Fractional Stochastic Gradient Algorithm for Time-Delayed Models With Piece-Wise Linear Input Using Self-Organizing Maps Method

JIA TANG
Wuxi Vocational College of Science and Technology, Wuxi 214122, China
e-mail: wuxibangbang@163.com

This work was supported in part by the Natural Science Foundation for Colleges and Universities in Jiangsu Province under Grant 18KJB120009, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20131109.

ABSTRACT
Although stochastic gradient algorithm can identify linear systems with high efficiency, it is inefficient for nonlinear systems for the difficulty in the step-size designing. To overcome this dilemma, this paper proposes a fractional stochastic gradient algorithm for systems with piece-wise linear input. First, the nonlinear system is transformed into a polynomial nonlinear model, then the parameters and time-delay are estimated iteratively based on the fractional stochastic gradient algorithm and self-organizing maps method. In addition, to increase the convergence rates of the fractional stochastic gradient algorithm, a multi-innovation fractional stochastic gradient algorithm is developed. Convergence analysis and simulation examples are introduced to show the effectiveness of the proposed algorithms.

INDEX TERMS
System identification, fractional stochastic gradient algorithm, multi-innovation, piece-wise linear, nonlinear model, self-organizing maps.

I. INTRODUCTION
Parameter estimation plays an important role in control theory and application [1], [2], [3]. A robust controller or an accurate predicted model is designed based on accurate parameters of the dynamic systems [4], [5], [6], [7]. The Stochastic Gradient (SG) and Least Squares (LS) algorithms are two classical identification algorithms. The LS algorithm updates the parameters by solving a derivative function, it has faster convergence rates but with the cost of heavier computational efforts when compared with the SG algorithm [8], [9]. In addition, if the considered model has a complex structure, the derivative function may not have analytic solutions. Therefore, the LS algorithm can be inefficient for systems with complex nonlinear structures.

The SG algorithm is a perfect alternative of the LS algorithm. It does not require function calculation and has less computational efforts. The basic idea of the SG algorithm is first to design a direction, which is usually called negative gradient direction; and then to compute a suitable step-size for the direction [10], [11], [12]. Actually, the direction and step-size are two main factors in the SG algorithm designing. To improve the efficiency of the SG algorithm, one can choose a better direction, e.g., the conjugate SG algorithm [13], and the momentum SG algorithm [14]; or can design a suitable step-size, e.g., the forgetting factor SG algorithm [15], and the modified SG algorithm [16]. However, the SG algorithm has slow convergence rates due to its zigzagging nature [17].

The Fractional SG (FSG) algorithm, which was proposed in [18] and [19], is an outstanding modified SG algorithm. It introduces a fractional gradient direction in the SG algorithm. Thanks to this additional direction, the convergence rates of the FSG algorithm are faster than those of the SG algorithm. For example, Muhammad et al proposed a two-stage Fractional Least Mean Square (FLMS) identification algorithm for CARMA systems, where the systems are decomposed into two parts, and the two parts are
identified by the FLMS algorithm [20]. Naveed et al developed two modified fractional SG algorithms for power signal modeling, and these two algorithms have less computational efforts than the FSG algorithm [21]. Recently, Xu et al provided a momentum-based FSG and an adaptive-based FSG algorithms for time-delayed ARX models, these two algorithms have fast convergence rates and less computational efforts [22]. However, all the above work assumed that the models are linear.

In engineering practices, the linear model usually cannot well catch the dynamics of the systems, while the nonlinear model can [23], [24], and [25]. Therefore, nonlinear model identification is more important. In the past few decades, a plethora of SG algorithms are proposed for different kinds of nonlinear models, e.g., the polynomial nonlinear model [26], [27], the separable nonlinear model [28], [29], and the hard nonlinear model [30], [31]. Among them, the hard nonlinear model identification is more difficult. Because designing a suitable step-size is challenging for the special structures of the hard nonlinear models.

This paper proposes an FSG algorithm for time-delayed nonlinear models with piece-wise linear input. By using the self-organizing maps method, the time-delay can be estimated. Then, based on the estimated time-delay, the parameters are updated using the FSG algorithm. To improve the convergence rates, a multi-innovation FSG algorithm is also proposed. Compared with the previous work, the method in this paper has the following contributions: (1) the method does not need to decompose the model into two sub-models, thus has a simpler structure than the structure in [20]; (2) the method involves more directions in each sampling instant, thus has faster convergence rates than the method in [21]; (3) the considered model is more difficult than the model in [22].

The paper is organized as follows: Section II describes the time-delayed nonlinear model and the fractional calculus. Section III proposes the FSG Self-Organizing Maps (FSG-SOM) algorithm. Section IV gives the properties of the FSG-SOM algorithm. Two simulation examples are provided in Section V. Finally, conclusions are presented in Section VI.

