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

Multi-Innovation Stochastic Gradient Algorithm for Output Error Systems Based on the Auxiliary Model

Dongqing Wang, Feng Ding, Peter X. Liu

Abstract— **This paper combines the multi-innovation theory with the auxiliary model identification idea to present the auxiliary model based multi-innovation stochastic gradient algorithm by expanding the scalar innovation to an innovation vector and introducing the innovation length. Convergence analysis in the stochastic framework indicates that the parameter estimation error consistently converges to zero under certain excitation condition. Finally, we illustrate and test the proposed algorithm with an example.**

I. PROBLEM FORMULATION

Consider the output error systems [1], as depicted in Figure 1,

$$y(t)=\frac{B(z)}{A(z)}u(t)+v(t), \tag{1}$$

where $u(t)$ is the system input, $x(t):=\frac{B(z)}{A(z)}u(t)$ is the true output or noise-free output, $v(t)$ is a white noise with zero mean, $y(t)$ is the measurement of $x(t)$, z^{-1} represents a unit delay operator $[z^{-1}y(t)=y(t-1)]$, $A(z)$ and $B(z)$ are polynomials of degrees n_a and n_b, and represented as

$$A(z)=1+a_1z^{-1}+a_2z^{-2}+\cdots+a_{n_a}z^{-n_a},$$
$$B(z)=b_1z^{-1}+b_2z^{-2}+\cdots+b_{n_b}z^{-n_b}.$$

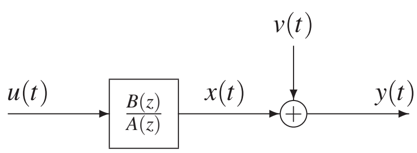

Fig. 1. The output error system with white noises

For the output error systems in (1), the parameter estimates given by the recursive least squares algorithm is biased [2]. In order to get consistently unbiased parameter estimate, many identification methods were published, e.g., the biased compensation least squares algorithms [2]–[4], auxiliary model based least squares identification methods [5], auxiliary model based stochastic gradient identification methods [6], etc. The biased compensation least squares and auxiliary model least squares identification methods have fast convergence rates but require computing the covariance matrices, causing an increased computational complexity.

On the other hand, the stochastic gradient (SG) algorithm requires less computation but has a slower convergence rate than the least squares ones. In order to reduce computational complexity and improve the convergence rate of the SG algorithm, this paper combines the multi-innovation identification theory [7] with the auxiliary model identification idea [5] to study identification problems for output error systems. Difficulties of identification for the output error systems are that there exist unmeasurable true outputs or noise-free outputs in the information vector. This paper, by means of the auxiliary model identification idea, establishes an auxiliary model by using the measurable information and replaces the unknown variables in the information vector with the outputs of the auxiliary model, and presents an auxiliary model based multi-innovation stochastic gradient (AM-MISG) algorithm, thus the identification problems can be solved. The AM-MISG identification method can enhance the parameter estimation accuracy and convergence rates by enlarging the innovation length. The advantage of the AM-MISG algorithm is that it does not involve the covariance matrices [7].

The convergence analysis of identification algorithms are generally based on the stochastic process theory and martingale theory [8]–[11], Ding and Chen discussed the performances of the auxiliary model based stochastic gradient algorithm for dual-rate systems [6] and multi-innovation stochastic gradient algorithm for a linear regression model [7] using the stochastic martingale theory. This paper studies the convergence properties of the AM-MISG algorithm also by using the stochastic martingale theory. While the shows that through a simulation study, the AM-MISG algorithm can improve convergence rate, the underlying PE condition is too strong, and may be even unrealistic in practice [11].

Briefly, the paper is organized as follows. Section II derives an AM-MISG algorithm for output error systems. Section III studies the convergence performance of the AM-MISG algorithm. Section IV provides an illustrative example. Finally, concluding remarks are given in Section V.

This work was supported by the Shandong Province Colleges and Universities Outstanding Young Teachers in Domestic Visiting Scholars Project and by the National Natural Science Foundation of P. R. China (60874020).

D.Q. Wang is with the College of Automation Engineering, Qingdao University (Jiangnan University), Qingdao, P.R. China 266071. dqwang64@163.com

F. Ding is with the Control Science and Engineering Research Center at the Jiangnan University, Wuxi, P.R. China 214122. fding@jiangnan.edu.cn

P.X. Liu is with the Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada K1S 5B6. He is currently a Visiting Professor at the School of Mechatronics Engineering, Nanchang University, Nanchang, China 330031. xpliu@sce.carleton.ca.

