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,

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

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

\[
A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + \cdots + a_{n_a} z^{-n_a},
\]

\[
B(z) = b_1 z^{-1} + b_2 z^{-2} + \cdots + b_{n_b} z^{-n_b}.
\]

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.
II. THE ALGORITHM DESCRIPTION

Define the middle variable,
\[ x(t) := \frac{B(z)}{A(z)} u(t). \] (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 \( \theta \) and the information vector \( \phi_0(t) \) as
\[ \theta := [a_1, a_2, \cdots, a_{n_a}, b_1, b_2, \cdots, b_{n_b}]^T \in \mathbb{R}^p, \]
\[ \phi_0(t) := [-x(t-1), -x(t-2), \cdots, -x(t-n_a)], \]
\[ u(t-1), -u(t-2), \cdots, u(t-n_b)]^T \in \mathbb{R}^n. \] (3)

Equations (1)-(2) can be written as
\[ x(t) = \phi_0(t) \theta, \]
\[ y(t) = x(t) + v(t) = \phi_0(t) \theta + v(t). \] (4)

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

Let \( E \) denote an expectation operator, \( \hat{\theta}(t) \) be the estimate of \( \theta \) at time \( t \), and the norm of the matrix \( X \) is defined by \( \|X\|^2 := \text{tr}(XX^T) \). Defining and minimizing a quadratic cost function like in [10],
\[ J(\theta) = E[\|y(t) - \phi_0(t) \theta\|^2], \]
leads to the following stochastic gradient algorithms of estimating \( \theta \) [10],
\[ \hat{\theta}(t) = \hat{\theta}(t-1) + \frac{\phi_0(t)}{r(t)} [y(t) - \phi_0(t) \hat{\theta}(t-1)], \] (6)
\[ r(t) = r(t-1) + \|\phi_0(t)\|^2, r(0) = 1. \] (7)

However, the algorithm in (6)-(7) is impossible to realize because the information vector \( \phi_0(t) \) 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 \( \phi_0(t) \) are replaced with the outputs \( \hat{x}(t-i) \) of an auxiliary model in Figure 2, where \( \phi_a(t) \) and \( \theta_a(t) \) represent the information vector and parameter vector of the auxiliary model, respectively. Let
\[ \phi(t) := [-\hat{x}(t-1), -\hat{x}(t-2), \cdots, -\hat{x}(t-n_a)], \]
\[ u(t-1), u(t-2), \cdots, u(t-n_b)]^T \in \mathbb{R}^n. \] (8)

Here, we take \( \phi(t) \) to be the information vector \( \phi_a(t) \) of the auxiliary model, and \( \hat{\theta}(t) \) to be the parameter vector \( \theta_a(t) \) of the auxiliary model, thus we have
\[ \hat{x}(t) = \phi^T(t) \hat{\theta}(t). \] (9)

Replacing the unknown \( \phi_0(t) \) in (6)-(7) with \( \phi(t) \) gives the auxiliary model based stochastic gradient identification algorithms (the AM-SG algorithm for short):
\[ \hat{\theta}(t) = \hat{\theta}(t-1) + \frac{\phi(t)}{r(t)} e(t), \] (10)
\[ e(t) = y(t) - \phi^T(t) \hat{\theta}(t-1), \] (11)
\[ r(t) = r(t-1) + \|\phi(t)\|^2, r(0) = 1. \] (12)

Although the above AM-SG algorithm in (8)-(12) can estimate the parameter vector \( \theta \), 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 \( e(t) := y(t) - \phi^T(t) \hat{\theta}(t-1) \in \mathbb{R}^1 \) in (10) is called the innovation and a scalar value \( [10] \), 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 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
\[ e(t-i) = y(t-i) - \phi^T(t-i) \hat{\theta}(t-i-1). \]

In general, one thinks that the estimate \( \hat{\theta}(t-1) \) is closer to \( \theta \) 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
\[ E(p,t) := \begin{bmatrix} y(t) - \phi^T(t) \hat{\theta}(t-1) \\ y(t-1) - \phi^T(t-1) \hat{\theta}(t-1) \\ \vdots \\ y(t-p+1) - \phi^T(t-p+1) \hat{\theta}(t-1) \end{bmatrix} \in \mathbb{R}^P. \]

Define the stacked output vector \( Y(p,t) \) and information matrix \( \Phi(p,t) \) as
\[ Y(p,t) := [y(t), y(t-1), \cdots, y(t-p+1)]^T \in \mathbb{R}^P, \]
\[ \Phi(p,t) := [\phi(t), \phi(t-1), \cdots, \phi(t-p+1)] \in \mathbb{R}^{n \times P}. \]

The innovation vector \( E(p,t) \) can be expressed as
\[ E(p,t) = Y(p,t) - \Phi^T(p,t) \hat{\theta}(t-1). \]

Since \( E(1,t) = e(t), \Phi(1,t) = \phi(t), Y(1,t) = y(t) \), the AM-SG algorithm in (10)-(12) may be equivalently expressed as
\[ \hat{\theta}(t) = \hat{\theta}(t-1) + \frac{\Phi(1,t)}{r(t)} E(1,t), \] (13)
\[ E(1,t) = Y(1,t) - \Phi^T(1,t) \hat{\theta}(t-1), \] (14)
\[ r(t) = r(t-1) + \|\Phi(1,t)\|^2, r(0) = 1. \] (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):

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

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

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

$$\Phi(p,t) = \{\varphi(t), \varphi(t-1), \cdots, \varphi(t-p+1)\},$$ (19)

