^{t}\frac{2}{r(i)}[\xi(i)-\eta(i)]^{\mathrm{T}}W^{-1}(i)[\hat{v}(i)-v(i)] \geq 0, \text{ a.s.} \end{aligned}$$

Since $V(t-1)$, $r(t)$, $\xi(t)-\eta(t)$, $W(t)$, $\psi(t)$, $\varphi(t)$ and $e(t)-v(t)$ are uncorrelated with $v(t)$, and are $\mathscr{F}_{t-1}$ measurable, adding (19) by $S(t)$ and taking the conditional expectation with respect to $\mathscr{F}_{t-1}$ and using Assumptions (A1)-(A2) give

$$\begin{aligned} &\mathrm{E}[V(t)+S(t)|\mathscr{F}_{t-1}] = V(t-1)+S(t-1) \\ &+2\mathrm{tr}\left\{\frac{\psi(t)\psi^{\mathrm{T}}(t)+\|\varphi(t)\|^2 I_m}{r^2(t)}W^{-1}(t)\right\}\sigma_v^2 r^{\varepsilon}(t) \\ &-\mathrm{E}\left[e^{\mathrm{T}}(t)\frac{\psi(t)\psi^{\mathrm{T}}(t)+\|\varphi(t)\|^2 I_m}{r^2(t)}e(t)|\mathscr{F}_{t-1}\right]. \end{aligned} \tag{21}$$

Since the sum of the third term on the right-hand side from $t=1$ to $t=\infty$ is finite according to Lemma 3, applying the martingale convergence theorem to (21) shows that $V(t)$ converges a.s. to a finite random variable, say V_0, i.e.,

$$V(t)+S(t) = \|\tilde{\alpha}(t)\|^2+\|\tilde{\theta}(t)\|^2+S(t)\to V_0<\infty, \text{ a.s.,} \tag{22}$$

and also

$$\sum_{t=1}^{\infty} e^{\mathrm{T}}(t)\frac{\psi(t)\psi^{\mathrm{T}}(t)+\|\varphi(t)\|^2 I_m}{r^2(t)}e(t)<\infty, \text{ a.s.}$$

This means

$$\begin{aligned} &\sum_{t=1}^{\infty}\frac{\|\psi^{\mathrm{T}}(t)e(t)\|^2+\|\varphi(t)\|^2\|e(t)\|^2}{r^2(t)} \\ &\quad = \sum_{t=1}^{\infty}\|\Delta\tilde{\alpha}(t)\|^2+\|\Delta\tilde{\theta}(t)\|^2<\infty, \text{ a.s.} \end{aligned} \tag{23}$$

From (13) and (14), we have

$$\begin{aligned} \tilde{\alpha}(t) &= \tilde{\alpha}(t-i)-\sum_{j=0}^{i-1}\frac{\psi^{\mathrm{T}}(t-j)}{r(t-j)}e(t-j) \\ &= \tilde{\alpha}(t-i)-\sum_{j=0}^{i-1}\Delta\tilde{\alpha}(t-j), \end{aligned} \tag{24}$$

$$\begin{aligned} \tilde{\theta}(t) &= \tilde{\theta}(t-i)+\sum_{j=0}^{i-1}\frac{\varphi(t-j)}{r(t-j)}e^{\mathrm{T}}(t-j) \\ &= \tilde{\theta}(t-i)+\sum_{j=0}^{i-1}\Delta\tilde{\theta}(t-j), \ i\geq 1. \end{aligned} \tag{25}$$

Using (11)-(12) and (23), it is easy to get

$$\begin{aligned} &\|\tilde{\alpha}(t)-\tilde{\alpha}(t-i)\|^2 = \|\hat{\alpha}(t)-\hat{\alpha}(t-i)\|^2 \\ &= \left\|\sum_{j=0}^{i-1}\Delta\tilde{\alpha}(t-j)\right\|^2 \leq i\sum_{j=0}^{i-1}\|\Delta\tilde{\alpha}(t-j)\|^2<\infty, \text{ a.s.,} \\ &\|\tilde{\theta}(t)-\tilde{\theta}(t-i)\|^2 = \|\hat{\theta}(t)-\hat{\theta}(t-i)\|^2 \\ &= \left\|\sum_{j=0}^{i-1}\Delta\tilde{\theta}(t-j)\right\|^2 \leq i\sum_{j=0}^{i-1}\|\Delta\tilde{\theta}(t-j)\|^2<\infty, \text{ a.s.} \end{aligned}$$

