2005 American Control Conference
June 8-10, 2005. Portland, OR, USA

FrB15.2

On Consistency of Stochastic Gradient Algorithms for ARMAX Models with Disturbances

Feng Ding and Tongwen Chen

Abstract— **In this paper, we give a new method to prove in detail the consistency of the residual based and innovation based stochastic gradient algorithms for identifying CARMA/ARMAX models with disturbances under weaker conditions on statistical properties of the noise, e.g., the mean value is non-zero, and the variance is time-varying, and/or high-order moments are possibly nonexistent. The analysis indicates that the parameter estimation error is consistently bounded, and consistently converges to zero under persistent excitation conditions.**

Keywords: Identification, parameter estimation, convergence properties, stochastic gradient, stochastic approximation.

I. INTRODUCTION

For decades, a great deal of work has been published on convergence analysis of identification methods such as least squares (LS) and stochastic gradient (SG) algorithms. However, further research is still required for the following reason:

> Most convergence results of LS and SG (based adaptive control) algorithms make strict assumptions. For example, references [1]–[9] all assume that the process noise is stationary with zero-mean, constant variance, and its high-order moments exist, which is usually not the case in practice.

Therefore, exploring the properties of the SG algorithm under weaker conditions is still open and also the goal in this paper. We will frame our study in the CARMA (or ARMAX) models with a time-delay and a deterministic disturbance, and assume that the process noise has non-zero mean and bounded time-varying variance, and possibly no high-order moments. Under such weaker assumptions, the validity of all existing convergence results on parameter estimation and adaptive control is questionable.

The SG algorithms are a class of important stochastic approximation methods, which have received much attention and have been widely applied in many areas [10]–[14], including adaptive control and optimization [7]–[9], [15]. Under the so-called weaker assumptions, Chen and Caines studied the properties of the SG algorithm by using the ordinary differential equation method [15]. The idea was to transform the discrete recursive equation of the parameter estimation error into a continuous-time differential equation. But the main result (Theorem 1 in [15]) on the parameter estimation convergence of the SG algorithm used conflicting assumptions, see the next section.

This paper is organized as follows. In Section II, we discuss the problem formulation and the conflicting assumptions in [15]. In Sections III and IV, we study the performance of the residual based SG algorithm and the innovation based SG algorithms, respectively. In Section V, we present an illustrative example for the results in this paper. Finally, concluding remarks are given in Section VI.

II. THE PROBLEM FORMULATION

Consider the discrete-time system described by a CARMA model with a disturbance,

$$A(z)y(t) = z^{-d}B(z)u(t) + D(z)v'(t) + f',$$

where $\{y(t)\}$ and $\{u(t)\}$ are the system input and output sequences, $\{v'(t)\}$ is a random noise sequence with mean value $\mathrm{E}[v'(t)] = \mu$ and bounded time-varying variance $\sigma_v^2(t)$, d the time delay, f' the deterministic disturbance, z^{-1} represents the unit backward shift operator: $z^{-1}y(t) = y(t-1)$, and $A(z)$, $B(z)$ and $D(z)$ are polynomials in z^{-1} with

$$\begin{aligned}
A(z) &= 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_{n_a} z^{-n_a},\\
B(z) &= b_0 + b_1 z^{-1} + b_2 z^{-2} + \cdots + b_{n_b} z^{-n_b},\\
D(z) &= 1 + d_1 z^{-1} + d_2 z^{-2} + \cdots + d_{n_d} z^{-n_d}.
\end{aligned}$$

Let $v(t) = v'(t) - \mu$, $f = \mu(1 + d_1 + d_2 + \cdots + d_{n_d}) + f'$, then $v(t)$ have zero mean value and variance $\sigma_v^2(t)$. Hence

$$A(z)y(t) = z^{-d}B(z)u(t) + D(z)v(t) + f. \qquad (1)$$

Let $n_0 := n_a + n_b + 2 + n_d$ and define the extended parameter vector θ and the information vector $\varphi_0(t)$ as

$$\begin{aligned}
\theta &= [a_1, \cdots, a_{n_a}, b_0, \cdots, b_{n_b}, d_1, \cdots, d_{n_d}, f]^{\mathrm{T}},\\
\varphi_0(t) &= [-y(t-1), -y(t-2), \cdots, -y(t-n_a),\\
&\quad u(t-d), u(t-d-1), \cdots, u(t-d-n_b),\\
&\quad v(t-1), \cdots, v(t-n_d), 1]^{\mathrm{T}} \in \mathbb{R}^{n_0}.
\end{aligned}$$

Then Equation (1) can be rewritten as

$$y(t) = \varphi_0^{\mathrm{T}}(t)\theta + v(t). \qquad (2)$$

The objective of this paper is, by means of simple stochastic gradient algorithms, to estimate the parameter vector θ by

This research was supported by the Natural Sciences and Engineering Research Council of Canada and the National Natural Science Foundation of China (60474039).

F. Ding is with the Control Science and Engineering Research Center at the Southern Yangtze University, Wuxi 214122, China (`fding@sytu.edu.cn`), and is currently a Research Associate at the University of Alberta, Edmonton, Canada. `fding@ece.ualberta.ca`

T. Chen is with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada T6G 2V4. `tchen@ece.ualberta.ca`

0-7803-9098-9/05/$25.00 ©2005 AACC

utilizing the observations $\{u(i), y(i):\ i \le t\}$, and to study the properties of the SG algorithms.