II. PROBLEM STATEMENT
A. NONLINEAR MODEL WITH PIECE-WISE LINEAR INPUT
Consider the following time-delayed model with piece-wise linear input [32],

\[ y_t = B(z)f(u_{t-\tau}) + v_t, \]

where \( y_t \) is the system output, \( u_t \) is the system input with bounded values that is persistently excited, \( v_t \) is the Gaussian white noise with zero mean, \( B(z) \) is a polynomial which is written by

\[ B(z) = 1 + b_1z^{-1} + b_2z^{-2} + \cdots + b_nz^{-n}, \]

where \( z^{-1} - r \) is the time delay, \( f(u_t) \) is the piece-wise linear input which is shown in FIGURE 1, where \( k_1 \) and \( k_2 \) are unknown constants. Then, \( f(u_t) \) is written by

\[ f(u_t) = \begin{cases} k_1u_t, & u_t \geq 0, \\ k_2u_t, & u_t < 0. \end{cases} \] (2)

Define a switching function,

\[ h(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0. \end{cases} \] (3)

Based on Equations (2) and (3), it gives rise to

\[ f(u(t)) = k_1h(u_t)u_t + k_2h(-u_t)u_t. \]

Then, the time-delayed model is written by

\[ y_t = \psi_{t-\tau}^T \vartheta + v_t, \]

\[ \psi_{t-\tau} = [h(u_{t-\tau})u_{t-\tau}, h(-u_{t-\tau})u_{t-\tau-1}, h(u_{t-\tau})u_{t-\tau-1}, \cdots, h(u_{t-\tau-\eta})u_{t-\tau-\eta}, h(-u_{t-\tau-\eta})u_{t-\tau-\eta}] \in \mathbb{R}^{2\eta + 2}, \]

\[ \vartheta = [k_1, k_2, b_1k_1, b_1k_2, \cdots, b_nk_1, b_nk_2]^T \in \mathbb{R}^{2n+2}. \]

Using the SG algorithm for the above model, we have

\[ \vartheta_t = \vartheta_{t-1} + \frac{1}{r_t}[y_t - \psi_{t-\tau}^T \vartheta_{t-1}], \]

\[ r_t = r_{t-1} + \psi_{t-\tau}^T \psi_{t-\tau}, \quad r_0 = 0. \]

Two challenges exist when identify the time-delayed model: (1) the information vector contains unknown variables because of the unknown time-delay \( \tau \); (2) the convergence rates of the SG algorithm are slow for its zigzagging nature.

Remark 1: Based on the parameter vector \( \vartheta \), once we obtain its estimates, we can recovery the parameters \( b_i \) based on \( b_i = \frac{b_i}{k_1} \) or we can get \( b_i = k_i k_2 \).

B. FRACTIONAL DERIVATIVE
To increase the convergence rates, Newton iterative algorithm is a good choice. However, the Newton method involves matrix inverse in each iteration which may lead to heavy computational efforts or even make the algorithm divergent for an ill-conditioned information matrix. Then, we use the

\[ FSG-SOM \] algorithm. Two simulation examples are provided (FSG-SOM) algorithm. Section IV gives the properties of the
FSG algorithm to establish the linkage between the SG and Newton algorithms.

The Riemann-Liouville (RL) fractional derivative, the Caputo (CAP) fractional derivative, and the Grünwald-Letnikov (GL) fractional derivative are three classical methods which can obtain the fractional derivative.

The three classical fractional derivation methods are defined as

\[ \text{RL derivative:} \quad D^\alpha_a f(t) = \frac{1}{\Gamma(n-\alpha)} \frac{d^n}{dt^n} \int_a^t (t-s)^{n-\alpha-1} f(s) \, ds, \]

where \( \Gamma(\alpha) = \alpha \times \Gamma(\alpha-1) \).

2: The CAP fractional derivative

\[ \text{CAP derivative:} \quad D^\alpha_a f(t) = \frac{1}{\Gamma(n-\alpha)} \int_a^t (t-s)^{n-\alpha-1} f(s) \, ds. \]

3: The GL fractional derivative

\[ \text{GL derivative:} \quad D^\alpha_a f(t) = \lim_{h \to 0} \frac{1}{h^\alpha} \sum_{k=0}^{n} (-1)^k \Gamma(\alpha + 1) \frac{f(x - kh)}{k! \Gamma(\alpha - k + 1)}. \]

Take \( f(x) = x^2 \) for an example, the \( \alpha \)-order fractional derivative of \( f(x) \) is written by

\[ f^\alpha(x) = \frac{\Gamma(3)}{\Gamma(3-\alpha)} x^{2-\alpha}. \]

III. FSG BASED SOM ALGORITHM FOR TIME-DELAYED NONLINEAR MODEL

In this section, we use the SOM method to obtain the time-delay and then apply the FSG algorithm to estimate the parameters.

A. SOM METHOD

Since we have no knowledge of the time-delay. Let the upper bound of the time-delay be \( M \), then we can get \( M \) sub-models,

\[ y_t = \psi_{t-1}^T \theta + v_t, \]

\[ \vdots \]

\[ y_t = \psi_{t-M}^T \theta + v_t. \]

Assume that the parameters in the sampling instant \( t - 1 \) are \( \hat{\theta}_{t-1} \), then the \( M \) residual errors are computed by

\[ \epsilon_{t-1}^1 = \| y_t - \psi_{t-1}^T \hat{\theta}_{t-1} \|, \]

\[ \vdots \]

\[ \epsilon_{t-1}^M = \| y_t - \psi_{t-M}^T \hat{\theta}_{t-1} \|. \]

Choosing the smallest error among all the residual errors yields

\[ \epsilon_{t-1}^{\text{index}} = \arg \min \{ \epsilon_{t-1}^1, \epsilon_{t-1}^2, \ldots, \epsilon_{t-1}^M \}. \]

For example, if the \( \text{index} = 4 \) in the sampling instant \( t - 1 \), then we can get that the time-delay is \( \tilde{t}_{t-1} = 4 \).