Summing from $t=i$ to $t=\infty$ and using (23) give

$$\sum_{t=i}^{\infty}\|\tilde{\alpha}(t)-\tilde{\alpha}(t-i)\|^2 \leq i\sum_{j=0}^{i-1}\sum_{t=i}^{\infty}\|\Delta\tilde{\alpha}(t-j)\|^2<\infty,$$

$$\sum_{t=i}^{\infty}\|\tilde{\theta}(t)-\tilde{\theta}(t-i)\|^2 \leq i\sum_{j=0}^{i-1}\sum_{t=i}^{\infty}\|\Delta\tilde{\theta}(t-j)\|^2<\infty.$$

This indicates that the estimation differences are convergent and the relation in (22) shows that the estimation errors are bounded. This proves Theorem 1. □

Further, according to the definitions of $W(t)$ and $S(t)$, from (22), we have

$$\begin{aligned} &\sum_{i=1}^{t}\frac{\|\hat{v}(i)-v(i)\|^2}{r(i)} \\ &\leq \sum_{i=1}^{t}\frac{[\hat{v}(i)-v(i)]^{\mathrm{T}}W^{-1}(i)[\hat{v}(i)-v(i)]}{r(i)}<\infty, \text{ a.s.} \end{aligned} \tag{26}$$

Hence, using the Kronecker Lemma (Lemma D.5.5 in [12]) gives

$$\lim_{t\to\infty}\frac{1}{r(t)}\sum_{i=1}^{t}\|\hat{v}(i)-v(i)\|^2=0, \text{ a.s.,}$$

or

$$\sum_{i=1}^{t}\|\hat{v}(i)-v(i)\|^2=o(r(t)), \text{ a.s.} \tag{27}$$

Since $D(z)$ is strictly stable, applying Lemma B.3.3 in [12] to (20) and using the above inequality give

$$\sum_{i=1}^{t}\frac{\|\xi(i)-\eta(i)\|^2}{r(i)}<\infty, \text{ a.s.,}$$

or

$$\lim_{t\to\infty}\frac{1}{r(t)}\sum_{i=1}^{t}\|\xi(i)-\eta(i)\|^2=0, \text{ a.s.} \tag{28}$$

Let $\psi_{0k}(t)$ and $\psi_k(t)$ be the kth row of $\psi_0(t)$ and $\psi(t)$, respectively. Define the extended information vectors,

$$\phi_{0k}(t) := \begin{bmatrix}\psi_{0k}^{\mathrm{T}}(t)\\ \varphi(t)\end{bmatrix}, \ \phi_k(t) := \begin{bmatrix}\psi_k^{\mathrm{T}}(t)\\ \varphi(t)\end{bmatrix}, \ k=1,2,\cdots,m,$$

and the data product moment matrices,

$$R_k(t) := \sum_{i=1}^{t}\phi_k(i)\phi_k^{\mathrm{T}}(i), \ R_{0k}(t) := \sum_{i=1}^{t}\phi_{0k}(i)\phi_{0k}^{\mathrm{T}}(i).$$

Further, let θ_k and $\tilde{\theta}_k(t)$ be the kth column of θ and $\tilde{\theta}(t)$, respectively, and

$$r_0(t) := 1+\sum_{i=1}^{t}\|\psi_0(i)\|^2+\sum_{i=1}^{t}\|\varphi(i)\|^2.$$

The following is to establish the consistent convergence of the parameter estimation.

Theorem 2: For the system in (2) and the HESG algorithm in (3)-(8), suppose that the conditions of Theorem 1 hold, $r(t) \to \infty$, and

$$(A4) \quad \limsup_{t\to\infty} \frac{r_0(t)}{\lambda_{\min}[R_{0k}(t)]} < \infty, \text{ a.s.}$$

Then the parameter estimation error given by the HESG algorithm consistently converges to zero, i.e.,

$$\|\hat{\alpha}(t)-\alpha\|^2 + \|\hat{\theta}_k(t)-\theta_k\|^2$$
$$= o\left(\frac{r_0(t)}{\lambda_{\min}[R_{0k}(t)]}\right) \to 0, \text{ a.s., } k=1,2,\cdots,m.$$

This means

$$\lim_{t\to\infty}\hat{\alpha}(t)=\alpha, \text{ a.s., and } \lim_{t\to\infty}\hat{\theta}(t)=\theta, \text{ a.s.}$$

The proof is omitted due to the limited space but available from the authors.