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

II. The algorithm description

Define the middle variable,

$$x(t) := \frac{B(z)}{A(z)} u(t). \tag{2}$$

Referring to Figure 1, $x(t)$ is the unknown noise-free outputs (i.e., true outputs), $y(t)$ is the measurements of $x(t)$ corrupted by the additive noise $v(t)$.

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

$$\begin{aligned}\theta &:= [a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b}]^{\mathrm{T}} \in \mathbb{R}^n,\\ \varphi_0(t) &:= [-x(t-1), -x(t-2), \cdots, -x(t-n_a),\\ &\quad u(t-1), -u(t-2), \cdots, u(t-n_b)]^{\mathrm{T}} \in \mathbb{R}^n. \end{aligned} \tag{3}$$

Equations (1)-(2) can be written as

$$x(t) = \varphi_0^{\mathrm{T}}(t)\theta, \tag{4}$$

$$\begin{aligned} y(t) &= x(t) + v(t)\\ &= \varphi_0^{\mathrm{T}}(t)\theta + v(t). \end{aligned} \tag{5}$$

Equation (5) is the identification model for the output error systems.

Let E denote an expectation operator, $\hat{\theta}(t)$ be the estimate of θ at time t, and the norm of the matrix X is defined by $\|X\|^2 := \mathrm{tr}[XX^{\mathrm{T}}]$. Defining and minimizing a quadratic cost function like in [10],

$$J(\theta) = \mathrm{E}[\|y(t) - \varphi_0^{\mathrm{T}}(t)\theta\|^2],$$

leads to the following stochastic gradient algorithms of estimating θ [10],

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \frac{\varphi_0(t)}{r(t)}[y(t) - \varphi_0^{\mathrm{T}}(t)\hat{\theta}(t-1)], \tag{6}$$

$$r(t) = r(t-1) + \|\varphi_0(t)\|^2, \ r(0) = 1. \tag{7}$$

However, the algorithm in (6)-(7) is impossible to realize because the information vector $\varphi_0(t)$ contains the unknown inner variables $x(t-i)$. The solution here is based on the auxiliary model identification idea [5], [6]: these unknown $x(t-i)$ in $\varphi_0(t)$ are replaced with the outputs $\hat{x}(t-i)$ of an auxiliary model in Figure 2, where $\varphi_{\mathrm{a}}(t)$ and $\theta_{\mathrm{a}}(t)$ represent the information vector and parameter vector of the auxiliary model, respectively. Let

$$\begin{aligned}\varphi(t) &:= [-\hat{x}(t-1), -\hat{x}(t-2), \cdots, -\hat{x}(t-n_a),\\ &\quad u(t-1), u(t-2), \cdots, u(t-n_b)]^{\mathrm{T}} \in \mathbb{R}^n. \end{aligned} \tag{8}$$

Here, we take $\varphi(t)$ to be the information vector $\varphi_{\mathrm{a}}(t)$ of the auxiliary model, and $\hat{\theta}(t)$ to be the parameter vector $\theta_{\mathrm{a}}(t)$ of the auxiliary model, thus we have

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

Replacing the unknown $\varphi_0(t)$ in (6)-(7) with $\varphi(t)$ gives the auxiliary model based stochastic gradient identification algorithms (the AM-SG algorithm for short):

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \frac{\varphi(t)}{r(t)} e(t), \tag{10}$$

$$e(t) = y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t-1), \tag{11}$$

$$r(t) = r(t-1) + \|\varphi(t)\|^2, \ r(0) = 1. \tag{12}$$

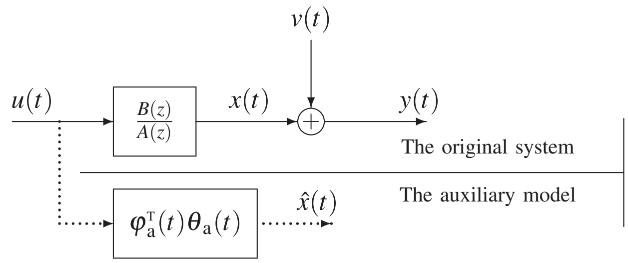

Fig. 2. The output error system with the auxiliary model

Although the above AM-SG algorithm in (8)-(12) can estimate the parameter vector θ, its convergence rate is very poor (see the example later). The following is to derive a multi-innovation stochastic gradient algorithm by expanding the innovation length to improve the parameter estimation accuracy.

Since the scalar quantity $e(t) := y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t-1) \in \mathbb{R}^1$ in (10) is called the innovation and a scalar value [1], the objective of this work is to expand this scalar innovation $e(t) \in \mathbb{R}^1$ to an innovation vector $E(p,t) \in \mathbb{R}^p$ and to present an auxiliary model based multi-innovation stochastic gradient algorithm. The details are as follows.