Notice that the noise terms $v(t-i),\ i=1,2,\cdots,n_d$ in the information vector $\varphi_0(t)$ are unmeasurable. Thus, we replace them by the estimated residuals $\hat{v}(t-i)$ or the innovations $e(t-i)$, and $\varphi_0(t)$ by $\varphi(t)$, then obtain the residual based SG algorithm and the innovation based SG algorithm, respectively.

Let us introduce some notation first. The symbol I stands for an identity of appropriate dimensions; the norm of a column vector x is defined as $\|x\|^2 = x^{\mathrm{T}}x$; for $g(t) \ge 0$, we write $f(t) = O(g(t))$ or $f(t) \sim g(t)$ if there exists a positive constant δ_1 such that $|f(t)| \le \delta_1 g(t)$.

We assume that $\{v(t), \mathscr{F}_t\}$ is a martingale difference sequence defined on a probability space $\{\Omega, \mathscr{F}, P\}$, where $\{\mathscr{F}_t\}$ is the σ algebra sequence generated by $\{v(t)\}$, i.e., $\mathscr{F}_t = \sigma(v(t), v(t-1), v(t-2), \cdots,)$ or $\mathscr{F}_t = \sigma(y(t), y(t-1), y(t-2), \cdots,)$ for the deterministic input sequence $\{u(t)\}$. The sequence $\{v(t)\}$ satisfies [16]

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

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

$$(A3) \qquad \limsup_{t\to\infty} \frac{1}{t}\sum_{i=1}^{t} v^2(i) \le \sigma_v^2 < \infty, \text{ a.s.}$$

That is, $\{v(t)\}$ is a random noise with zero mean and bounded time-varying variances. Thus, the system in (2) may be non-stationary.

Reference [15] assumed that

$$(C1) \qquad \mathrm{E}[v^2(t)|\mathscr{F}_{t-1}] \le \delta_\varepsilon r^\varepsilon(t), \text{ a.s., } 0 \le \varepsilon < 1;$$

and there exist an integer sequence $\{t_0, t_1, t_2, \cdots, t_s, \cdots\}$, $t_0 = 0$, $t_s^* := t_{s+1} - t_s \ge \dim\theta$ and a positive constant c not depending on t such that for all $s = 1, 2, 3, \cdots$, the following persistent excitation (PE) condition holds:

$$(C2) \qquad \sum_{t=t_{s-1}}^{t_s-1} \frac{\varphi(t)\varphi^{\mathrm{T}}(t)}{r(t)} \ge cI, \text{ a.s.}$$

Here, δ_ε is some positive constant and $r(t)$ and $\varphi(t)$ are defined later.

Condition $0 \le \varepsilon < 1$ in (C1) implies that (C1) holds for all $\varepsilon \in (0,1)$ not for some $\varepsilon \in [0,1)$. We will show that (C2) cannot hold for $\varepsilon \in (0,1)$, i.e., Assumption (C2) holds only for $\varepsilon = 0$. Consider the example:

$$y(t) = ay(t-1) + bu(t) + v(t),\ \mathrm{E}[v^2(t)|\mathscr{F}_{t-1}] = t^{1/2}, \text{ a.s.}$$

We assume that the input $u(t)$ is taken as a pseudo-random binary sequence, i.e., $u(t) = \pm 1$. Then

$$y(t) = O(t^{1/4}),\ \varphi(t) = \begin{bmatrix} y(t-1) \\ u(t) \end{bmatrix} = \begin{bmatrix} O(t^{1/4}) \\ \pm 1 \end{bmatrix},$$
$$r(t) = \sum_{i=1}^{t} \|\varphi(i)\|^2 = \sum_{i=1}^{t} [y^2(i-1) + u^2(i)]$$
$$= O((t-1)^{3/2}) + t \ \sim\ t^{3/2} + t > t.$$

Thus, take $\delta_\varepsilon = 1$, Assumption (C1) holds. However,

$$\sum_{t=t_{s-1}}^{t_s-1} \frac{\varphi(t)\varphi^{\mathrm{T}}(t)}{r(t)}$$
$$\sim \sum_{t=t_{s-1}}^{t_s-1} \frac{1}{t^{3/2}} \begin{bmatrix} y^2(t-1) & y(t-1)u(t) \\ u(t)y(t-1) & u^2(t) \end{bmatrix}$$
$$\sim \sum_{t=t_{s-1}}^{t_s-1} \frac{1}{t^{3/2}} \begin{bmatrix} y^2(t-1) & \pm y(t-1) \\ \pm y(t-1) & 1 \end{bmatrix}$$
$$= \begin{bmatrix} \sum\limits_{t=t_{s-1}}^{t_s-1} \frac{y^2(t-1)}{t^{3/2}} & \pm \sum\limits_{t=t_{s-1}}^{t_s-1} \frac{y(t-1)}{t^{3/2}} \\ \pm \sum\limits_{t=t_{s-1}}^{t_s-1} \frac{y(t-1)}{t^{3/2}} & \sum\limits_{t=t_{s-1}}^{t_s-1} \frac{1}{t^{3/2}} \end{bmatrix}.$$

Since the series $\sum_{t=1}^{\infty} \frac{1}{t^{3/2}}$ is convergent, there does not exist any sequence $\{t_s\}$ even if $t_s^* \to \infty$ such that the element (2, 2) on the right-hand side of the above expression is greater than any small positive number c for any s. That is, Assumption (C2) cannot hold for $\varepsilon > 0$. So $\varepsilon = 0$. The PE condition in (C2) may also be modified as follows: There exist a positive constant c and an integer N such that the following inequality holds:

$$(A4) \qquad \frac{1}{N}\sum_{i=0}^{N-1} \varphi(t-i)\varphi^{\mathrm{T}}(t-i) \ge cI, \text{ a.s., for all } t.$$

This is referred to as the strong PE condition.

