An auxiliary model based stochastic gradient algorithm for multivariable output error systems

DING Feng[1, 3] LIU Peter X.[2, 3]

Abstract The identification problem of multivariable output error systems is considered in this paper. By constructing an auxiliary model using available input-output data and by replacing the unknown inner variables in the information vector with the outputs of the auxiliary model, an auxiliary model based stochastic gradient (AM-SG) identification algorithm is presented. Convergence analysis using the martingale convergence theorem indicates that the parameter estimates given by the AM-SG algorithm converge to their true values. In order to improve the convergence rate of the AM-SG algorithm, the AM-SG algorithm with a forgetting factor is given. The simulation results confirm the theoretical findings.

Key words Recursive identification, parameter estimation, stochastic gradient, auxiliary model identification idea, multivariable systems, convergence properties, martingale convergence theorem

Parameter estimation has had wide applications in many areas, including signal processing, adaptive prediction and control, time-series analysis, process modelling, and so on. In the area of system identification, Zheng used the bias correction method in the identification of linear dynamic error-in-variable systems [1]; Yang and Zhang made comparisons of some bias compensation methods and other identification approaches for Box-Jenkins models [2], and Zhang and Yang presented a bias compensation recursive least squares identification for output error systems with colored noises [3]; Zong et al studied the iterative identification problem related to control design [4]; Zhong and Song discussed the hierarchical optimization identification for linear state space systems [5].

The least squares identification algorithms have fast convergence rates. Recently, Wang presented an auxiliary model based recursive extended least squares identification method for output error moving average systems [6]. However, the stochastic gradient (SG) parameter estimation algorithms have less computation load and have received much attention in self-tuning control and system identification. In the literature, many gradient based identification approaches were reported. For example, Ding and Chen proposed a hierarchical stochastic gradient algorithm for multivariable systems [7] and a multi-innovation stochastic gradient algorithm for linear regression model [8]. Wang and Ding developed an auxiliary model based multi-innovation generalized extended stochastic gradient identification algorithm for Box-Jenkins models [9] using the multi-innovation identification theory [8], but no convergence analysis was carried out. Also, Wang and Ding gave an extended stochastic gradient identification algorithm for Hammerstein-Wiener nonlinear ARMAX Systems [10].

This paper studies the gradient based identification approach and convergence for multivariable systems with output measurement noises. For such a system, the difficulty of identification is that the information vector contains unmeasurable variables. Our solution is to use the auxiliary model technique [11, 12]: to replace these unknown variables with the outputs of the auxiliary model, to present an auxiliary model based stochastic gradient (AM-SG) identification algorithm and further to analyze the convergence of the proposed algorithms. In order to improve the convergence rate of the gradient based algorithm, an AM-SG algorithm with a forgetting factor is given (the AM-FFSG algorithm for short). Compared with the auxiliary mode based recursive least squares algorithm, the AM-FFSG algorithm requires less computation burden. The simulation results indicate that if we choose appropriately the forgetting factor, the AM-FFSG algorithm can achieve a faster convergence rate and the parameter estimation accuracy is closer to that of the least squares algorithm. The AM-SG algorithm with the unknown information vector for the output error systems in this paper differs from the standard stochastic gradient algorithm in [13] which assumes that each entry of the information vector is known.

Briefly, the rest of the paper are organized as follows. Section 1 simply describes the identification problem to be discussed in the paper. Section 2 derives a basic identification algorithm for multivariable systems based on the auxiliary model technique. Sections 3 analyzes the performance of the proposed algorithm. Section 4 presents an illustrative example for the results in this paper. Finally, concluding remarks are given in Section 5.

1 Problem formulation

Consider a multivariable (i.e. MIMO: multi-input multi-output) output error system

$$\boldsymbol{x}(t) = G(z)\boldsymbol{u}(t), \tag{1}$$

$$\begin{aligned}\boldsymbol{y}(t) &= \boldsymbol{x}(t) + \boldsymbol{v}(t), \\ &= G(z)\boldsymbol{u}(t) + \boldsymbol{v}(t),\end{aligned} \tag{2}$$

which is different from the equation error systems (CAR/ARX model) in [13], where $\boldsymbol{u}(t) \in \mathbf{R}^r$ is the system input vector, $\boldsymbol{x}(t) \in \mathbf{R}^m$ the system output vector (the true output or noise-free output), $\boldsymbol{y}(t) \in \mathbf{R}^m$ is the measurement of $\boldsymbol{x}(t)$ contaminated by the noise $\boldsymbol{v}(t) \in \mathbf{R}^m$, as depicted in Figure 1, $G(z) \in \mathbf{R}^{m\times r}$ the transfer matrix of the system with z^{-1} representing the unit delay operator z^{-1} $[z^{-1}\boldsymbol{u}(t) = \boldsymbol{u}(t-1)]$.

According to the matrix polynomial theory [14], any strictly proper rational fraction matrix can be decomposed into a matrix fraction description: $G(z) = A^{-1}(z)B(z)$, where $A(z)$ and $B(z)$ are polynomial matrices in z^{-1} and defined as

$$A(z) = I + A_1z^{-1} + A_2z^{-2} + \cdots + A_{n_a}z^{-n_a} \in \mathbf{R}^{m\times m},$$

$$B(z) = B_1z^{-1} + B_2z^{-2} + \cdots + B_{n_b}z^{-n_b} \in \mathbf{R}^{m\times r}.$$

Received December 15, 2009; in revised form February 6, 2009
Supported by the National Natural Science Foundation of P. R. China (60874020)
1. School of Communication and Control Engineering, Jiangnan University, Wuxi, China 214122. fding@jiangnan.edu.cn, fding@sce.carleton.ca. 2. School of Mechatronics Engineering, Nanchang University, Nanchang, China 330031. xpliu@sce.carleton.ca. 3. Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada K1S 5B6. xpliu@sce.carleton.ca.
DOI: 10.1360/aas-007-xxxx