Define an innovation vector consisting of $e(t-i)$, $i = 0, 1, \cdots, p-1$, as follows:

$$E(p,t) = \begin{bmatrix} e(t) \\ e(t-1) \\ \vdots \\ e(t-p+1) \end{bmatrix} \in \mathbb{R}^p,$$

i.e., multi-innovation (p represents innovation length) and

$$e(t-i) = y(t-i) - \varphi^{\mathrm{T}}(t-i)\hat{\theta}(t-i-1).$$

In general, one thinks that the estimate $\hat{\theta}(t-1)$ is closer to θ than $\hat{\theta}(t-i)$ at time $t-i$ $(i = 2, 3, 4, \cdots, p-1)$. Thus, the innovation vector is taken more reasonably to be

$$E(p,t) := \begin{bmatrix} y(t) - \varphi^{\mathrm{T}}(t)\hat{\theta}(t-1) \\ y(t-1) - \varphi^{\mathrm{T}}(t-1)\hat{\theta}(t-1) \\ \vdots \\ y(t-p+1) - \varphi^{\mathrm{T}}(t-p+1)\hat{\theta}(t-1) \end{bmatrix} \in \mathbb{R}^p.$$

Define the stacked output vector $Y(p,t)$ and information matrix $\Phi(p,t)$ as

$$\begin{aligned} Y(p,t) &:= [y(t), y(t-1), \cdots, y(t-p+1)]^{\mathrm{T}} \in \mathbb{R}^p,\\ \Phi(p,t) &:= [\varphi(t), \varphi(t-1), \cdots, \varphi(t-p+1)] \in \mathbb{R}^{n\times p}. \end{aligned}$$

The innovation vector $E(p,t)$ can be expressed as

$$E(p,t) = Y(p,t) - \Phi^{\mathrm{T}}(p,t)\hat{\theta}(t-1).$$

Since $E(1,t) = e(t)$, $\Phi(1,t) = \varphi(t)$, $Y(1,t) = y(t)$, the AM-SG algorithm in (10)-(12) may be equivalently expressed as

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \frac{\Phi(1,t)}{r(t)} E(1,t), \tag{13}$$

$$E(1,t) = Y(1,t) - \Phi^{\mathrm{T}}(1,t)\hat{\theta}(t-1), \tag{14}$$

$$r(t) = r(t-1) + \|\Phi(1,t)\|^2, \ r(0) = 1. \tag{15}$$

Here, the multi-innovation length p is equal to 1. Referring to [7], we replace the 1's in the above three equations with p to obtain the following auxiliary model based multi-innovation stochastic gradient algorithm with the innovation length p (the AM-MISG algorithm for short):

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \frac{\Phi(p,t)}{r(t)} E(p,t), \quad (16)$$

$$E(p,t) = Y(p,t) - \Phi^{\mathrm{T}}(p,t)\hat{\theta}(t-1), \quad (17)$$

$$r(t) = r(t-1) + \|\Phi(p,t)\|^2, \ r(0) = 1, \quad (18)$$

$$\Phi(p,t) = [\varphi(t), \varphi(t-1), \cdots, \varphi(t-p+1)], \quad (19)$$

$$Y(p,t) = [y(t), y(t-1), \cdots, y(t-p+1)]^{\mathrm{T}}, \quad (20)$$

$$\varphi(t) = [-\hat{x}(t-1), -\hat{x}(t-2), \cdots, -\hat{x}(t-n_a), u(t-1), u(t-2), \cdots, u(t-n_b)]^{\mathrm{T}}, \quad (21)$$

$$\hat{x}(t) = \varphi^{\mathrm{T}}(t)\hat{\theta}(t). \quad (22)$$

When the innovation length $p = 1$, the AM-MISG Algorithm degrades to the AM-SG algorithm.

III. Convergence analysis

Let us introduce some notations first. $\lambda_{\min}[X]$ represents the minimum eigenvalue of the X. For $g(t) \geq 0$, we write $f(t) = O(g(t))$ if there exist positive constants δ_1 and t_0 such that $|f(t)| \leq \delta_1 g(t)$ for $t \geq t_0$ and $f(t) = o(g(t))$ if $f(t)/g(t) \to 0$ as $t \to \infty$.