Remark 2: The SOM method updates the time-delay with a binary choice, the weights of the sub-model are 1 or 0. This is different from the method in the expectation maximization (EM) algorithm which assigns different weights for each sub-model.

Let

\[ M = \epsilon_{t-1}^{\text{max}}, \]

where

\[ \epsilon_{t-1}^{\text{max}} = \arg \max \{ \epsilon_{t-1}^1, \epsilon_{t-1}^2, \ldots, \epsilon_{t-1}^M \}. \]

For example, in the EM algorithm, for the sub-model \( j \), its cost function is written by

\[ J(\theta | \tau = j) = \frac{\epsilon_{t-1}^{\text{max}} - \epsilon_{t-1}^j}{2(M \epsilon_{t-1}^{\text{max}} - \sum_{i=1}^{M} \epsilon_{t-1}^i)} [y_t - \psi_{t-j}^T \theta]^2. \]

To obtain the parameters in the sampling instant \( t \), the cost function is written by

\[ J(\theta | \tau = \tilde{t}_{t-1}) = \frac{1}{2} [y_t - \psi_{t-\tilde{t}_{t-1}}^T \theta]^2. \]

With a comparison, the cost function of the SOM method is

\[ J(\theta | \tau = \tilde{t}_{t-1}) = \frac{1}{2} [y_t - \psi_{t-\tilde{t}_{t-1}}^T \theta]^2. \]

Remark 3: Equations (8) and (9) show that the EM algorithm has heavier computational efforts than the SOM method. However, the EM algorithm assigns different weights for each sub-model, it is more stable than the SOM method.

B. FSG ALGORITHM

Once the time-delay in the sampling instant \( t - 1 \) is obtained, the parameters are estimated next. The parameter estimates of the FSG algorithm in the sampling instant \( t \) are updated by

\[ \frac{\partial J(\theta | \tau = \tilde{t}_{t-1})}{\partial \theta} - \frac{\partial^\alpha J(\theta | \tau = \tilde{t}_{t-1})}{\partial \theta^\alpha}, \]

where

\[ \frac{\partial J(\theta | \tau = \tilde{t}_{t-1})}{\partial \theta} = -\psi_{t-\tilde{t}_{t-1}} [y_t - \psi_{t-\tilde{t}_{t-1}} \theta_{t-1}], \]

and

\[ \frac{\partial^\alpha J(\theta | \tau = \tilde{t}_{t-1})}{\partial \theta^\alpha} = -\psi_{t-\tilde{t}_{t-1}} [y_t - \psi_{t-\tilde{t}_{t-1}} \theta_{t-1}] \frac{\partial^\alpha \theta}{\partial \theta^\alpha}. \]

Remark 4: The FSG algorithm performs an additional fractional gradient direction in the SG algorithm, which involves more information. Thus, it has faster convergence rates than the SG algorithm.
Then, Equation (10) is transformed into
\[
\vartheta_t = \vartheta_{t-1} + r_1 \psi_{t-\hat{t}_{t-1}}[y_t - \psi_{t-\hat{t}_{t-1}}^T \vartheta_{t-1}] + r_2 \frac{1}{\Gamma(2-\alpha)} \vartheta_{t-1}^{1-\alpha} \vartheta_{t-1}^{-1} + \vartheta_{t-1}^{1-\alpha} \vartheta_{t-1}^{-1} \vartheta_{t-1}^{-1}
\]
\[
= \vartheta_{t-1} + \left[ r_1 + r_2 \frac{1}{\Gamma(2-\alpha)} \vartheta_{t-1}^{1-\alpha} \right] \psi_{t-\hat{t}_{t-1}}[y_t - \psi_{t-\hat{t}_{t-1}}^T \vartheta_{t-1}].
\]
Equation (11) is simplified as
\[
\vartheta_t = \vartheta_{t-1} + \beta \psi_{t-\hat{t}_{t-1}}[y_t - \psi_{t-\hat{t}_{t-1}}^T \vartheta_{t-1}].
\]
Remark 5: Assume that \( \beta \) is the step-size and \( \psi_{t-\hat{t}_{t-1}}[y_t - \psi_{t-\hat{t}_{t-1}}^T \vartheta_{t-1}] \) is the negative gradient direction. The FSG algorithm is equivalent to the SG algorithm.