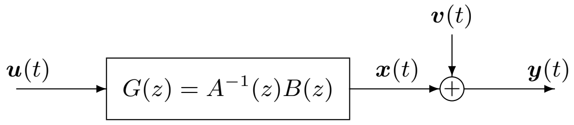

Figure 1 The output-error system

Here, the inner variable $\boldsymbol{x}(t)$ is unknown, $\boldsymbol{y}(t)$ the measurement output vector and $\boldsymbol{v}(t)$ the observation noise vector with zero mean. $\{\boldsymbol{u}(t),\boldsymbol{y}(t)\}$ are the measurement input-output data and $A_i\in\mathbf{R}^{m\times m}$ and $B_i\in\mathbf{R}^{m\times r}$ are the parameter matrices to be identified. Assume that n_a and n_b are known and that $\boldsymbol{u}(t)=\mathbf{0}$, $\boldsymbol{y}(t)=\mathbf{0}$ and $\boldsymbol{v}(t)=\mathbf{0}$ for $t\leqslant 0$. The system in (1)-(2) contains $m^2n_a+mrn_b=:S_1$ parameters. The objective is to present an auxiliary model stochastic gradient algorithm to estimate the unknown parameter matrices A_i and B_i using the input-output data $\{\boldsymbol{u}(t),\boldsymbol{y}(t)\}$.

The model in (1)-(2) may be equivalently written as an MIMO ARMAX model [15, 16]:

$$A(z)\boldsymbol{y}(t)=B(z)\boldsymbol{u}(t)+D(z)\boldsymbol{v}(t),\ D(z)=A(z).$$

This model can be identified by using the extended least squares (ELS) or extended stochastic gradient (ESG) algorithm [16]. Using the ELS algorithm to estimate the parameters of such a special ARMAX model indeed requires identifying m^2n_a more parameters than the actual model parameters. Although the noise model $D(z)$ equals $A(z)$, their estimates are different. In other words, the size of the parameter matrix increases, so this directly leads to a larger computational burden. Therefore, exploring computationally efficient identification approach is the goal of this paper. The following is to derive an auxiliary model identification algorithms with less computation.

2 Basic Algorithms

Let us introduce some notations first. The symbol I (I_m) stands for an identity matrix of appropriate sizes (of $m\times m$); 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 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 square matrix X, respectively; for $g(t)\geqslant 0$, we write $f(t)=O(g(t))$ if there exists positive constants δ_1 and t_0 such that $|f(t)|\leqslant\delta_1 g(t)$ for $t\geqslant t_0$.

Let $n:=mn_a+rn_b$. Define the parameter matrix θ and information vector $\boldsymbol{\varphi}_0(t)$ as

$$\begin{aligned}\theta^{\mathrm{T}}&:=[A_1,A_2,\cdots,A_{n_a},B_1,B_2,\cdots,B_{n_b}]\in\mathbf{R}^{m\times n},\\ \boldsymbol{\varphi}_0(t)&:=[-\boldsymbol{x}^{\mathrm{T}}(t-1),-\boldsymbol{x}^{\mathrm{T}}(t-2),\cdots,-\boldsymbol{x}^{\mathrm{T}}(t-n_a),\\ &\quad\ \boldsymbol{u}^{\mathrm{T}}(t-1),\boldsymbol{u}^{\mathrm{T}}(t-2),\cdots,\boldsymbol{u}^{\mathrm{T}}(t-n_b)]^{\mathrm{T}}\in\mathbf{R}^{n}.\end{aligned}$$

Then from (1) to (2), we have

$$\begin{aligned}\boldsymbol{x}(t)&=\theta^{\mathrm{T}}\boldsymbol{\varphi}_0(t),\\ \boldsymbol{y}(t)&=\theta^{\mathrm{T}}\boldsymbol{\varphi}_0(t)+\boldsymbol{v}(t).\end{aligned}\tag{3}$$

Here, a difficulty arises in that the information vector $\boldsymbol{\varphi}_0(t)$ contains unknown $\boldsymbol{x}(t-i)$ so that the standard stochastic gradient (SG) methods cannot be applied to (3) *directly*. The objective of this work is to establish an auxiliary model by using the available data $\{\boldsymbol{u}(t),\boldsymbol{y}(t)\}$, and to present auxiliary model based SG algorithms by using the output $\boldsymbol{x}_{\mathrm{a}}(t)$ of this auxiliary model in place of the unknown $\boldsymbol{x}(t)$, and to make performance analysis of the algorithms involved.

Let $\hat{\theta}(t)$ be the estimate of θ at time t:

$$\hat{\theta}^{\mathrm{T}}(t)=[\hat{A}_1(t),\cdots,\hat{A}_{n_a}(t),\hat{B}_1(t),\cdots,\hat{B}_{n_b}(t)],$$

and use the entries of the estimate $\hat{\theta}(t)$ to form the polynomials:

$$\begin{aligned}\hat{A}(z)&=I+\hat{A}_1(t)z^{-1}+\hat{A}_2(t)z^{-2}+\cdots+\hat{A}_{n_a}(t)z^{-n_a},\\ \hat{B}(z)&=\hat{B}_1(t)z^{-1}+\hat{B}_2(t)z^{-2}+\cdots+\hat{B}_{n_b}(t)z^{-n_b}.\end{aligned}$$

In terms of $\hat{A}(z)$ and $\hat{B}(z)$, we construct an auxiliary model:

$$\boldsymbol{x}_{\mathrm{a}}(t)=G_{\mathrm{a}}(z)\boldsymbol{u}(t),\ G_{\mathrm{a}}(z):=\hat{A}^{-1}(z)\hat{B}(z),\tag{4}$$

where $G_{\mathrm{a}}(z)$ denotes the estimate of $G(z)$ and is used as the transfer function matrix of the the auxiliary model.