III. The residual based SG algorithm

Let $\hat{\theta}(t)$ denote the estimate of θ. The residual based SG algorithm of estimating θ is as follows:

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \frac{\varphi(t)}{r(t)}[y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t-1)], \quad (3)$$

$$r(t) = r(t-1) + \varphi^{\mathrm{T}}(t)\varphi(t), \quad r(0) = 1, \quad (4)$$

$$\varphi(t) = [-y(t-1), -y(t-2), \cdots, -y(t-n_a), u(t-d), u(t-d-1), \cdots, u(t-d-n_b), \hat{v}(t-1), \cdots, \hat{v}(t-n_d), 1]^{\mathrm{T}}, \quad (5)$$

$$\hat{v}(t) = y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t). \quad (6)$$

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

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

$$\tilde{\theta}(t) = \hat{\theta}(t) - \theta, \quad (7)$$
$$e(t) = y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t-1).$$

It follows that

$$\hat{v}(t) = \frac{r(t-1)}{r(t)} e(t).$$

The following lemma is required to establish the main convergence results.

Lemma 1: For $g(t) \ge 0$ and $h(t) \ge 0$, assume that $\lim_{t\to\infty} g(t) = g$, $\sum_{t=1}^{\infty} h(t) = \infty$, and $\sum_{t=1}^{\infty} g(t)h(t) < \infty$. Then $g = 0$.

The proof of Lemma 1 is straightforward and hence omitted.

Theorem 1: For the system in (2) and the SG algorithm in (3)-(6), assume that Conditions (A1) to (A4) hold, $\sum_{t=1}^{\infty} r^{-1}(t) = \infty$, a.s., and $D(z)$ is a strictly positive real function. Then the parameter estimation vector $\hat{\theta}(t)$ in (3) consistently converges to the true parameter vector θ.

Proof We use the martingale convergence theorem to prove this theorem. Let

$$\Delta\tilde{\theta}(t) = \frac{\varphi(t)}{r(t)} e(t), \ \tilde{y}(t) = -\varphi^{\mathrm{T}}(t)\tilde{\theta}(t). \tag{8}$$

Substituting (8) into (3) and using (7), we easily obtain

$$\tilde{\theta}(t) = \tilde{\theta}(t-1) + \frac{\varphi(t)}{r(t)} e(t) = \tilde{\theta}(t-1) + \Delta\tilde{\theta}(t). \tag{9}$$

Taking the norm and using (9) and (8) give

$$\begin{aligned} \|\tilde{\theta}(t)\|^2 &= \|\tilde{\theta}(t-1)\|^2 - \frac{2\tilde{y}(t)[\hat{v}(t)-v(t)]}{r(t-1)} \\ &\quad -\frac{2\tilde{y}(t)v(t)}{r(t-1)} - \frac{\|\varphi(t)\|^2}{r^2(t)} e^2(t). \end{aligned} \tag{10}$$

From the definitions of $\hat{v}(t)$, we have

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

It follows that

$$\begin{aligned} \tilde{y}(t) &= [D(z)-\rho][\hat{v}(t)-v(t)] + \rho[\hat{v}(t)-v(t)] \\ &= y_1(t) + \rho[\hat{v}(t)-v(t)], \end{aligned}$$

where

$$y_1(t) = D_1(z)[\hat{v}(t)-v(t)], \quad D_1(z) = D(z) - \rho.$$

Here, $\tilde{y}(t)$ may be regarded as the output of the linear system $D_1(z)$ driven by $\hat{v}(t)-v(t)$. 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 [16], the following inequalities hold:

$$\begin{aligned} &\sum_{i=1}^{t} \frac{2\tilde{y}_1(i)[\hat{v}(i)-v(i)]}{r(i-1)} \geq 0, \ \text{a.s.}, \\ S(t) &:= \sum_{i=1}^{t} \frac{2\tilde{y}_1(i)[\hat{v}(i)-v(i)]}{r(i-1)} + \rho \sum_{i=1}^{t} \frac{2[\hat{v}(i)-v(i)]^2}{r(i-1)} \\ &= \sum_{i=1}^{t} \frac{2\tilde{y}(i)[\hat{v}(i)-v(i)]}{r(i-1)} \geq 0, \ \text{a.s.} \end{aligned}$$