C. FSG BASED SOM ALGORITHM
In the first step, we obtain the time-delay based on the estimated parameters using the SOM method, that is
\[
e_{t-1}^{\text{index}} = \arg \min \{e_{t-1}^1, e_{t-1}^2, \cdots, e_{t-1}^M\},
\]
\[
\hat{t}_{t-1} = \text{index}.
\]
In the second step, we update the parameters based on the estimated time-delay using the FSG algorithm,
\[
\vartheta_t = \vartheta_{t-1} + \beta \psi_{t-\hat{t}_{t-1}}[y_t - \psi_{t-\hat{t}_{t-1}}^T \vartheta_{t-1}].
\]
\[
\beta = \left[ r_1 + r_2 \frac{1}{\Gamma(2-\alpha)} \vartheta_{t-1}^{1-\alpha} \frac{1}{\Gamma(2-\alpha)} \vartheta_{t-1}^{-1} \right].
\]

The steps of the FSG-SOM algorithm are listed as follows.

A. CONVERGENCE ANALYSIS OF THE FSG-SOM ALGORITHM
Theorem 1: For the FSG-SOM algorithm proposed in (13)-(16), the cost functions in the sampling instants \( t - 1 \) and \( t \) are \( J(\vartheta_{t-1}, \hat{t}_{t-1}) \) and \( J(\vartheta_t, \hat{t_t}) \), respectively. Then, the following inequality holds,
\[
J(\vartheta_t, \hat{t_t}) \leq J(\vartheta_{t-1}, \hat{t}_{t-1}).
\]
Proof: Assume that the cost function in the sampling instant \( t - 1 \) is
\[
J(\vartheta_{t-1}, \hat{t}_{t-1}) = \frac{1}{2} [y_t - \psi_{t-\hat{t}_{t-1}}^T \vartheta_{t-1}]^2.
\]
For a fixed time-delay \( \hat{t}_{t-1} \), if we choose a suitable step-size \( \beta \) for the FSG algorithm, we can guarantee that
\[
\frac{1}{2} [y_t - \psi_{t-\hat{t}_{t-1}}^T \vartheta_{t-1}]^2 \leq \frac{1}{2} [y_t - \psi_{t-\hat{t}_{t-1}}^T \vartheta_{t-1}]^2.
\]
That is
\[
J(\vartheta_t, \hat{t_t}) \leq J(\vartheta_{t-1}, \hat{t}_{t-1}).
\]
According to SOM method, we obtain
\[
\hat{t}_t = \arg \min \left\{ \frac{1}{2} [y_t - \psi_{t-\hat{t}_{t-1}}^T \vartheta_{t-1}]^2, \cdots, \frac{1}{2} [y_t - \psi_{t-M}^T \vartheta_{t-1}]^2 \right\}
\]
If follows that
\[
J(\vartheta_t, \hat{t_t}) \leq J(\vartheta_{t-1}, \hat{t}_{t-1}).
\]

The proof is completed.
Remark 7: Theorem 1 only can guarantee that the estimates converge to a stable point. If \( J(\vartheta, \tau) \) is a strictly convex function, it can ensure the estimates to converge to the global optimal point.

FSG-SOM algorithm
\[
\text{Initialize} \ \vartheta_0 = 1/p_0 \text{ with } p_0 \text{ being a column vector whose entries are all unity and } p_0 = 10^6, \text{ and } y_t = 0, u_t = 0, t \leq 0
\]
\[
\text{Repeat}
\]
\[
\text{for } t = 1, 2, \cdots, \text{ do}
\]
\[
\text{Collect } u(t) \text{ and } y(t)
\]
\[
\text{Compute } \hat{t}_{t-1} \text{ based on Equations (13) and (14)}
\]
\[
\text{Form } \psi_{t-\hat{t}_{t-1}}
\]
\[
\text{Compute } \beta \text{ according to (16)}
\]
\[
\text{Update } \vartheta_t \text{ by (15)}
\]
\[
\text{end}
\]
\[
\text{until convergence}
\]
B. THE RANGE OF THE STEP-SIZE $\beta$

In Theorem 1, we should choose a suitable step-size $\beta$ to keep

$$J(\theta_t, \hat{\theta}_{t-1}) \leq J(\hat{\theta}_{t-1}, \hat{\theta}_{t-1}).$$

Subtracting the true value $\hat{\theta}$ on both sides of Equation (15) yields

$$e_t = \theta_t - \hat{\theta} = \theta_{t-1} - \theta + \beta \psi_{t-\hat{\theta}_{t-1}}[y_t - \psi^T_{t-\hat{\theta}_{t-1}} \theta_{t-1}]$$

$$= e_{t-1} - \beta \psi_{t-\hat{\theta}_{t-1}}[\psi^T_{t-\hat{\theta}_{t-1}} e_{t-1}]$$

$$= (I - \beta \psi_{t-\hat{\theta}_{t-1}} \psi^T_{t-\hat{\theta}_{t-1}}) e_{t-1}. \quad (17)$$

To keep the algorithm convergent, one should choose a suitable step-size $\beta$ to make

$$\|I - \beta \psi_{t-\hat{\theta}_{t-1}} \psi^T_{t-\hat{\theta}_{t-1}}\| < 1.$$

Remark 8: In order to keep the FSG algorithm convergent, the step-size $\beta$ should keep the convergence factor $\gamma = \|I - \beta \psi_{t-\hat{\theta}_{t-1}} \psi^T_{t-\hat{\theta}_{t-1}}\| < 1$. Since $\psi_{t-\hat{\theta}_{t-1}}$ is a vector, there was not much way for the choice of $\beta$.