Equation (4) may be also written as a matrix form,

$$\begin{aligned}\boldsymbol{x}_{\mathrm{a}}(t)&=\hat{\theta}^{\mathrm{T}}(t)\boldsymbol{\varphi}(t),\\ \boldsymbol{\varphi}(t)&:=[-\boldsymbol{x}_{\mathrm{a}}^{\mathrm{T}}(t-1),\cdots,-\boldsymbol{x}_{\mathrm{a}}^{\mathrm{T}}(t-n_a),\\ &\quad\ \boldsymbol{u}^{\mathrm{T}}(t-1),\cdots,\boldsymbol{u}^{\mathrm{T}}(t-n_b)]^{\mathrm{T}}.\end{aligned}$$

If we use $\boldsymbol{\varphi}(t)$ to replace $\boldsymbol{\varphi}_0(t)$ in (3), then the identification problem of θ can be solved. Using this idea, we can obtain an auxiliary model based stochastic gradient (AM-SG) algorithm of estimating the parameter matrix θ of the multivariable systems in

$$\hat{\theta}(t)=\hat{\theta}(t-1)+\frac{\boldsymbol{\varphi}(t)}{r(t)}\boldsymbol{e}^{\mathrm{T}}(t),\tag{5}$$

$$\boldsymbol{e}(t)=\boldsymbol{y}(t)-\hat{\theta}^{\mathrm{T}}(t-1)\boldsymbol{\varphi}(t),\tag{6}$$

$$r(t)=r(t-1)+\|\boldsymbol{\varphi}(t)\|^2,\ r(0)=1,\tag{7}$$

$$\begin{aligned}\boldsymbol{\varphi}(t)&=[-\boldsymbol{x}_{\mathrm{a}}^{\mathrm{T}}(t-1),\cdots,-\boldsymbol{x}_{\mathrm{a}}^{\mathrm{T}}(t-n_a),\\ &\quad\ \boldsymbol{u}^{\mathrm{T}}(t-1),\cdots,\boldsymbol{u}^{\mathrm{T}}(t-n_b)]^{\mathrm{T}},\end{aligned}\tag{8}$$

$$\boldsymbol{x}_{\mathrm{a}}(t)=\hat{\theta}^{\mathrm{T}}(t)\boldsymbol{\varphi}(t).\tag{9}$$

The initial value is generally chosen to be a small real matrix, e.g., $\hat{\theta}(0)=10^{-6}\mathbf{1}_{m\times n}$.

3 Main convergence results

We assume that $\{\boldsymbol{v}(t),\mathcal{F}_t\}$ is a martingale difference vector sequence defined on a probability space $\{\Omega,\mathcal{F},P\}$, where $\{\mathcal{F}_t\}$ is the σ algebra sequence generated by the observation data up to and including time t [16]. The sequence $\{\boldsymbol{v}(t)\}$ satisfies:

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

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

Lemma 1. [13] For the algorithm in (5)-(9), the following inequality holds:

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

Theorem 1. For the system in (3) and algorithm in (5)-(9), define the data product moment matrices,

$$Q(t):=\sum_{i=1}^{t}\boldsymbol{\varphi}(i)\boldsymbol{\varphi}^{\mathrm{T}}(i),$$

and assume that (A1) and (A2) hold, $A(z)$ is a strictly positive real matrix, $r(t)\to\infty$. Then the parameter estimation matrix $\hat{\theta}(t)$ consistently converges to θ.

The stochastic martingale theory is one of the main tools of analyzing the convergence of identification algorithms [11, 12]. The following proves this theorem by formulating a martingale process and by using the martingale convergence theorem in [16].

Proof Define the parameter estimation error matrix:

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

Using (5) gives

$$\tilde{\theta}(t) = \tilde{\theta}(t-1) + \frac{\boldsymbol{\varphi}(t)}{r(t)}\boldsymbol{e}^{\mathrm{T}}(t). \tag{10}$$

Let

$$\begin{aligned}\tilde{\boldsymbol{y}}(t) &:= -\tilde{\theta}^{\mathrm{T}}(t)\boldsymbol{\varphi}(t),\\ \boldsymbol{\eta}(t) &:= \boldsymbol{y}(t) - \hat{\theta}^{\mathrm{T}}(t)\boldsymbol{\varphi}(t).\end{aligned} \tag{11}$$

Using (6), it follows that

$$\begin{aligned}\boldsymbol{\eta}(t) &= \frac{r(t-1)}{r(t)}\boldsymbol{e}(t) = \boldsymbol{y}(t) - \boldsymbol{x}_{\mathrm{a}}(t)\\ &= \boldsymbol{x}(t) + \boldsymbol{v}(t) - \boldsymbol{x}_{\mathrm{a}}(t).\end{aligned} \tag{12}$$

Taking the norm of both sides of (10) and using (11) yield

$$\begin{aligned}\|\tilde{\theta}(t)\|^2 &= \|\tilde{\theta}(t-1)\|^2 - \frac{2\tilde{\boldsymbol{y}}^{\mathrm{T}}(t)[\boldsymbol{\eta}(t)-\boldsymbol{v}(t)]}{r(t-1)}\\ &+\frac{2\boldsymbol{\varphi}^{\mathrm{T}}(t)\tilde{\theta}(t-1)\boldsymbol{v}(t)}{r(t-1)} + \frac{2\|\boldsymbol{\varphi}(t)\|^2}{r(t-1)r(t)}[\boldsymbol{e}(t)-\boldsymbol{v}(t)]^{\mathrm{T}}\boldsymbol{v}(t)\\ &+\frac{2\|\boldsymbol{\varphi}(t)\|^2\|\boldsymbol{v}(t)\|^2}{r(t-1)r(t)} - \frac{\|\boldsymbol{\varphi}(t)\|^2}{r^2(t)}\|\boldsymbol{e}(t)\|^2.\end{aligned} \tag{13}$$

From (12), we have

$$\begin{aligned}A(z)[\boldsymbol{\eta}(t)-\boldsymbol{v}(t)] &= A(z)\boldsymbol{x}(t) - A(z)\boldsymbol{x}_{\mathrm{a}}(t)\\ &= B(z)\boldsymbol{u}(t) - A(z)\boldsymbol{x}_{\mathrm{a}}(t) = \theta^{\mathrm{T}}\boldsymbol{\varphi}(t) - \boldsymbol{x}_{\mathrm{a}}(t)\\ &= \theta^{\mathrm{T}}\boldsymbol{\varphi}(t) - \hat{\theta}^{\mathrm{T}}(t)\boldsymbol{\varphi}(t) = -\tilde{\theta}^{\mathrm{T}}(t)\boldsymbol{\varphi}(t) = \tilde{\boldsymbol{y}}(t).\end{aligned} \tag{14}$$