Condition (A4) shows that the data product moment matrices $R_{0k}(t)$, $k=1,2,\cdots,m$, have bounded condition numbers, i.e., the information vectors $\phi_{0k}(t)$, consisting of the input-output and noise data, are persistently exciting. If the input vector $u(t)$ is taken as a pseudo-random binary sequence or uncorrelated random signal vector sequence, and $v(t)$ as a white noise vector sequence with zero mean and constant variances $[\sigma_v^2(1), \sigma_v^2(2), \cdots, \sigma_v^2(m)]$ (i.e., $\varepsilon = 0$), then the weak persistent excitation condition (A4) is automatically satisfied because $u(t)$ is a persistent excitation signal vector and the white noises are best persistent excitation signals [1].

Earlier convergence analysis of identification algorithms assume that the noises are independent and identically distributed random sequences with finite 4th-order moments and the input and output signals have finite nonzero power [11], or that the noises are stationary and ergodic and have constant variances [4], or that the noise variances and high-order moments exist and are finite [5]–[10]. We think that studying consistent convergence of second-order moments of parameter estimation errors, it is neither necessary nor reasonable to assume existence of higher-order moments [2], [3], [13], [14]. The assumptions in (A1) and (A2) imply that $v(t)$ is a non-stationary uncorrelated noise vector with zero mean. Theorems 1 and 2 do not require such assumptions as stationarity and ergodicity, or existence of higher-order moments. Thus, the process in (1) can be possibly non-stationary and non-ergodic. The convergence analysis in this paper again does not assume that the noise variances are finite. Therefore, assumptions (A1)-(A4) represent the weakest conditions, to our best knowledge, which guarantee the consistent convergence of the parameter estimation errors for stochastic gradient algorithms.

When $\varepsilon = 0$ in (A2), i.e., the stationary noise case, if there exist positive constants c_1, c_2 and t_0 such that, for $t \geq t_0$, the following weak persistent excitation condition (bounded condition number) holds:

$$(A4') \quad c_1 I \leq \frac{1}{t}\sum_{i=1}^{t}\phi_k(i)\phi_k^{\mathrm{T}}(i) \leq c_2 I, \text{ a.s., } \quad k=1,2,\cdots,m.$$

Then $\|\hat{\alpha}(t)-\alpha\|^2 + \|\hat{\theta}(t)-\theta\|^2 \to 0$, a.s.

The HESG algorithm has low computational effort, but its convergence is slow, just like the stochastic gradient algorithm of scalar systems in [12]. In order to improve the convergence rate and tracking performance of the HESG algorithm, we introduce a forgetting factor λ and obtain the HESG algorithm with a forgetting factor (FFHESG algorithm for short) as follows:

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

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \frac{\varphi(t)}{r(t)}[y(t)+\psi(t)\hat{\alpha}(t-1) - \hat{\theta}^{\mathrm{T}}(t-1)\varphi(t)]^{\mathrm{T}}, \tag{30}$$

$$r(t) = \lambda\, r(t-1) + \|\psi(t)\|^2 + \|\varphi(t)\|^2, \quad r(0)=1,\ 0\leq\lambda\leq 1, \tag{31}$$

$$\varphi(t) = [u^{\mathrm{T}}(t),\ u^{\mathrm{T}}(t-1),\ \cdots,\ u^{\mathrm{T}}(t-n)], \tag{32}$$

$$\psi(t) = [y(t-1),\ y(t-2),\ \cdots,\ y(t-n),\ -\hat{v}(t-1),\ \hat{v}(t-2),\ \cdots,\ -\hat{v}(t-n)], \tag{33}$$

$$\hat{v}(t) = y(t) + \psi(t)\hat{\alpha}(t) - \hat{\theta}^{\mathrm{T}}(t)\varphi(t). \tag{34}$$

When $\lambda = 1$, the FFHESG algorithm reduces to the HESG algorithm; when $\lambda = 0$, the FFHESG algorithm is the hierarchical projection algorithm.