Adding both sides of (10) by $S(t)$ gives

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

Since $S(t-1)$, $\varphi^{\mathrm{T}}(t)\tilde{\theta}(t-1)$, $r(t-1)$, $\varphi(t)$, $e(t)-v(t)$ and $r(t)$ are uncorrelated with $v(t)$ and are $\mathscr{F}_{t-1}$ measurable, taking the conditional expectation on both sides of the above equation with respect to $\mathscr{F}_{t-1}$ and using (A1)-(A3) give

$$\begin{aligned} &\mathrm{E}[\|\tilde{\theta}(t)\|^2 + S(t)|\mathscr{F}_{t-1}] = \|\tilde{\theta}(t-1)\|^2 + S(t-1) \\ &\quad + \frac{\|\varphi(t)\|^2\sigma_v^2(t)}{r(t-1)r(t)} - \mathrm{E}\left[\frac{\|\varphi(t)\|^2}{r^2(t)} e^2(t)|\mathscr{F}_{t-1}\right] \\ &\leq \|\tilde{\theta}(t-1)\|^2 + S(t-1) + \frac{\|\varphi(t)\|^2\sigma_v^2}{r(t-1)r(t)} \\ &\quad - \mathrm{E}\left[\frac{\|\varphi(t)\|^2}{r^2(t)} e^2(t)|\mathscr{F}_{t-1}\right], \ \text{a.s.} \end{aligned} \tag{11}$$

Since the sum of the right-hand second last term of (11) for t from 1 to ∞ is finite (a.s.):

$$\sum_{t=1}^{\infty} \frac{\|\varphi(t)\|^2}{r^2(t)} \leq \sum_{t=1}^{\infty} \frac{\|\varphi(t)\|^2}{r(t-1)r(t)} = \frac{1}{r(0)} - \frac{1}{r(\infty)} < \infty,$$

applying the martingale convergence theorem (Lemma D.5.3 in [16]) to (11), we conclude that $\|\tilde{\theta}(t)\|^2 + S(t)$ converges a.s. to a finite random variable, say, C; i.e.,

$$\lim_{t\to\infty} \|\tilde{\theta}(t)\|^2 + S(t) = C < \infty, \ \text{a.s.}, \tag{12}$$

and also

$$\sum_{t=1}^{\infty} \|\Delta\tilde{\theta}(t)\|^2 = \sum_{t=1}^{\infty} \frac{\|\varphi(t)\|^2}{r^2(t)} e^2(t) < \infty, \ \text{a.s.} \tag{13}$$

It follows that

$$\sum_{t=1}^{\infty} \frac{\tilde{y}^2(t)}{r(t-1)} < \infty, \ \text{a.s.}, \ \sum_{t=1}^{\infty} \frac{[\hat{v}(t)-v(t)]^2}{r(t-1)} < \infty, \ \text{a.s.} \tag{14}$$

Equation (12) implies that the parameter estimation error is consistently bounded without the assumption of persistent excitation.

From (9) and (13), it is not difficult to get

$$\begin{aligned} \tilde{\theta}(t) &= \tilde{\theta}(t-1) + \frac{\varphi(t)}{r(t)} e(t) \\ &= \tilde{\theta}(t-j) + \sum_{i=0}^{j-1} \frac{\varphi(t-i)}{r(t-i)} e(t-i); \end{aligned}$$

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

Summing from $t = j$ to ∞ and using (13) give

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

Hence, we have

$$\begin{aligned} &\sum_{t=1}^{\infty} \frac{[e(t)-v(t)]^2}{r(t-1)} = \sum_{i=1}^{\infty} \frac{[y(t)-\varphi^{\mathrm{T}}(t)\hat{\theta}(t-1)-v(t)]^2}{r(t-1)} \\ &= \sum_{i=1}^{\infty} \frac{[\hat{v}(t)-v(t)+\varphi^{\mathrm{T}}(t)(\hat{\theta}(t)-\hat{\theta}(t-1))]^2}{r(t-1)} \\ &\leq \sum_{i=1}^{\infty} \frac{2[\hat{v}(t)-v(t)]^2}{r(t-1)} + \sum_{i=1}^{\infty} \frac{2[\varphi^{\mathrm{T}}(t)(\hat{\theta}(t)-\hat{\theta}(t-1))]^2}{r(t-1)} \\ &\leq \sum_{i=1}^{\infty} \frac{2[\hat{v}(t)-v(t)]^2}{r(t-1)} + C_1 \sum_{i=1}^{\infty} \|\hat{\theta}(t) - \hat{\theta}(t-1)\|^2 < \infty. \end{aligned}$$

Here, $C_1 < \infty$. Taking the norm on both sides of (9) gives

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

Taking the conditional expectation on both sides of the above equation with respect to $\mathscr{F}_{t-1}$ yields

$$\begin{aligned}\mathrm{E}[\|\tilde{\theta}(t)\|^2|\mathscr{F}_{t-1}] &= \|\tilde{\theta}(t-1)\|^2 + \tfrac{2\varphi^{\mathrm{T}}(t)\tilde{\theta}(t-1)}{r(t)}[e(t)-v(t)] \\ &\quad + \tfrac{\|\varphi(t)\|^2}{r^2(t)}[e(t)-v(t)]^2 + \tfrac{\|\varphi(t)\|^2}{r^2(t)}v^2(t), \text{ a.s.}\end{aligned}$$

Since the last three terms on the right-hand side of the above equation are absolutely summable, $\|\tilde{\theta}(t)\|^2$ converge a.s. to a finite random variable.