Since $A(z)$ is strictly positive real, referring to Appendix C in [16], the following inequality holds,

$$S(t) := \sum_{i=1}^{t}\frac{2\tilde{\boldsymbol{y}}^{\mathrm{T}}(i)[\boldsymbol{\eta}(i)-\boldsymbol{v}(i)]}{r(t-1)} \geqslant 0,\ \text{a.s.}$$

Let $W(t) := \|\tilde{\theta}(t)\|^2 + S(t)$. Adding both sides of (13) by $S(t)$ gives

$$\begin{aligned}W(t) &= W(t-1) + \frac{2\boldsymbol{\varphi}^{\mathrm{T}}(t)\tilde{\theta}(t-1)\boldsymbol{v}(t)}{r(t-1)}\\ &+\frac{2\|\boldsymbol{\varphi}(t)\|^2}{r(t-1)r(t)}[\boldsymbol{e}(t)-\boldsymbol{v}(t)]^{\mathrm{T}}\boldsymbol{v}(t) + \frac{2\|\boldsymbol{\varphi}(t)\|^2\|\boldsymbol{v}(t)\|^2}{r(t-1)r(t)}\\ &-\frac{\|\boldsymbol{\varphi}(t)\|^2}{r^2(t)}\|\boldsymbol{e}(t)\|^2.\end{aligned}$$

Since $S(t-1)$, $\boldsymbol{\varphi}^{\mathrm{T}}(t)\tilde{\theta}(t-1)$, $r(t-1)$, $\boldsymbol{\varphi}(t)$, $r(t)$ and $\boldsymbol{e}(t)-\boldsymbol{v}(t)$ are uncorrelated with $\boldsymbol{v}(t)$ and $\mathcal{F}_{t-1}$ measurable, taking the conditional expectation of both sides of the above equation with respect to $\mathcal{F}_{t-1}$ and using (A1)-(A2) yield

$$\begin{aligned}\mathrm{E}[W(t)|\mathcal{F}_{t-1}] &= W(t-1) + \frac{\|\boldsymbol{\varphi}(t)\|^2\sigma^2 r^{\epsilon}(t-1)}{r(t-1)r(t)}\\ &-\mathrm{E}\left[\frac{\|\boldsymbol{\varphi}(t)\|^2}{r^2(t)}\|\boldsymbol{e}(t)\|^2|\mathcal{F}_{t-1}\right],\ \text{a.s.}\end{aligned} \tag{15}$$

The summation of the right-hand third term of the above equation from $t=1$ to $t=\infty$ is finite [16], i.e.,

$$\sigma^2\sum_{t=1}^{\infty}\frac{\|\boldsymbol{\varphi}(t)\|^2}{[r(t-1)]^{1-\epsilon}r(t)} < \infty,\ \text{a.s.},\ 1-\epsilon > 0.$$

Applying the martingale convergence theorem (Lemma D.5.3 in [16]) to (15) to get that $W(t)$ a.s. converges to a finite random variable, say, C, i.e.,

$$\lim_{t\to\infty}\|\tilde{\theta}(t)\|^2 + S(t) = C < \infty,\ \text{a.s.}, \tag{16}$$

and also

$$\sum_{t=1}^{\infty}\frac{\|\boldsymbol{\varphi}(t)\|^2}{r^2(t)}\|\boldsymbol{e}(t)\|^2 < \infty,\ \text{a.s.} \tag{17}$$

Hence

$$\sum_{t=1}^{\infty}\frac{\|\tilde{\boldsymbol{y}}(t)\|^2}{r(t-1)} < \infty,\ \text{a.s.},\quad \sum_{t=1}^{\infty}\frac{\|\boldsymbol{\eta}(t)-\boldsymbol{v}(t)\|^2}{r(t-1)} < \infty,\ \text{a.s.} \tag{18}$$

Using the Kronecker lemma (Lemma D.5.5 in [16]), it follows that

$$\begin{aligned}&\lim_{t\to\infty}\frac{1}{r(t-1)}\sum_{i=1}^{t}\|\tilde{\boldsymbol{y}}(i)\|^2 = 0,\ \text{a.s.},\\ &\lim_{t\to\infty}\frac{1}{r(t-1)}\sum_{i=1}^{t}\|\boldsymbol{\eta}(t)-\boldsymbol{v}(t)\|^2 = 0,\ \text{a.s.}\end{aligned}$$

Equation (16) shows that the parameter estimation error is consistent bounded. From (10), we have

$$\tilde{\theta}(t) = \tilde{\theta}(t-i) + \sum_{j=0}^{i-1}\frac{\boldsymbol{\varphi}(t-j)}{r(t-j)}\boldsymbol{e}^{\mathrm{T}}(t-j),\ i\geqslant 1. \tag{19}$$

Thus, we have

$$\begin{aligned}\sum_{t=i}^{\infty}\|\tilde{\theta}(t)-\tilde{\theta}(t-i)\|^2 &= \sum_{t=i}^{\infty}\|\hat{\theta}(t)-\hat{\theta}(t-i)\|^2\\ &= \sum_{t=i}^{\infty}\left\|\sum_{j=0}^{i-1}\frac{\boldsymbol{\varphi}(t-j)}{r(t-j)}\boldsymbol{e}^{\mathrm{T}}(t-j)\right\|^2\\ &\leqslant i\sum_{j=0}^{i-1}\sum_{t=i}^{\infty}\frac{\|\boldsymbol{\varphi}(t-j)\|^2}{r^2(t-j)}\|\boldsymbol{e}(t-j)\|^2 < \infty,\ \text{a.s.},\ i<\infty,\end{aligned}$$