If $\gamma$ satisfies

$$\gamma = \|I - \beta \sum_{i=0}^{l-1} \psi_{t-i-\hat{\theta}_{t-1}} \psi^T_{t-i-\hat{\theta}_{t-1}}\|,$$

there exist lots of methods to choose a suitable step-size $\beta$. In this case, the FSG algorithm can be modified as

$$\theta_t = \theta_{t-1} + \beta \psi_{t-\hat{\theta}_{t-1}}[y_t - \psi^T_{t-\hat{\theta}_{t-1}} \theta_{t-1}]$$

$$+ \beta \psi_{t-\hat{\theta}_{t-1}}[y_{t-1} - \psi^T_{t-\hat{\theta}_{t-1}} \theta_{t-1}]$$

$$+ \cdots + \beta \psi_{t-1+1-\hat{\theta}_{t-1}}[y_{t-1+1-\hat{\theta}_{t-1}} - \psi^T_{t-1+1-\hat{\theta}_{t-1}} \theta_{t-1}]. \quad (18)$$

Clearly, this is the same as the multi-innovation algorithm. It also demonstrates that the Multi-innovation FSG (M-FSG) algorithm has a faster convergence rate than the FSG algorithm but with the cost of heavier computational efforts.

Remark 9: The M-FSG algorithm involves more directions in each iteration than the FSG algorithm, thus, it has faster convergence rates with cost of more computational efforts.

The steps of the M-FSG-SOM algorithm are listed as follows.

**Theorem 2:** Assume $l = 2n + 2$ and $\sum_{i=0}^{l-1} \psi_{t-i-\hat{\theta}_{t-1}} \psi^T_{t-i-\hat{\theta}_{t-1}}$ is nonsingular. The step-size $\beta$ is chosen as

$$\beta = \left(\sum_{i=0}^{l-1} \psi_{t-i-\hat{\theta}_{t-1}} \psi^T_{t-i-\hat{\theta}_{t-1}}\right)^{-1}.$$

Then, the best estimates can be obtained in only one iteration.

**Proof:** Substituting the true value $\hat{\theta}$ on both sides of Equation (18) yields

$$e_t = \theta_t - \hat{\theta} = \theta_{t-1} - \theta + \beta \psi_{t-\hat{\theta}_{t-1}}[y_t - \psi^T_{t-\hat{\theta}_{t-1}} \theta_{t-1}]$$

$$+ \beta \psi_{t-1-i-\hat{\theta}_{t-1}}[y_{t-1} - \psi^T_{t-1-i-\hat{\theta}_{t-1}} \theta_{t-1}] + \cdots +$$

$$+ \beta \psi_{t-1+1-i-\hat{\theta}_{t-1}}[y_{t-1+1-i-\hat{\theta}_{t-1}} - \psi^T_{t-1+1-i-\hat{\theta}_{t-1}} \theta_{t-1}],$$

$$\approx e_{t-1} - \beta \psi_{t-\hat{\theta}_{t-1}} \psi^T_{t-\hat{\theta}_{t-1}} e_{t-1} - \beta \psi_{t-1-i-\hat{\theta}_{t-1}} \psi^T_{t-1-i-\hat{\theta}_{t-1}} e_{t-1} -$$

$$- \cdots - \beta \psi_{t-1+1-i-\hat{\theta}_{t-1}} \psi^T_{t-1+1-i-\hat{\theta}_{t-1}} e_{t-1}$$

$$= \left[I - \beta \sum_{i=0}^{l-1} \psi_{t-i-\hat{\theta}_{t-1}} \psi^T_{t-i-\hat{\theta}_{t-1}}\right] e_{t-1}. \quad (19)$$

Substituting $\beta = \left(\sum_{i=0}^{l-1} \psi_{t-i-\hat{\theta}_{t-1}} \psi^T_{t-i-\hat{\theta}_{t-1}}\right)^{-1}$ into the above equation obtains

$$e_t = \left[I - \left(\sum_{i=0}^{l-1} \psi_{t-i-\hat{\theta}_{t-1}} \psi^T_{t-i-\hat{\theta}_{t-1}}\right)^{-1}\right]$$

$$\times \sum_{i=0}^{l-1} \psi_{t-i-\hat{\theta}_{t-1}} \psi^T_{t-i-\hat{\theta}_{t-1}} e_{t-1}. \quad (20)$$

Clearly, if $l = 2n + 2$ and $\sum_{i=0}^{l-1} \psi_{t-i-\hat{\theta}_{t-1}} \psi^T_{t-i-\hat{\theta}_{t-1}}$ is nonsingular, we have

$$e_t = 0,$$

which means that we can get the best estimates in only one iteration.

**Remark 10:** If the number of the information vectors satisfies $l > 2n + 2$, and the step-size is chosen as

$$\beta = \left(\sum_{i=0}^{l-1} \psi_{t-i-\hat{\theta}_{t-1}} \psi^T_{t-i-\hat{\theta}_{t-1}}\right)^{-1},$$

then the M-FSG algorithm is the same as the LS algorithm where the number of the collected data is $l$.