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

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., $\mathscr{F}_t = \sigma(v(t), v(t-1), v(t-2), \cdots,)$ [10]. The sequence $\{v(t)\}$ satisfies

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

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

Theorem 1: For the system in (5) and the AM-MISG algorithm in (16)-(22), define

$$R(t) := \sum_{i=1}^{t} \Phi(p,i)\Phi^{\mathrm{T}}(p,i),$$

and assume that (A1) and (A2) hold, $A(z)$ is strictly positive real and

$$(A3) \quad r(t) = O(\lambda_{\min}[R(t)]), \ \text{a.s.}$$

Then, the parameter estimation vector $\hat{\theta}(t)$ consistently converges to the true parameter vector θ.

The proof is omitted but available from the authors.

IV. Example

Consider the following output error model,

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

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} = 1 - 0.60 z^{-1} + 0.40 z^{-2},$$

$$B(z) = b_1 z^{-1} + b_2 z^{-2} = -0.20 z^{-1} + 0.80 z^{-2},$$

$$\theta = [a_1, a_2, b_1, b_2]^{\mathrm{T}} = [-0.60, 0.40, -0.20, 0.80]^{\mathrm{T}}.$$

The input $\{u(t)\}$ is taken as an uncorrelated persistent excitation signal sequence with zero mean and unit variance $\sigma_u^2 = 1.00^2$ and $\{v(t)\}$ as a white noise sequence with zero mean and variance $\sigma^2 = 0.10^2$, the corresponding noise-to-signal ratio is $\delta_{\mathrm{ns}} = 11.24\%$. Applying the AM-MISG algorithm with different innovation length $p = 1$, 2, and 5 to estimate the parameters of this system, the parameter estimates and their errors are shown in Table I and the estimation errors $\delta := \|\hat{\theta}(t) - \theta\|/\|\theta\|$ versus t are shown in Figure 3.

From Table I and Figure 3, the parameter estimates converge fast to their true values for large p. The parameter estimation errors become (generally) smaller and smaller with the data length t increasing. This shows that the proposed algorithm is effective.

V. Conclusions

According to the auxiliary model identification idea and the multi-innovation theory, an auxiliary model based multi-innovation stochastic gradient identification algorithm is developed for output error systems. By using the outputs of an auxiliary models to replace the unknown variables in the information vector, the multi-innovation algorithm repeatedly uses innovations by expanding the scalar innovation to the innovation vector and introducing the innovation length. The algorithm proposed requires less computation than existing RLS algorithms and has high accurate parameter estimation than AM-SG algorithm. Convergence analysis shows that the parameter estimation error consistently converges to zero under certain excitation condition.

References

[1] L. Ljung, System Identification: Theory for the User, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[2] W.X. Zheng, "On a least-squares-based algorithm for identification of stochastic linear systems," *IEEE Transactions on Signal Processing*, vol. 46, no. 6, pp. 1631-1638, 1998.

[3] W.X. Zheng, "A bias correction method for identification of linear dynamic errors-in-variables models," *IEEE Transactions on Automatic Control*, vol. 47, no. 7, pp. 1142-1147, 2002.

[4] W.X. Zheng, "Parameter estimation of stochastic linear systems with noisy input," *International Journal of Systems Science*, vol. 35, no. 3, pp. 185-190, 2004.

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

[6] 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.

[7] F. Ding, T. Chen, "Performance analysis of multi-innovation gradient type identification methods," *Automatica*, vol. 43, no. 1, pp. 1-14, 2007.

[8] T.L. Lai, 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.

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

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

[11] Lai T L, Wei C Z. 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.

TABLE I

THE PARAMETER ESTIMATES AND ERRORS ($\sigma^2 = 0.10^2$, $\delta_{ns} = 11.24\%$)

Algorithms	t	a_1	a_2	b_1	b_2	δ (%)
AM-SG	100	-0.15168	0.05247	-0.15359	0.50719	58.42788
(AM-MISG,	200	-0.20049	0.07145	-0.16443	0.54845	52.60657
	300	-0.22737	0.08445	-0.18157	0.56502	49.49584
$p = 1$)	500	-0.25693	0.09403	-0.17869	0.59336	46.04954
	1000	-0.30184	0.11304	-0.18951	0.62565	41.00311
	1500	-0.32361	0.12108	-0.18901	0.64176	38.65996
	2000	-0.33849	0.12878	-0.18877	0.65281	36.93897
	2500	-0.35229	0.13475	-0.19025	0.66210	35.45282
	3000	-0.36042	0.13965	-0.18998	0.66647	34.53358
AM-MISG	100	-0.28918	0.29401	-0.21237	0.64757	33.06925
$p = 2$	200	-0.36741	0.29036	-0.20461	0.67859	25.96192
	300	-0.40500	0.29634	-0.21546	0.68750	22.66928
	500	-0.44380	0.29521	-0.20643	0.71378	18.89825
	1000	-0.49097	0.30548	-0.20809	0.73697	14.39313
	1500	-0.51056	0.31045	-0.20534	0.74743	12.52036
	2000	-0.52182	0.31716	-0.20332	0.75474	11.19285
	2500	-0.53202	0.32128	-0.20322	0.75985	10.18209
	3000	-0.53679	0.32527	-0.20227	0.76223	9.57905
AM-MISG	100	-0.56730	0.36327	-0.20702	0.78656	4.69725
$p = 5$	200	-0.58706	0.37746	-0.20175	0.78915	2.57627
	300	-0.59380	0.38877	-0.20436	0.79454	1.33381
	500	-0.59634	0.38717	-0.20087	0.80295	1.24988
	1000	-0.59774	0.39190	-0.20045	0.80372	0.84066
	1500	-0.59965	0.39242	-0.20089	0.80062	0.70014
	2000	-0.59825	0.39697	-0.19945	0.80105	0.33674
	2500	-0.59957	0.39518	-0.19935	0.80013	0.44597
	3000	-0.59857	0.39725	-0.19941	0.79932	0.29504
True values		-0.60000	0.40000	-0.20000	0.80000	