From (9), we have

$$\tilde{\theta}(t+j) = \tilde{\theta}(t) + \sum_{i=1}^{j}\Delta\tilde{\theta}(t+i). \qquad (15)$$

Replacing t in (8) with $t+j$ gives

$$\varphi^{\mathrm{T}}(t+j)\tilde{\theta}(t+j) = -\tilde{y}(t+j).$$

By using (15), it follows that

$$\varphi^{\mathrm{T}}(t+j)\left[\tilde{\theta}(t) + \sum_{i=1}^{j}\Delta\tilde{\theta}(t+i)\right] = -\tilde{y}(t+j),$$

or

$$\varphi^{\mathrm{T}}(t+j)\tilde{\theta}(t) = -\tilde{y}(t+j) - \varphi^{\mathrm{T}}(t+j)\sum_{i=1}^{j}\Delta\tilde{\theta}(t+i).$$

Squaring and summing for j from $j=0$ to $j=N-1$ give

$$\begin{aligned}&\tilde{\theta}^{\mathrm{T}}(t)\left[\sum_{j=t}^{t+N-1}\varphi(j)\varphi^{\mathrm{T}}(j)\right]\tilde{\theta}(t) \\ &= \sum_{j=0}^{N-1}\left[\tilde{y}(t+j) + \varphi^{\mathrm{T}}(t+j)\sum_{i=1}^{j}\Delta\tilde{\theta}(t+i)\right]^2 \\ &\le 2\sum_{j=0}^{N-1}\left[\tilde{y}^2(t+j) + [\varphi^{\mathrm{T}}(t+j)\sum_{i=1}^{j}\Delta\tilde{\theta}(t+i)]^2\right]\end{aligned}$$

$$\begin{aligned}&\le 2\sum_{j=0}^{N-1}[\tilde{y}^2(t+j) + \|\varphi^{\mathrm{T}}(t+j)\|^2\|\sum_{i=1}^{j}\Delta\tilde{\theta}(t+i)\|^2] \\ &\le 2\sum_{j=0}^{N-1}[\tilde{y}^2(t+j) + N\|\varphi^{\mathrm{T}}(t+j)\|^2\sum_{i=0}^{N-1}\|\Delta\tilde{\theta}(t+i)\|^2].\end{aligned}$$

Using (A4) and dividing $r(t)$, and noting that for $0 \le j \le N-1$, $\|\varphi(t+j)\|/r(t) \le c_1 < \infty$, a.s., we get easily

$$\begin{aligned}\tfrac{\|\tilde{\theta}(t)\|^2}{r(t)} &\le \tfrac{2}{Nc}\sum_{j=0}^{N-1}\tfrac{\tilde{y}^2(t+j)}{r(t)} \\ &\quad + \tfrac{2}{Nc}\sum_{j=0}^{N-1}\tfrac{N\|\varphi(t+j)\|^2}{r(t)}\sum_{i=0}^{N-1}\|\Delta\tilde{\theta}(t+i)\|^2 \\ &\le \tfrac{2}{Nc}\sum_{j=0}^{N-1}\left[\tfrac{\tilde{y}^2(t+j)}{r(t)} + Nc_1\sum_{i=0}^{N-1}\|\Delta\tilde{\theta}(t+i)\|^2\right] \\ &\le \tfrac{2}{Nc}\sum_{j=0}^{N-1}\tfrac{\tilde{y}^2(t+j)}{r(t)} + \tfrac{2Nc_1}{c}\sum_{i=0}^{N-1}\|\Delta\tilde{\theta}(t+i)\|^2 \\ &= \tfrac{2}{Nc}\sum_{j=0}^{N-1}\tfrac{\tilde{y}^2(t+j)}{r(t)} + \tfrac{2Nc_1}{c}\sum_{j=0}^{N-1}\|\Delta\tilde{\theta}(t+j)\|^2, \text{ a.s.}\end{aligned}$$

Summing for t from 1 to ∞ gives

$$\begin{aligned}\sum_{t=1}^{\infty}\tfrac{\|\tilde{\theta}(t)\|^2}{r(t)} \quad &\le \quad \tfrac{2}{Nc}\sum_{j=0}^{N-1}\sum_{t=1}^{\infty}\tfrac{\tilde{y}^2(t+j)}{r(t)} \\ &\quad + \tfrac{2Nc_1}{c}\sum_{j=0}^{N-1}\sum_{t=1}^{\infty}\tfrac{\|\Delta\tilde{\theta}(t+j)\|^2}{r(t)}, \text{ a.s.}\end{aligned}$$

Using the relation, $r(t) \sim r(t+j)$, for $0 \le j \le N-1$, and (13) and (14), we have

$$\sum_{t=1}^{\infty}\frac{\|\tilde{\theta}(t)\|^2}{r(t)} < \infty, \text{ a.s.} \qquad (16)$$

Since $\|\tilde{\theta}(t)\|^2$ converges to a finite random variable, and $\sum\limits_{t=1}^{\infty} r^{-1}(t) = \infty$, the application of Lemma 1 yields

$$\lim_{t\to\infty}\|\tilde{\theta}(t)\|^2 = 0, \text{ a.s.}$$

This proves Theorem 1. □

Similarly, using (C2) instead of (A4), we also obtain the same result. In fact, we have