**Remark 11:** Theorem 2 shows that the larger the number of directions is, the faster convergence rates the M-FSG algorithm will have. However, the larger number of directions will lead to heavier computational efforts.

V. EXAMPLES

A. EXAMPLE 1

Consider a nonlinear model with piece-wise linear input,

$$y_t = (1 + 0.5e^{-1})f(u_{t-\tau}) + v_t,$$
The nonlinear model is simplified as
\[ y_t = (1 + 0.5z^{-1})(0.4h(u_{t-\tau})u_{t-\tau} + 0.2h(-u_{t-\tau})u_{t-\tau} + 0.2h(u_{t-1})u_{t-1-\tau} + 0.1h(-u_{t-1-\tau})u_{t-1-\tau-1}) + \nu_t, \]
\[ \psi(t - \tau) = [h(u_{t-\tau})u_{t-\tau}, h(-u_{t-\tau})u_{t-\tau}, h(u_{t-1})u_{t-1-\tau}, h(-u_{t-1-\tau})u_{t-1-\tau-1})], \]
\[ \vartheta = [\vartheta_1, \vartheta_2, \vartheta_3, \vartheta_4]^T = [0.4, 0.2, 0.2, 0.1]^T. \]

Assume that the time delay is \( \tau = 2 \) and assign \( M = 5 \).

In simulation, we collect 1000 sets of input and output data and assume \( \alpha = 1.2 \). Use the FSG and SG algorithms for this nonlinear model. The parameter estimates and their estimation errors \( \delta = \frac{\|\vartheta_t - \vartheta\|}{\|\vartheta\|} \) are shown in Table 1 and FIGURE 2. The time-delay estimates are shown in FIGURE 3. In addition, Monte Carlo simulations (with 100 different noise seeds) were performed based on the FSG algorithm. The parameter estimation errors are shown in FIGURE 4.

| Algorithms | \( t \) | \( \vartheta_1 \) | \( \vartheta_2 \) | \( \vartheta_3 \) | \( \vartheta_4 \) | \( \delta (%) \) |
|------------|-------|-------------|-------------|-------------|-------------|-------------|
| SG         | 100   | 0.35592     | 0.16219     | 0.28182     | 0.09142     | 20.13961    |
|            | 200   | 0.36054     | 0.16738     | 0.27267     | 0.08930     | 17.90733    |
|            | 400   | 0.36698     | 0.17389     | 0.26360     | 0.08854     | 15.42497    |
|            | 600   | 0.36959     | 0.17656     | 0.25955     | 0.08823     | 14.36436    |
|            | 800   | 0.37160     | 0.17854     | 0.25661     | 0.08808     | 13.58459    |
|            | 1000  | 0.37296     | 0.17993     | 0.25402     | 0.08765     | 12.96771    |
| FSG        | 100   | 0.38592     | 0.19606     | 0.19723     | 0.12495     | 5.81030     |
|            | 200   | 0.38740     | 0.19774     | 0.19629     | 0.12380     | 5.45518     |
|            | 400   | 0.38815     | 0.19851     | 0.19522     | 0.12204     | 5.10379     |
|            | 600   | 0.38853     | 0.19883     | 0.19492     | 0.12136     | 4.95914     |
|            | 800   | 0.38884     | 0.19907     | 0.19474     | 0.12087     | 4.85279     |
|            | 1000  | 0.38917     | 0.19940     | 0.19469     | 0.12059     | 4.77460     |
| True Values|       | 0.40000     | 0.20000     | 0.20000     | 0.10000     |              |

FIGURE 2. The parameter estimation errors.

FIGURE 3. The time-delay estimates.

FIGURE 4. The parameter estimation errors.

From this simulation, we can get the following findings:
(1) The parameter estimates of the FSG and SG algorithms can asymptotically converge to the true values;
(2) The convergence rates of the FSG algorithm are faster than those of the SG algorithm;
(3) Based on the estimates of \( \vartheta_1 \) and \( \vartheta_2 \), we can get \( k_1 \) and \( k_2 \); and based on \( \vartheta_1 \) and \( \vartheta_3 \), we can compute \( b_1 \);
(4) FIGURE 4 shows that the FSG algorithm is robust to the noise.

B. AN OPEN CHANNEL SYSTEM

An open channel system, shown in FIGURE 5, is proposed in this subsection, where \( R \) is the radius, \( x \) is the length of the channel, \( u_t \) is the discharge at the upstream end, \( y_t \) is the discharge at the downstream end, the slope is \( \beta \). In this example, two slopes \( \beta = 10 \) and \( \beta = 15 \) degrees are assigned to the
open channel system, that leads to two different inputs [33],

\[ y(t) = (1 + 0.4097e^{-1})f(u_{t-\tau}) + v(t). \]

We collect 1000 sets of input-output data using Matlab software, where the sequence \{u(t)\} is generated by

\[ u = i d i n p u t (L', ' R B S', [0, 1], [-1, 1]), \]

the data from \( t = 1 : 500 \) belong to \( \beta = 10 \), and those from \( 501 : 1000 \) belong to \( \beta = 15 \). Let \( \sigma = \frac{1}{2}(\text{mean}(u(1 : 500) + \text{mean}(501 : 1000))) \), and the true input are \( u = u - \sigma \).