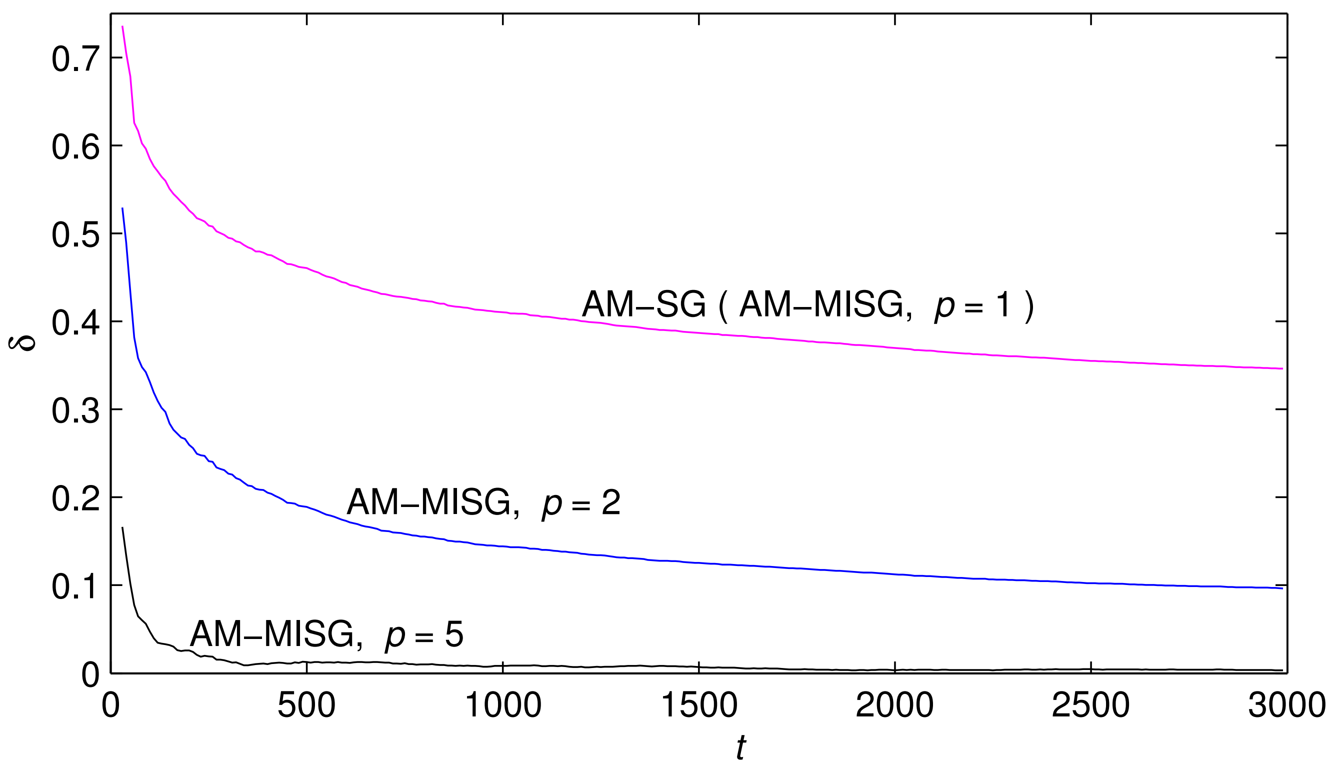

Fig. 3. The estimation errors versus data length t ($\sigma^2 = 0.10^2$)