IV. Example

In this section, we present an example to illustrate the performance of the proposed algorithms. Consider the following 2-input and 2-output system:

$$\begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} - 0.85\begin{bmatrix} y_1(t-1) \\ y_2(t-1) \end{bmatrix}$$
$$= \begin{bmatrix} 2.00 & 1.00 \\ 1.00 & 2.00 \end{bmatrix}\begin{bmatrix} u_1(t-1) \\ u_2(t-1) \end{bmatrix} + \begin{bmatrix} v_1(t) \\ v_2(t) \end{bmatrix} + 0.60\begin{bmatrix} v_1(t-1) \\ v_2(t-1) \end{bmatrix}.$$

$$\alpha = \begin{bmatrix} \alpha_1 \\ d_1 \end{bmatrix} = \begin{bmatrix} -0.85 \\ 0.60 \end{bmatrix},$$
$$\theta^{\mathrm{T}} = Q_1 = \begin{bmatrix} \beta_{11}(1) & \beta_{12}(1) \\ \beta_{21}(1) & \beta_{22}(1) \end{bmatrix} = \begin{bmatrix} 2.00 & 1.00 \\ 1.00 & 2.00 \end{bmatrix}.$$

In simulation, the inputs $u_1(t)$ and $u_2(t)$ both are taken as two independent persistent excitation sequences with zero mean and unit variances, and $v_1(t)$ and $v_2(t)$ as two white noise sequences with zero mean and variances $\sigma_v^2(1)$ and $\sigma_v^2(2)$, respectively. Applying the HESG algorithm with a forgetting factor to estimate the parameters of this system, the estimation errors δ with different forgetting factors ($\lambda =$

$1.00, 0.95, 0.90)$ versus t are shown in Figs. 1 and 2, where the relative parameter estimation error δ is defined as

$$\delta = \sqrt{\frac{\|\hat{\alpha}(t)-\alpha\|^2 + \|\hat{\theta}(t)-\theta\|^2}{\|\alpha\|^2 + \|\theta\|^2}} \times 100\%,$$

Changing the noise variances $\sigma_v^2(1)$ and $\sigma_v^2(2)$ results in adjusting the noise-to-signal ratios $\delta_{\rm ns}(1)$ and $\delta_{\rm ns}(2)$ of the two output channels. When $\sigma_v^2(1) = 0.50^2$ and $\sigma_v^2(2) = 0.50^2$, the noise-to-signal ratios are $\delta_{\rm ns}(1) = 34.50\%$ and $\delta_{\rm ns}(2) = 34.50\%$; when $\sigma_v^2(1) = 1.50^2$ and $\sigma_v^2(2) = 1.50^2$, the noise-to-signal ratios are $\delta_{\rm ns}(1) = 103.49\%$ and $\delta_{\rm ns}(2) = 103.49\%$.

From Figs. 1 and 2, we can draw the following conclusions:

- It is clear that the estimation errors δ are becoming smaller (in general) as t increases. In other words, increasing data length generally results in smaller parameter estimation errors.
- A high noise level results in a slow rate of convergence for the parameter estimation.
- As long as we choose appropriate forgetting factors, the estimation errors δ are becoming smaller (in general) as t increases.
- From Figs. 1 and 2, we can see that as the forgetting factor λ is increased, the estimation error becomes larger. However, 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. Thus, a better strategy 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 can be obtained.
- The simulation results confirm the stated theoretical results in Section 3.

V. Conclusions

According to the hierarchical identification principle, hierarchical extended stochastic gradient algorithms are presented for MIMO ARMAX-like systems. The analysis indicates that the algorithms proposed have good convergence properties under weaker conditions. Finally, the simulation results verify the theoretical findings.

References

[1] L. Ljung, *"System Identification: Theory for the User,"* Englewood Cliffs, NJ: Prentice-Hall, 1999.

[2] F. Ding and T. Chen, "Hierarchical gradient-based identification of multivariable discrete-time systems," *Automatica*, vol. 41, no. 2, pp. 315-325, 2005.

[3] F. Ding and T. Chen, "Hierarchical least squares identification methods for multivariable systems," *IEEE Transactions on Automatic Control*, vol. 50, no. 3, pp. 397-402, 2005.

[4] V. Solo, "The convergence of AML," *IEEE Transactions on Automatic Control*, vol. 24, no. 6, pp. 958-962, 1979.