The piece-wise input is written by

\[
 f(u_t) = \begin{cases} 
 0.52 u_t, & u_t > 0, \\
 0.27 u_t, & u_t < 0. 
\end{cases} 
\]  

Then, the open channel system is written by

\[
 y_t = 0.52 h(u_{t-\tau})u_{t-\tau} + 0.27 h(-u_{t-\tau})u_{t-\tau} + 0.2130 h(u_{t-\tau-1})u_{t-\tau-1} + 0.1106 h(-u_{t-\tau-1})u_{t-\tau-1} + v_t. 
\]

In simulation, we manually impose the time-delay as \( \tau = 1, M = 5 \) and \( \alpha = 0.9 \). Apply the SG, FSG and M-FSG algorithms for the considered nonlinear model. The parameter estimates and their estimation errors \( \delta = \frac{\|\hat{\vartheta}_t - \vartheta\|}{\|\vartheta\|} \) are shown in FIGURE 6 and Table 2. The time-delay estimates are shown in FIGURE 7. The numbers of iterations with almost the same estimation errors using the three algorithms are shown Table 3.

This simulation example shows that all the three algorithms can catch the true values. Among these three algorithms, the M-FSG algorithm has the fastest convergence rates, then is the FSG algorithm, and the SG algorithm has the slowest convergence rates.
VI. CONCLUSION
An FSG algorithm is proposed for piece-wise linear model in this paper. The algorithm has faster convergence rates than the traditional SG algorithm because it introduces a fractional gradient in the SG algorithm. The properties demonstrate that the multi-innovation FSG (M-FSG) algorithm is more efficient than the FSG algorithm. In application, we can choose different numbers of innovations on a case by case basis.

Although the FGS algorithm has several advantages over the SG algorithm, there are some interesting topics need to be further discussed. For example, how to choose a suitable α? and can this algorithm be extended to systems with other kind of nonlinear structures. These topics remain as open problems in future.

Data Availability Statement: All data generated or analyzed during this study are included in this article.