$$\begin{aligned}&\tilde{\theta}^{\mathrm{T}}(t_s)[\sum_{t=t_s}^{t_{s+1}-1}\tfrac{\varphi(t)\varphi^{\mathrm{T}}(t)}{r(t)}]\tilde{\theta}(t_s) \\ &= \sum_{j=0}^{t_s^*-1}\tfrac{1}{r(t_s+j)}[\tilde{y}(t_s+j) + \varphi^{\mathrm{T}}(t_s+j)\sum_{i=1}^{j}\Delta\tilde{\theta}(t_s+i)]^2 \\ &\le 2\sum_{j=0}^{t_s^*-1}\left[\tfrac{\tilde{y}^2(t_s+j)}{r(t_s+j)} + \tfrac{\|\varphi^{\mathrm{T}}(t_s+j)\|^2}{r(t_s+j)}\|\sum_{i=1}^{j}\Delta\tilde{\theta}(t_s+i)\|^2\right] \\ &\le 2\sum_{j=0}^{t_s^*-1}[\tfrac{\tilde{y}^2(t_s+j)}{r(t_s+j)} + \tfrac{t_s^*\|\varphi^{\mathrm{T}}(t_s+j)\|^2}{r(t_s+j)}\sum_{i=0}^{t_s^*-1}\|\Delta\tilde{\theta}(t_s+i)\|^2].\end{aligned}$$

If furthermore assume that as $s \to \infty$,

$$\tfrac{t_s^*\|\varphi^{\mathrm{T}}(t_s+j)\|^2}{r(t_s+j)} < \infty, \text{ a.s.},$$

then using (13) and (14), we have

$$\lim_{s\to\infty}\tilde{\theta}^{\mathrm{T}}(t_s)[\sum_{t=t_s}^{t_{s+1}-1}\tfrac{\varphi(t)\varphi^{\mathrm{T}}(t)}{r(t)}]\tilde{\theta}(t_s) = 0, \text{ a.s.}$$

By using (C2), it follows that $\tilde{\theta}(t_s) \to 0$. Noting that $\|\tilde{\theta}(t)\|^2$ converges to a finite random variable, we have $\tilde{\theta}(t) \to 0$, a.s. as $t \to \infty$.

IV. THE INNOVATION BASED SG ALGORITHM

The innovation based SG algorithm is as follows:

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \frac{\varphi(t)}{r(t)}[y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t-1)], \tag{17}$$

$$r(t) = r(t-1) + \varphi^{\mathrm{T}}(t)\varphi(t), \quad r(0) = 1, \tag{18}$$

$$\begin{aligned} \varphi(t) = [\ & -y(t-1) \ \cdots \ -y(t-n_a) \\ & u(t-d) \ \cdots \ u(t-d-n_b) \\ & e(t-1) \ \cdots \ e(t-n_d) \quad 1 \ ]^{\mathrm{T}}, \end{aligned} \tag{19}$$

$$e(t) = y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t-1) \text{ (innovation)}. \tag{20}$$

Theorem 2: For the system in (2) and the SG algorithm in (17)-(20), assume that Conditions (A1) to (A4) hold, $\sum_{t=1}^{\infty} r^{-1}(t) = \infty$, a.s., and $D(z) - 1/2$ is strictly positive real. Then $\hat{\theta}(t)$ in (17) consistently converges to θ, i.e., $\hat{\theta}(t) \to \theta$ as $t \to \infty$.

Proof Similarly as in the preceding section, we have

$$\tilde{\theta}(t) = \tilde{\theta}(t-1) + \frac{\varphi(t)}{r(t)}e(t).$$

Taking the norm gives

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

Let

$$\tilde{x}(t) = -\varphi^{\mathrm{T}}(t)\tilde{\theta}(t-1) - \frac{1}{2}[e(t)-v(t)].$$

It is easy to get

$$\begin{aligned} \|\tilde{\theta}(t)\|^2 &= \|\tilde{\theta}(t-1)\|^2 - \frac{2[e(t)-v(t)]}{r(t)}\tilde{x}(t) \\ &\quad + \frac{\|\varphi(t)\|^2}{r^2(t)}v^2(t) + \frac{2\varphi^{\mathrm{T}}(t)\tilde{\theta}(t-1)}{r(t)}v(t) \\ &\quad + \frac{2\|\varphi(t)\|^2}{r^2(t)}[e(t)-v(t)]v(t), \text{ a.s.} \end{aligned} \tag{21}$$

From the definitions of $e(t)$ in (20), we have

$$\begin{aligned} & D(z)[e(t)-v(t)] \\ &\quad = D(z)e(t) - A(z)y(t) + z^{-d}B(z)u(t) + f \\ &\quad = e(t) - y(t) + \varphi^{\mathrm{T}}(t)\theta \\ &\quad = -\varphi^{\mathrm{T}}(t)\hat{\theta}(t-1) + \varphi^{\mathrm{T}}(t)\theta \\ &\quad = -\varphi^{\mathrm{T}}(t)\tilde{\theta}(t-1). \end{aligned}$$

Hence

$$\begin{aligned} \tilde{x}(t) &= D(z)[e(t)-v(t)] - \tfrac{1}{2}[e(t)-v(t)] \\ &= [D(z) - \tfrac{1}{2}][e(t)-v(t)]. \end{aligned}$$