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

$$\varphi(t) = [-\hat{x}(t-1), -\hat{x}(t-2), \cdots, -\hat{x}(t-n_d)],$$ (21)

$$\hat{x}(t) = \Phi^T(t)\hat{\theta}(t).$$ (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. $\lambda_{\text{min}}[X]$ represents the minimum eigenvalue of the $X$. For $g(t) \geq 0$, we write $f(t) = O(g(t))$ if there exist positive constants $\delta_l$ and $t_0$ such that $|f(t)| \leq \delta_l g(t)$ 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 $\{v(t), \mathcal{F}_t\}$ is a martingale difference sequence defined on a probability space $\{\Omega, \mathcal{F}, P\}$, where $\{\mathcal{F}_t\}$ is the $\sigma$ algebra sequence generated by $v(t)$, i.e., $\mathcal{F}_t = \sigma\{v(t), v(t-1), \cdots\}$ [10]. The sequence $\{v(t)\}$ satisfies

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

(A2) $E[|v(t)|^2|\mathcal{F}_{t-1}] \leq \sigma_v^2(t) \leq \sigma_v^2 < \infty$, a.s.

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

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

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

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

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

The proof is omitted but available from the authors.

### IV. Example

Consider the following output error model,

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

$$A(z) = 1 + a_1z^{-1} + a_2z^{-2} = 1 - 0.60z^{-1} + 0.40z^{-2},$$

$$B(z) = b_1z^{-1} + b_2z^{-2} = -0.20z^{-1} + 0.80z^{-2},$$

$$\theta = [a_1, a_2, b_1, b_2]^T = [-0.60, 0.40, -0.20, 0.80]^T.$$
TABLE I
THE PARAMETER ESTIMATES AND ERRORS ($\sigma^2 = 0.102$, $\delta_{ns} = 11.24\%$)

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>$t$</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$\delta$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AM-SG</td>
<td>100</td>
<td>-0.13168</td>
<td>0.03247</td>
<td>-0.13359</td>
<td>0.50719</td>
<td>38.42788</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>-0.20049</td>
<td>0.07145</td>
<td>-0.16443</td>
<td>0.54845</td>
<td>52.60657</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>-0.22737</td>
<td>0.08445</td>
<td>-0.18157</td>
<td>0.56502</td>
<td>49.49584</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>-0.25693</td>
<td>0.09403</td>
<td>-0.17869</td>
<td>0.59336</td>
<td>46.04954</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>-0.30184</td>
<td>0.11304</td>
<td>-0.18951</td>
<td>0.62565</td>
<td>41.00311</td>
</tr>
<tr>
<td></td>
<td>1500</td>
<td>-0.32361</td>
<td>0.12108</td>
<td>-0.18901</td>
<td>0.64176</td>
<td>38.65996</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>-0.33849</td>
<td>0.12878</td>
<td>-0.18877</td>
<td>0.65281</td>
<td>36.93897</td>
</tr>
<tr>
<td></td>
<td>2500</td>
<td>-0.35229</td>
<td>0.13475</td>
<td>-0.19025</td>
<td>0.66210</td>
<td>35.45282</td>
</tr>
<tr>
<td></td>
<td>3000</td>
<td>-0.36042</td>
<td>0.13965</td>
<td>-0.18998</td>
<td>0.66647</td>
<td>34.53358</td>
</tr>
</tbody>
</table>

| AM-MISG, $p = 1$ | 100 | -0.28918  | 0.29401   | -0.21237  | 0.64757   | 33.06925     |
| AM-MISG, $p = 2$ | 200 | -0.36741  | 0.29036   | -0.20461  | 0.67859   | 25.96192     |
| AM-MISG, $p = 5$ | 300 | -0.40500  | 0.29521   | -0.20643  | 0.71378   | 18.89825     |
| AM-MISG, $p = 2$ | 500 | -0.49097  | 0.30548   | -0.20809  | 0.73697   | 14.39313     |
| AM-MISG, $p = 5$ | 1000| -0.51056  | 0.31045   | -0.20534  | 0.74743   | 12.52036     |
| AM-MISG, $p = 2$ | 1500| -0.52182  | 0.31716   | -0.20332  | 0.75474   | 11.19285     |
| AM-MISG, $p = 5$ | 2000| -0.53202  | 0.32128   | -0.20322  | 0.75985   | 10.18209     |
| AM-MISG, $p = 2$ | 2500| -0.53679  | 0.32527   | -0.20227  | 0.76223   | 9.57905      |
| AM-MISG, $p = 5$ | 3000| -0.56730  | 0.36327   | -0.20702  | 0.78656   | 4.69725      |
| AM-MISG, $p = 2$ | 100 | -0.38760  | 0.37746   | -0.20175  | 0.78915   | 2.57627      |
| AM-MISG, $p = 5$ | 200 | -0.59634  | 0.38717   | -0.20087  | 0.80295   | 1.24988      |
| AM-MISG, $p = 2$ | 300 | -0.59974  | 0.39190   | -0.20045  | 0.80372   | 0.84066      |
| AM-MISG, $p = 5$ | 500 | -0.59965  | 0.39242   | -0.20089  | 0.80062   | 0.70014      |
| AM-MISG, $p = 2$ | 1000| -0.59825  | 0.39697   | -0.19945  | 0.80105   | 0.33674      |
| AM-MISG, $p = 5$ | 2000| -0.59957  | 0.39518   | -0.19935  | 0.80013   | 0.44597      |
| AM-MISG, $p = 2$ | 2500| -0.59857  | 0.39725   | -0.19941  | 0.79932   | 0.29504      |
| AM-MISG, $p = 5$ | 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 $t$ ($\sigma^2 = 0.10^2$)