$$\begin{aligned}\sum_{t=1}^{\infty}\frac{\|\boldsymbol{e}(t)-\boldsymbol{v}(t)\|^2}{r(t-1)} &= \sum_{i=1}^{\infty}\frac{\|\boldsymbol{y}(t)-\hat{\theta}^{\mathrm{T}}(t-1)\boldsymbol{\varphi}(t)-\boldsymbol{v}(t)\|^2}{r(t-1)}\\ &= \sum_{i=1}^{\infty}\frac{\|\boldsymbol{y}(t)-\hat{\theta}^{\mathrm{T}}(t)\boldsymbol{\varphi}(t)-\boldsymbol{v}(t)+[\hat{\theta}^{\mathrm{T}}(t)-\hat{\theta}^{\mathrm{T}}(t-1)]\boldsymbol{\varphi}(t)\|^2}{r(t-1)}\\ &= \sum_{i=1}^{\infty}\frac{\|\boldsymbol{\eta}(t)-\boldsymbol{v}(t)+[\tilde{\theta}^{\mathrm{T}}(t)-\tilde{\theta}^{\mathrm{T}}(t-1)]\boldsymbol{\varphi}(t)\|^2}{r(t-1)}\end{aligned}$$

$$\leqslant \sum_{i=1}^{\infty}\frac{2\|\boldsymbol{\eta}(t)-\boldsymbol{v}(t)\|^2}{r(t-1)}+\sum_{i=1}^{\infty}\frac{2\|[\tilde{\theta}^{\rm T}(t)-\tilde{\theta}^{\rm T}(t-1)]\boldsymbol{\varphi}(t)\|^2}{r(t-1)}$$
$$\leqslant \sum_{i=1}^{\infty}\frac{2\|\boldsymbol{\eta}(t)-\boldsymbol{v}(t)\|^2}{r(t-1)}+2\sum_{i=1}^{\infty}\frac{\|\boldsymbol{\varphi}(t)\|^2}{r(t-1)}\|\tilde{\theta}(t)-\tilde{\theta}(t-1)\|^2$$
$$=\sum_{i=1}^{\infty}\frac{2\|\boldsymbol{\eta}(t)-\boldsymbol{v}(t)\|^2}{r(t-1)}+2C_1\sum_{i=1}^{\infty}\|\tilde{\theta}(t)-\tilde{\theta}(t-1)\|^2$$
$$=: C_2<\infty,\ \text{a.s.},\ C_1<\infty.$$

Here, we have assumed that $\|\boldsymbol{\varphi}(t)\|^2 \leqslant C_1 r(t-1)$. Using the Kronecker lemma gives

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

From (19), we have

$$\tilde{\theta}(t-i)=\tilde{\theta}(t)-\sum_{j=0}^{i-1}\frac{\boldsymbol{\varphi}(t-j)}{r(t-j)}\boldsymbol{e}^{\rm T}(t-j). \tag{20}$$

Replacing t in (11) by $t-i$ yields

$$\boldsymbol{\varphi}^{\rm T}(t-i)\tilde{\theta}(t-i)=-\tilde{\boldsymbol{y}}^{\rm T}(t-i).$$

Using (20), we have

$$\boldsymbol{\varphi}^{\rm T}(t-i)\tilde{\theta}(t)=-\tilde{\boldsymbol{y}}^{\rm T}(t-i)+\boldsymbol{\varphi}^{\rm T}(t-i)\sum_{j=0}^{i-1}\frac{\boldsymbol{\varphi}(t-j)}{r(t-j)}\boldsymbol{e}^{\rm T}(t-j).$$

To some extent, the rest of the proof is similar to that of reference [13]. Squaring and using the relation, $(a+b)^2 \leqslant 2(a^2+b^2)$, yield

$$\|\boldsymbol{\varphi}^{\rm T}(t-i)\tilde{\theta}(t)\|^2 \leqslant 2\|\tilde{\boldsymbol{y}}(t-i)\|^2+2\|\boldsymbol{\varphi}(t-i)\|^2$$
$$\times\left\|\sum_{j=0}^{i-1}\frac{\boldsymbol{\varphi}(t-j)}{r(t-j)}\{[\boldsymbol{e}(t-j)-\boldsymbol{v}(t-j)]+\boldsymbol{v}(t-j)\}^{\rm T}\right\|^2.$$

Since $\boldsymbol{e}(t-j)-\boldsymbol{v}(t-j)$ is uncorrelated with $\boldsymbol{v}(t-j)$ and is $\mathcal{F}_{t-1}$ measurable, taking the conditional expectation on both sides with respect to $\mathcal{F}_{t-1}$ and using (A1)-(A2) give

$$\mathrm{E}[\|\boldsymbol{\varphi}^{\rm T}(t-i)\tilde{\theta}(t)\|^2|\mathcal{F}_{t-1}] \leqslant 2\|\tilde{\boldsymbol{y}}(t-i)\|^2$$
$$+2\|\boldsymbol{\varphi}(t-i)\|^2\sum_{j=0}^{i-1}\frac{\|\boldsymbol{\varphi}(t-j)\|^2}{r^2(t-j)}$$
$$\times\{\|\boldsymbol{e}(t-j)-\boldsymbol{v}(t-j)\|^2+\sigma^2 r^{\epsilon}(t-1)\}.$$

Summing for i from $i=0$ to $i=t-1$ of both sides and dividing by $r(t)$ yield

$$\frac{\mathrm{E}\{\mathrm{tr}[\tilde{\theta}^{\rm T}(t)Q(t)\tilde{\theta}(t)]|\mathcal{F}_{t-1}\}}{r(t)}=\frac{2}{r(t)}\sum_{i=1}^{t}\|\tilde{\boldsymbol{y}}(i)\|^2+S_1(t)+S_2(t),$$

where

$$S_1(t):=2\sum_{i=1}^{t-1}\frac{\|\boldsymbol{\varphi}(t-i)\|^2}{r(t)}\sum_{j=0}^{i-1}\frac{\|\boldsymbol{\varphi}(t-j)\|^2}{r^2(t-j)}\sigma^2 r^{\epsilon}(t-1),$$

$$S_2(t):=2\sum_{i=1}^{t-1}\frac{\|\boldsymbol{\varphi}(t-i)\|^2}{r(t)}\sum_{j=0}^{i-1}\frac{\|\boldsymbol{\varphi}(t-j)\|^2}{r^2(t-j)}$$
$$\times\|\boldsymbol{e}(t-j)-\boldsymbol{v}(t-j)\|^2.$$