[5] L. Guo, "Convergence and logarithm laws of self-tuning regulators," *Automatica*, vol. 31, no. 3, pp. 435-450, 1995.

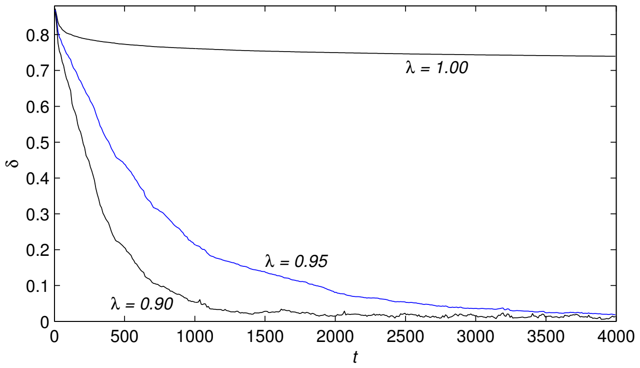

Fig. 1. The errors δ vs. t with different λ $[\sigma_v^2(1) = \sigma_v^2(2) = 0.50^2]$

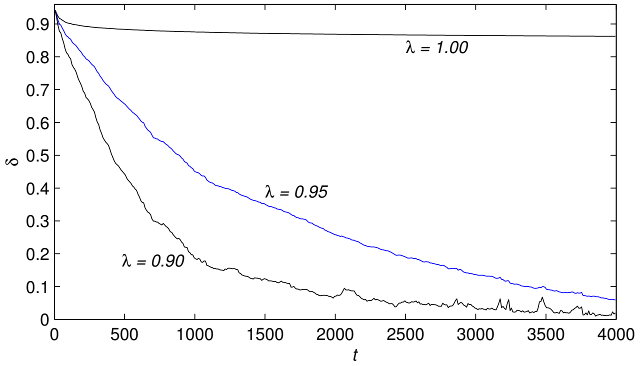

Fig. 2. The errors δ vs. t with different λ $[\sigma_v^2(1) = \sigma_v^2(2) = 1.50^2]$

[6] T.L. Lai and C.Z. Wei, "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems," *The Annals of Statistics*, vol. 10, no. 1, pp. 154-166, 1982.

[7] T.L. Lai and C.Z. Wei, "Extended least squares and their applications to adaptive control and prediction in linear systems," *IEEE Transactions on Automatic Control*, vol. 31, no. 10, pp. 898-906, 1986.

[8] T.L. Lai, and Z.L. Ying, "Recursive identification and adaptive prediction in linear stochastic systems," *SIAM Journal on Control and Optimization*, vol. 29, no. 5, pp. 1061-1090, 1991.

[9] W. Ren and P.K. Kumar, "Stochastic adaptive prediction and model reference control," *IEEE Transactions on Automatic Control*, vol. 39, no. 10, pp. 2047-2060, 1994.

[10] C.Z. Wei, "Adaptive prediction by least squares prediction in stochastic regression models," *The Annals of Statistics*, vol. 15, no. 4, pp. 1667-1682, 1987.

[11] L. Ljung, "Consistency of the least-squares identification method," *IEEE Transactions on Automatic Control*, vol. 21, no. 5, pp. 779-781, 1976.

[12] G.C. Goodwin and K.S. Sin, *Adaptive Filtering Prediction and Control*, Englewood Cliffs, NJ: Prentice-Hall, 1984.

[13] F. Ding and T. Chen, "Combined parameter and output estimation of dual-rate systems using an auxiliary model," *Automatica*, vol. 40, no. 10, pp. 1739-1748, 2004.

[14] F. Ding and T. Chen, "Performance bounds of forgetting factor least squares algorithm for time-varying systems with finite measurement data," *IEEE Transactions on Circuits and Systems-I: Regular Papers*, vol. 52, no. 3, pp. 555-566, 2005.

[15] F. Ding, T. Chen, "Identification of Hammerstein nonlinear ARMAX systems," *Automatica*, vol. 41, no. 9, pp. 1479-1489, 2005.

[16] F. Ding, T. Chen, "Parameter estimation of dual-rate stochastic systems by using an output error method," *IEEE Transactions on Automatic Control*, vol. 50, no. 9, pp. 1436-1441, 2005.