Let

$$S_1(t) = \sum_{i=1}^{t} \frac{2\tilde{x}(i)[e(i)-v(i)]}{r(i-1)}.$$

Noting that $S_1(t-1)$, $\varphi(t)$, $\tilde{\theta}(t-1)$, $r(t)$, and $e(t)-v(t)$ are uncorrelated with $v(t)$, and are $\mathscr{F}_{t-1}$ measurable, and $S_1(t) \geq 0$, adding both sides of (21) by $S_1(t)$ and taking the conditional expectation with respect to $\mathscr{F}_{t-1}$ and using (A1)-(A3) give

$$\begin{aligned} & \mathrm{E}[\|\tilde{\theta}(t)\|^2 + S_1(t)|\mathscr{F}_{t-1}] \\ &\quad \leq \|\tilde{\theta}(t-1)\|^2 + S_1(t-1) + \tfrac{\|\varphi(t)\|^2}{r^2(t)}\sigma_v^2(t), \text{ a.s.} \end{aligned}$$

Applying the martingale convergence theorem, we get

$$\lim_{t\to\infty} \|\tilde{\theta}(t)\|^2 + S_1(t) \to C_1 < \infty, \text{ a.s.},$$

Hence

$$\sum_{t=1}^{\infty} \frac{[\varphi^{\mathrm{T}}(t)\tilde{\theta}(t-1)]^2}{r(t-1)} < \infty, \text{ a.s.},$$

$$\sum_{t=1}^{\infty} \frac{[e(t)-v(t)]^2}{r(t-1)} < \infty, \text{ a.s.}$$

The rest of the proof is similar to that in Theorem 1, hence omitted. □

The SG algorithm has low computational effort, but its convergence rate is slow. In order to improve the convergence rate and tracking performance of the SG algorithm, we introduce a forgetting factor λ in the SG algorithm to get the SG algorithm with forgetting factor [refer to as the forgetting gradient algorithm (FG) for short] based on the estimation residuals:

$$\begin{aligned} \hat{\theta}(t) &= \hat{\theta}(t-1) + \frac{\varphi(t)}{r(t)}[y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t-1)], \\ r(t) &= \lambda r(t-1) + \varphi^{\mathrm{T}}(t)\varphi(t), 0 \leq \lambda < 1, \ r(0) = 1, \\ \varphi_s(t) &= [-y(t-1) \ -y(t-2) \ \cdots \ -y(t-n_a) \\ &\qquad u(t-d) \ u(t-d-1) \ \cdots \ u(t-d-n_b)]^{\mathrm{T}}, \\ \varphi(t) &= [\ \varphi_s^{\mathrm{T}}(t) \quad \hat{v}(t-1) \quad \cdots \quad \hat{v}(t-n_d) \quad 1 \ ]^{\mathrm{T}}, \\ \hat{v}(t) &= y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t). \end{aligned}$$

and the FG algorithm based on the innovations:

$$\begin{aligned} \hat{\theta}(t) &= \hat{\theta}(t-1) + \frac{\varphi(t)}{r(t)}[y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t-1)], \\ r(t) &= \lambda r(t-1) + \varphi^{\mathrm{T}}(t)\varphi(t), 0 \leq \lambda < 1, \ r(0) = 1, \\ \varphi_s(t) &= [-y(t-1) \ -y(t-2) \ \cdots \ -y(t-n_a) \\ &\qquad u(t-d) \ u(t-d-1) \ \cdots \ u(t-d-n_b)]^{\mathrm{T}}, \\ \varphi(t) &= [\varphi_s^{\mathrm{T}}(t) \quad e(t-1) \quad \cdots \quad e(t-n_d) \quad 1]^{\mathrm{T}}, \\ e(t) &= y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t-1). \end{aligned}$$

When $\lambda = 1$, the FG algorithm reduces to the SG algorithm; when $\lambda = 0$, the FG algorithm is the projection algorithm.

V. Example

An illustrative example is given in this section, which is based on computer simulation.

Assume that the simulated model takes the following form

$$
\begin{aligned}
&A(z)y(t) = B(z)u(t) + D(z)v(t),\\
&A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} = 1 - 1.60 z^{-1} + 0.80 z^{-2},\\
&B(z) = b_1 z^{-1} + b_2 z^{-2} = 0.85 z^{-1} + 0.65 z^{-2},\\
&D(z) = 1 + d_1 z^{-1} = 1 - 0.64 z^{-1},\\
&\theta = [a_1, a_2, b_1, b_2, d_1]^{\mathrm{T}} = [-1.60, 0.80, 0.85, 0.65, -0.65]^{\mathrm{T}}.
\end{aligned}
$$

Here $\{u(k)\}$ is taken as a persistent excitation signal sequence with zero mean and unit variance, i.e., $\mathrm{E}[u(t)] = 0$, $\mathrm{E}[u^2(t)] = 1.00^2$, and $\{v(k)\}$ as a white noise sequence with zero mean and variance $\sigma_v^2 = 1.00^2$. Apply the estimation residual based FG algorithm to estimate the parameters of this system, the parameter estimates are shown in Table I and the estimation error curves versus t with different forgetting factors are shown in Fig. 1, where $\delta_{\mathrm{ns}} = 35.75\%$ is the noise-to-signal ratio, TP denotes the true parameters.