Using Lemma 1, we have

$$S_1(t)=\frac{2}{r(t)}\sum_{i=2}^{t}\frac{[r(i-1)-r(0)]\|\boldsymbol{\varphi}(i)\|^2}{r^2(i)}\sigma^2 r^{\epsilon}(t-1)$$
$$\leqslant\frac{2}{[r(t)]^{1-\epsilon}}\sum_{i=2}^{t}\frac{\|\boldsymbol{\varphi}(i)\|^2}{r(i)}\sigma^2 \leqslant \frac{2\sigma^2\ln r(t)}{[r(t)]^{1-\epsilon}}\to 0,\ \text{a.s.},$$
$$S_2(t)=\frac{2}{r(t)}\sum_{i=2}^{t-1}\frac{[r(i-1)-r(0)]\|\boldsymbol{\varphi}(i)\|^2}{r^2(i)}\|\boldsymbol{e}(i)-\boldsymbol{v}(i)\|^2$$
$$\leqslant\frac{2}{r(t)}\sum_{i=2}^{t}\frac{\|\boldsymbol{\varphi}(i)\|^2}{r(i)}\|\boldsymbol{e}(i)-\boldsymbol{v}(i)\|^2$$
$$\leqslant\frac{2}{r(t)}\sum_{i=2}^{t}\|\boldsymbol{e}(i)-\boldsymbol{v}(i)\|^2\to 0,\ \text{a.s., as } t\to\infty.$$

Hence

$$\|\tilde{\theta}(t)\|^2=o\left(\frac{r(t)}{\lambda_{\min}[Q(t)]}\right),\ \text{a.s.}$$

This proves Theorem 1. □

The auxiliary model stochastic gradient (AM-SG) algorithm has low computational burden, but its convergence is slow, just like the stochastic gradient algorithm of scalar systems in [16]. In order to improve the convergence rate and tracking performance, introducing a forgetting factor λ in the AM-SG algorithm obtains the AM-SG algorithm with forgetting factor (the AM-FFSG algorithm for short) as follows:

$$\hat{\theta}(t)=\hat{\theta}(t-1)+\frac{\boldsymbol{\varphi}(t)}{r(t)}[\boldsymbol{y}^{\rm T}(t)-\boldsymbol{\varphi}^{\rm T}(t)\hat{\theta}(t-1)], \tag{21}$$

$$r(t)=\lambda r(t-1)+\|\boldsymbol{\varphi}(t)\|^2,\ 0<\lambda<1,\ r(0)=1, \tag{22}$$

$$\boldsymbol{\varphi}(t)=[-\boldsymbol{\varphi}^{\rm T}(t-1)\hat{\theta}(t-1),\cdots,-\boldsymbol{\varphi}^{\rm T}(t-n_a)\hat{\theta}(t-n_a),$$
$$\boldsymbol{u}^{\rm T}(t-1),\cdots,\boldsymbol{u}^{\rm T}(t-n_b)]^{\rm T}. \tag{23}$$

When $\lambda=1$, the AM-FFSG algorithm reduces to the AM-SG algorithm; when $\lambda=0$, the AM-FFSG algorithm is the auxiliary model projection algorithm.

For comparison, the following gives the auxiliary model based recursive least squares (AM-RLS) algorithm of estimating θ:

$$\hat{\theta}(t)=\hat{\theta}(t-1)+P(t)\boldsymbol{\varphi}(t)[\boldsymbol{y}^{\rm T}(t)-\boldsymbol{\varphi}^{\rm T}(t)\hat{\theta}(t-1)], \tag{24}$$

$$P(t)=P(t-1)-\frac{P(t-1)\boldsymbol{\varphi}(t)\boldsymbol{\varphi}^{\rm T}(t)P(t-1)}{1+\boldsymbol{\varphi}^{\rm T}(t)P(t-1)\boldsymbol{\varphi}(t)}. \tag{25}$$

4 Simulation tests

Consider the following 2-input and 2-output system (the output error system):

$$\begin{bmatrix}x_1(t)\\x_2(t)\end{bmatrix}+\begin{bmatrix}-0.50 & 0.30\\0.30 & -0.70\end{bmatrix}\begin{bmatrix}x_1(t-1)\\x_2(t-1)\end{bmatrix}$$
$$=\begin{bmatrix}2.00 & 0.80\\0.60 & 1.50\end{bmatrix}\begin{bmatrix}u_1(t-1)\\u_2(t-1)\end{bmatrix},$$
$$\begin{bmatrix}y_1(t)\\y_2(t)\end{bmatrix}=\begin{bmatrix}x_1(t)\\x_2(t)\end{bmatrix}+\begin{bmatrix}v_1(t)\\v_2(t)\end{bmatrix}.$$

In simulation, the inputs $\{u_1(t)\}$ and $\{u_2(t)\}$ 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 σ_1^2 and σ_2^2.

Applying the AM-SG, AM-FFSG and AM-RLS algorithms to estimate the parameters of this system, the parameter estimates are shown in Tables 1 to 3 and the estimation errors δ versus t are shown in Figures 2 and 3, where $\delta := \|\hat{\theta}(t)-\theta\|/\|\theta\|\times 100\%$ is the the parameter estimation error.

Changing the noise variances σ_1^2 and σ_2^2 can adjust the noise-to-signal ratios $\delta_{\rm ns}(1)$ and $\delta_{\rm ns}(2)$ of two output channels. When $\sigma_1^2 = 0.20^2$ and $\sigma_2^2 = 0.20^2$, the noise-to-signal ratios are $\delta_{\rm ns}(1) = 7.61\%$ and $\delta_{\rm ns}(2) = 7.99\%$; when $\sigma_1^2 = 1.00^2$ and $\sigma_2^2 = 1.00^2$, the noise-to-signal ratios are $\delta_{\rm ns}(1) = 38.06\%$ and $\delta_{\rm ns}(2) = 39.96\%$.

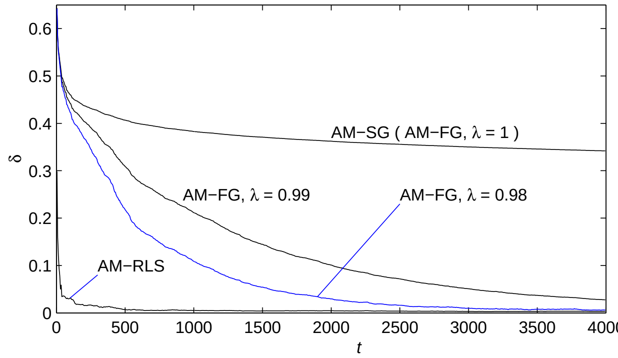

Figure 2 The estimation errors δ vs. t with different forgetting factors ($\sigma_1^2 = 0.20^2$ and $\sigma_2^2 = 0.20^2$)

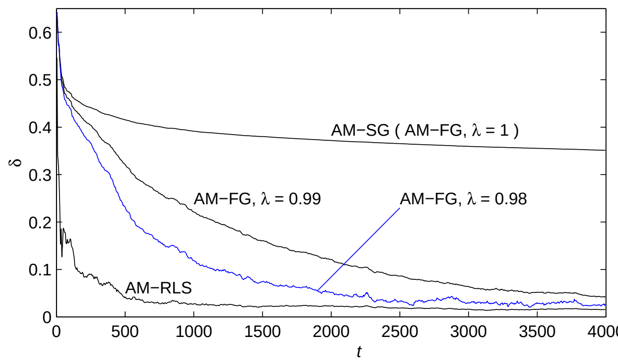

Figure 3 The estimation errors δ vs. t with different forgetting factors ($\sigma_1^2 = 1.00^2$ and $\sigma_2^2 = 1.00^2$)

From these simulation results in Tables 1 to 3 and Figures 2 and 3, we can draw the conclusions: 1. A lower noise level leads to a faster rate of convergence of the parameter estimates to the true parameters. 2. As long as an appropriate forgetting factor is chosen, the faster convergence rate can be achieved and the smaller estimation errors may be obtained. 3. The estimation errors δ are becoming smaller (in general) as the data length t increases. In other words, increasing data length generally results in smaller parameter estimation errors. 4. If we choose an appropriate forgetting factor, the parameter estimation error of the AM-FFSG algorithm is very close to that of the AM-RLS algorithm. 5. These show the effectiveness of the proposed theorem.

5 Conclusions

Using the auxiliary model technique, the auxiliary model based stochastic gradient algorithms are presented for MIMO systems. The convergence of the proposed algorithm is analyzed by using the martingale convergence theorem. The simulation results show that the proposed algorithms are effective.

References

1 Zheng WX. A bias correction method for identification of linear dynamic errors-in-variables models. *IEEE Transactions on Automatic Control*, 2002, **47**(7): 1142–1147.

2 Yang HZ, Zhang Y. Comparisons of bias compensation methods and other identification approaches for Box-Jenkins models. *Control Theory and Applications*, 2007, **24**(2): 215–222.

3 Zhang Y, Yang HZ. Bias compensation recursive least squares identification for output error systems with colored noises. *Acta Automatica Sinica*, 2007, **33**(10): 1053–1060.

4 Zong Q, Du LQ, Liu WJ. Iterative identification and control design based on vinnicombe distance. *Acta Automatica Sinica*, 2008, **34**(11): 1431–1436.

5 Zhong LS, Song ZH. Hierarchical optimization identification of LTI state-space systems by projected gradient search. *Acta Automatica Sinica*, 2008, **34**(6): 711–715.

6 Wang DQ. Recursive extended least squares identification method based on auxiliary models. *Control Theory and Applications*, 2009, **26**(1): 51–56.

7 Ding F, Chen T. Hierarchical gradient-based identification of multivariable discrete-time systems. *Automatica*, 2005, **41**(2): 315–325.

8 Ding F, Chen T. Performance analysis of multi-innovation gradient type identification methods. *Automatica*, 2007, **43**(1): 1–14.

9 Wang DQ, Ding F. Auxiliary models based multi-innovation generalized extended stochastic gradient algorithms. *Control and Decision*, 2008, **23**(9): 999–1003+1010.

10 Wang DQ, Ding F. Extended stochastic gradient identification algorithms for Hammerstein-Wiener ARMAX Systems. *Computers & Mathematics with Applications*, 2008, **56**(12): 3157–3164.

11 Ding F, Chen T. Combined parameter and output estimation of dual-rate systems using an auxiliary model. *Automatica*, 2004, **40**(10): 1739–1748.

12 Ding F, Chen T. Hierarchical least squares identification methods for multivariable systems. *IEEE Transactions on Automatic Control*, 2005, **50**(3): 397–402.

13 Ding F, Yang HZ, Liu F. Performance analysis of stochastic gradient algorithms under weak conditions. *Science in China Series F–Information Sciences*, 2008, **51**(9): 1269–1280.

14 Kailath T. Linear systems. Englewood Cliffs, NJ: Prentice-Hall, 1980.

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

16 Goodwin GC, Sin KS. Adaptive Filtering, Prediction and Control. Englewood Cliffs, New Jersey: Prentice-Hall, 1984.

DING Feng was born in Guangshui, Hubei Province. He received the B. Sc. degree from the Hubei University of Technology (Wuhan, 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, Xiangfan, China. From 1994 to 2002, he was with the Department of Automation at the Tsinghua University, Beijing, China and he was a Research Associate at the University of Alberta, Edmonton, Canada from 2002 to 2005.