TABLE I

The parameter estimates ($\lambda = 0.92$)

t	a_1	a_2	b_1	b_2	d_1	δ (%)
100	-1.55137	0.78450	0.62034	0.88793	-0.12258	28.25865
200	-1.51266	0.73322	0.62479	0.81972	-0.24184	22.93692
300	-1.44508	0.73932	0.65924	0.80982	-0.31527	20.25494
500	-1.53164	0.72927	0.69768	0.82605	-0.36074	17.27298
1000	-1.57689	0.78309	0.75764	0.81157	-0.47764	11.40200
1500	-1.53224	0.77323	0.77418	0.70454	-0.54926	6.84309
2000	-1.57709	0.76666	0.80901	0.68164	-0.61728	3.18873
2500	-1.59804	0.77695	0.88703	0.66650	-0.59664	2.92162
3000	-1.58981	0.77614	0.85406	0.65061	-0.63143	1.26732
TP	-1.60000	0.80000	0.85000	0.65000	-0.64000	

The parameter estimates ($\lambda = 1.00$)

t	a_1	a_2	b_1	b_2	d_1	δ (%)
100	-1.56664	0.78609	0.59016	0.88214	-0.07943	30.31538
200	-1.56476	0.78337	0.59052	0.88016	-0.08372	30.11719
300	-1.56072	0.78248	0.59230	0.87841	-0.08752	29.92107
500	-1.55965	0.77877	0.59425	0.87759	-0.09093	29.74825
1000	-1.56060	0.77388	0.59620	0.87672	-0.09592	29.51090
1500	-1.55926	0.77210	0.59747	0.87532	-0.09899	29.35361
2000	-1.55839	0.77100	0.59884	0.87435	-0.10105	29.23820
2500	-1.55694	0.77032	0.59995	0.87362	-0.10265	29.15028
3000	-1.55700	0.76935	0.60058	0.87312	-0.10381	29.08762
TP	-1.60000	0.80000	0.85000	0.65000	-0.64000	

From Table I and Fig. 1, we can see that the convergence rate of the SG algorithm is very slow and introducing a forgetting factor in the SG algorithm can improve the convergence performance of the SG algorithm.

VI. Conclusions

The performance of SG algorithms is analyzed for CARMA models with disturbances; the analysis method used in this paper can be easily extended to study the convergence of SG algorithms for other models, e.g., output error models. The consistency of the forgetting gradient algorithm requires further research.

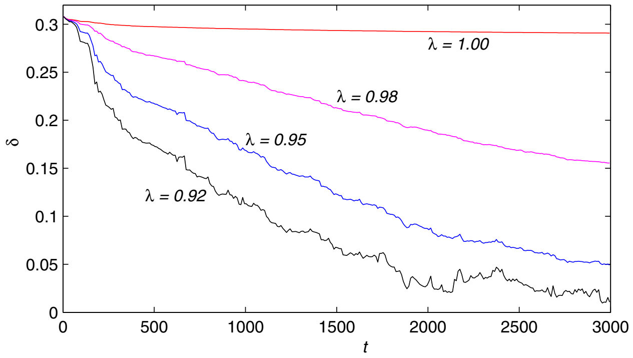

Fig. 1. The parameter estimation error δ vs. t

References

[1] 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.

[2] 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.

[3] 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.

[4] L. Guo and H.F. Chen, “The Astrom-Wittenmark self-tuning regulator revisited and ELS-based adaptive trackers,” *IEEE Transactions on Automatic Control*, vol. 36, no. 7, pp. 802-812, 1991.

[5] K. Toussi and W. Ren, “On the convergence least squares estimates in white noise,” *IEEE Transactions on Automatic Control*, vol. 39, no. 2, pp. 364-368, 1994.

[6] 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.

[7] P. Caines and S. Laforune, “Adaptive control with recursive identification for stochastic linear systems,” *IEEE Transactions on Automatic Control*, vol. 29, no. 4, pp. 312-321, 1984.

[8] M.S. Radenkovi and S. Stankovi, “Strong consistency of parameter estimates in direct self-tuning control algorithms based on stochastic approximation,” *Automatica*, vol. 26, no. 3, pp. 533-544, 1990.

[9] M.S. Radenkovic and A.N. Michel, “Almost sure rate of convergence of the parameter estimates in stochastic approximation algorithm,” *IEEE Transactions on Automatic Control*, vol. 45, no. 6, pp. 1161-1166, 2000.

[10] B. Delyon, General results on the convergence of stochastic algorithms, *IEEE Transactions on Automatic Control*, vol. 41, no. 9, pp. 1245-1255, 1996.

[11] H. Fang and H.F. Chen, “Stability and instability of limit points for stochastic approximation algorithms,” *IEEE Transactions on Automatic Control*, vol. 45, no. 3, pp. 413-420, 2000.

[12] R. Buche and H.J. Kushner, “Stochastic approximation and user adaptation in a competitive resource sharing system,” *IEEE Transactions on Automatic Control*, vol. 45, no. 5, pp. 844-853, 2000.

[13] R. Buche and H.J. Kushner, “Rate of convergence for constrained stochastic approximation algorithms,” *SIAM Journal on Control and Optimization*, vol. 40, no. 4, pp. 1011-1041, 2001.

[14] W. Greblicki, “Stochastic approximation in nonparametric identification of Hammerstein systems,” *IEEE Transactions on Automatic Control*, vol. 47, no. 11, pp. 1800-1810, 2002.

[15] H.F. Chen and P.E. Caines, “The strong consistency of the stochastic gradient algorithm of adaptive control,” *IEEE Transactions on Automatic Control*, vol. 30, no. 2, pp. 189-192, 1985.

[16] G.C. Goodwin and K.S. Sin, *Adaptive Filtering, Prediction and Control*. Englewood Cliffs, Prentice-hall, New Jersey, 1984.