He is now a Full Professor in the Control Science and Engineering Research Center at the Jiangnan University, Wuxi, China. His current research interests include model identification and adaptive control. He co-authored the book *Adaptive*

Table 1 The AM-SG estimates and errors ($\sigma_1^2 = 0.20^2$ and $\sigma_2^2 = 0.20^2$)

t	a_{11}	a_{12}	b_{11}	b_{12}	a_{21}	a_{22}	b_{21}	b_{22}	δ (%)
100	-0.59922	0.17140	1.48708	-0.30071	0.33764	-0.71091	0.31219	1.11802	46.03369
200	-0.60059	0.17094	1.51063	-0.24293	0.31021	-0.69470	0.32900	1.13286	43.72964
500	-0.59547	0.16413	1.54290	-0.16607	0.30202	-0.68306	0.34396	1.16092	40.65912
1000	-0.59658	0.16547	1.56834	-0.10671	0.29320	-0.68347	0.35884	1.18001	38.27283
2000	-0.59529	0.16723	1.59320	-0.05864	0.29220	-0.69037	0.37232	1.19823	36.23175
3000	-0.59712	0.16900	1.60756	-0.02985	0.29159	-0.69500	0.38084	1.20899	35.02159
4000	-0.59702	0.17187	1.61694	-0.00982	0.29265	-0.69681	0.38608	1.21616	34.18509
True values	-0.50000	0.30000	2.00000	0.80000	0.30000	-0.70000	0.60000	1.50000	

Table 2 The AM-FFSG estimates and errors with $\lambda = 0.99$ and $\lambda = 0.98$ ($\sigma_1^2 = 0.20^2$ and $\sigma_2^2 = 0.20^2$)

λ	t	a_{11}	a_{12}	b_{11}	b_{12}	a_{21}	a_{22}	b_{21}	b_{22}	δ (%)
0.99	100	-0.60777	0.17692	1.51146	-0.25885	0.32854	-0.70781	0.31572	1.14094	44.21498
	200	-0.61377	0.18746	1.55035	-0.16457	0.29378	-0.70023	0.34521	1.16656	40.41276
	500	-0.57355	0.18923	1.65060	0.06882	0.30173	-0.68181	0.39410	1.24891	30.90156
	1000	-0.55986	0.22650	1.75541	0.30436	0.28222	-0.69776	0.45650	1.32641	21.15147
	2000	-0.52576	0.26546	1.88925	0.55857	0.29103	-0.70166	0.53029	1.42339	10.09261
	3000	-0.51682	0.28012	1.94452	0.67591	0.29855	-0.69931	0.57385	1.46477	5.07976
	4000	-0.50970	0.29168	1.97003	0.73266	0.30409	-0.69537	0.58706	1.48132	2.74364
0.98	100	-0.61258	0.18144	1.53585	-0.21435	0.31580	-0.69947	0.32049	1.16407	42.30653
	200	-0.61699	0.20207	1.59079	-0.07974	0.27566	-0.70259	0.36541	1.20042	36.86039
	500	-0.54834	0.23167	1.75031	0.28612	0.31004	-0.68227	0.44480	1.32795	21.77915
	1000	-0.53044	0.26963	1.87190	0.54686	0.28936	-0.70362	0.51873	1.41278	10.87786
	2000	-0.50726	0.28972	1.96794	0.72901	0.29620	-0.69964	0.57536	1.48445	2.94576
	3000	-0.50522	0.29662	1.98799	0.77670	0.30052	-0.69631	0.60452	1.50000	0.96454
	4000	-0.50621	0.29923	1.99316	0.78825	0.30504	-0.69347	0.60168	1.49924	0.60126
True values		-0.50000	0.30000	2.00000	0.80000	0.30000	-0.70000	0.60000	1.50000	

Table 3 The AM-RLS estimates and errors ($\sigma_1^2 = 0.20^2$ and $\sigma_2^2 = 0.20^2$)

t	a_{11}	a_{12}	b_{11}	b_{12}	a_{21}	a_{22}	b_{21}	b_{22}	δ (%)
100	-0.50892	0.28087	2.02473	0.73734	0.29262	-0.69484	0.54565	1.48989	3.15273
200	-0.50403	0.30380	2.00898	0.77285	0.28644	-0.70340	0.57779	1.47845	1.56517
500	-0.50247	0.30326	2.00902	0.78733	0.29335	-0.69865	0.59075	1.49674	0.70049
1000	-0.49955	0.30134	2.00597	0.79384	0.29455	-0.70053	0.59248	1.49277	0.51227
2000	-0.50093	0.29993	1.99891	0.78932	0.29720	-0.70082	0.59432	1.49749	0.44703
3000	-0.50041	0.29945	1.99776	0.79155	0.29832	-0.70002	0.60076	1.49984	0.31355
4000	-0.50115	0.29986	1.99770	0.79179	0.30042	-0.70001	0.60161	1.49924	0.30776
True values	-0.50000	0.30000	2.00000	0.80000	0.30000	-0.70000	0.60000	1.50000	

Control Systems (Tsinghua University Press, Beijing, 2002), and published over 100 papers on modeling and identification as the first author.

http://www.cc.jiangnan.edu.cn/~fding/index.html
E-mails: fding@jiangnan.edu.cn, fding@sce.carleton.ca

Peter X. Liu received his B.Sc. and M.Sc. degrees from Northern Jiaotong University, China in 1992 and 1995, respectively, and Ph.D. degree from the University of Alberta, Canada in 2002. He has been with the Department of Systems and Computer Engineering, Carleton University, Canada since July 2002 and he is currently a Canada Research Chair Professor. His interest includes interactive networked systems and teleoperation, haptics, robotics, intelligent systems, context-aware intelligent networks, and their applications to biomedical engineering.

Dr. Liu has published more than 100 research articles. He serves as an Associate Editor for several journals including IEEE/ASME Transactions on Mechatronics, Intelligent Service Robotics, Control and Intelligent Systems and Int. J. of Advanced Media and Communication. He received 2007 Carleton Research Achievement Award, 2006 Province of Ontario Early Researcher Award, 2006 Carty Research Fellowship, the Best Conference Paper Award of the 2006 IEEE International Conference on Mechatronics and Automation, and 2003 Province of Ontario Distinguished Researcher Award. He has served in the Organization Committees of numerous conferences including being the General Chair of the 2008 IEEE International Workshop on Haptic Audio Visual Environments and their Applications, and the General Chair of 2005 IEEE International Conference on Mechatronics and Automation.

Dr. Liu is a member of the Professional Engineers of Ontario (P.Eng) and a senior member of IEEE.

E-mail: xpliu@sce.carleton.ca