REFERENCES
[1] X. Liu and X. Yang, “Identification of nonlinear state-space systems with skewed measurement noises,” IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 69, no. 11, pp. 4654–4662, Nov. 2022. doi: 10.1109/TCSI.2022.3193444.
[2] J. Chen, J. Ma, M. Gan, and Q. Zhu, “Multidirection gradient iterative algorithm: A unified framework for gradient iterative and least squares algorithms,” IEEE Trans. Autom. Control, vol. 67, no. 12, pp. 6770–6777, Dec. 2022. doi: 10.1109/TAC.2021.3132262.
[3] J. Chen, B. Huang, M. Gan, and C. L. P. Chen, “A novel reduced-order algorithm for rational models based on Arnoldi process and Krylov subspace,” Automatica, vol. 129, Jul. 2021, Art. no. 109663, doi: 10.1016/j.automatica.2021.109663.
[4] G. Pilloonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung, “Kernel methods in system identification, machine learning and function estimation: A survey,” Automatica, vol. 50, no. 3, pp. 657–682, Mar. 2014.
[5] B. Sinquin and M. Verhaegen, “KASID: Large-scale subspace identification with Kroncker modeling,” IEEE Trans. Autom. Control, vol. 64, no. 3, pp. 960–975, Mar. 2019.
[6] B. Ma, T. Chen, and L. Ljung, “On asymptotic properties of hyperparameter estimators for kernel-based regularization methods,” Automatica, vol. 94, pp. 381–395, Aug. 2018.
[7] J. Xu, Y. Xing, and J. Costa-Castello, “Adaptive estimation of time-varying parameters with application to roto-magnet plant,” IEEE Trans. Syst., Man, Cybern. Syst., vol. 51, no. 2, pp. 731–741, Feb. 2021, doi: 10.1109/TSMC.2018.2882844.
[8] T. Chen, M. S. Andersen, L. Ljung, A. Chiuso, and G. Pilloonetto, “System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques,” IEEE Trans. Autom. Control, vol. 59, no. 11, pp. 2933–2945, Nov. 2014.
[9] J. Guo, B. Mu, L. Y. Wang, G. Yin, and L. Xu, “Decision-based system identification and adaptive resource allocation,” IEEE Trans. Autom. Control, vol. 62, no. 5, pp. 2166–2179, May 2017.
[10] Y. Zhou, X. Zhang, and F. Ding, “Partially-coupled nonlinear parameter optimization algorithm for a class of multivariate hybrid models,” Appl. Math. Comput., vol. 414, Feb. 2022, Art. no. 126663.
[11] H. Xu, F. Ding, and B. Champagne, “Joint parameter and time-delay estimation for a class of nonlinear time-series models,” IEEE Signal Process. Lett., vol. 29, pp. 947–951, 2022.
[12] L. Xu, F. Ding, and E. Yang, “Auxiliary model multinnovation stochastic gradient parameter estimation methods for nonlinear sandwich systems,” Int. J. Robust Nonlinear Control, vol. 31, no. 1, pp. 148–165, Jan. 2021.
[13] S. Abbasbandy, A. Jafarian, and R. Ezzati, “Conjugate gradient method for fuzzy symmetric positive definite system of linear equations,” Appl. Math. Comput., vol. 171, no. 2, pp. 1184–1191, Dec. 2005.
[14] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.
[15] Q. Liu, Y. Xiao, F. Ding, and T. Hayat, “Decomposition-based overparameterization forgetting factor stochastic gradient algorithm for Hammerstein–Wiener nonlinear systems with non-uniform sampling,” Int. J. Robust Nonlinear Control, vol. 31, no. 12, pp. 6007–6024, Aug. 2021.
[16] H. Zhao, L. Lv, and Y. Ji, “Improved redundant rule-based stochastic gradient algorithm for time-delayed models using lasso regression,” IEEE Access, vol. 10, pp. 3336–3342, 2022.
[17] Y. Ji and L. Lv, “Two identification methods for a nonlinear membership function,” Complexity, vol. 2021, pp. 1–7, Apr. 2021.
[18] M. A. Z. Raja, Y. He, Z. A. Khan, and J. A. T. Machado, “Design of fractional hierarchical gradient descent algorithm for parameter estimation of nonlinear control autoregressive systems,” Chaos, Solitons Fractals, vol. 157, Apr. 2022, Art. no. 11913.
[19] N. I. Chaudhary, M. A. Z. Raja, Y. He, Z. A. Khan, and J. A. T. Machado, “Design of multi innovation fractional LMS algorithm for parameter estimation of input nonlinear control autoregressive systems,” Appl. Math. Model., vol. 93, pp. 412–425, May 2021.
[20] M. A. Z. Raja and N. I. Chaudhary, “Two-stage fractional least mean square algorithm for parameter estimation of CARMA systems,” Signal Process., vol. 107, pp. 327–339, Feb. 2015.
[21] N. I. Chaudhary, S. Zuhair, and M. A. Z. Raja, “A new computing approach for power signal modeling using fractional adaptive algorithms,” ISA Trans., vol. 68, pp. 189–202, May 2017.
[22] T. Xu, J. Chen, Y. Pu, and L. Guo, “Fractional-based stochastic gradient algorithms for time-delayed ARX models,” Circuits, Syst., Signal Process., vol. 41, no. 4, pp. 1895–1912, Apr. 2022.
[23] G. Y. Chen, M. Gan, C. L. P. Chen, and H. X. Li, “A regularized variable projection algorithm for separable nonlinear least-squares problems,” IEEE Trans. Autom. Control, vol. 64, no. 2, pp. 526–537, Feb. 2019.
[24] F. Bianchi, V. Breschi, D. Piga, and L. Pirolli, “Model structure selection for switched NARX system identification: A randomized approach,” Automatica, vol. 125, Mar. 2021, Art. no. 109415.
[25] Y. Chen, Y. Liu, J. Chen, and J. Ma, “A novel identification method for a class of closed-loop systems based on basis pursuit de-noising,” IEEE Access, vol. 8, pp. 99648–99654, 2020.
[26] Q. Liu and F. Ding, “Gradient-based recursive parameter estimation for a periodically nonuniformly sampled-data Hammerstein–Wiener system based on the key-term separation,” Int. J. Adapt. Control Signal Process., vol. 35, no. 10, pp. 1970–1989, Oct. 2021.
[27] J. Ma, B. Huang, and F. Ding, “Iterative identification of Hammerstein parameter varying systems with parameter uncertainties based on the variational Bayesian approach,” IEEE Trans. Syst., Man, Cybern. Syst., vol. 50, no. 1, pp. 1035–1045, Mar. 2020.
[28] L. Xu, “Separable multi-innovation Newton iterative modeling algorithm for multi-frequency signals based on the sliding measurement window,” Circuits, Syst., Signal Process., vol. 41, no. 2, pp. 805–830, Feb. 2022.
[29] L. Xu, F. Ding, L. Wan, and J. Sheng, “Separable multi-innovation stochastic gradient estimation algorithm for the nonlinear dynamic responses of systems,” Int. J. Adapt. Control Signal Process., vol. 34, no. 7, pp. 937–954, Jul. 2020.
[30] X. Wang, F. Ding, A. Alsaedi, and T. Hayat, “Auxiliary model-based iterative parameter estimation for a nonlinear output-error system with saturation and dead-zone nonlinearity,” Int. J. Robust Nonlinear Control, vol. 31, no. 9, pp. 4262–4286, Jun. 2021.
[31] F. Giri and E. W. Bai, Block-Oriented Nonlinear System Identification. Berlin, Germany: Springer, 2010.
[32] Y. Fan and X. Liu, “Data filtering-based multi-innovation forgetting gradient algorithms for input nonlinear FIR-MA systems with piecewise-linear characteristics,” J. Franklin Inst., vol. 358, no. 18, pp. 9818–9840, Dec. 2021.
[33] L. Bako, K. Boukhourouba, E. Duvieilla, and S. Lecoeuche, “A recursive identification algorithm for switched linear/affine models,” Nonlinear Anal., Hybrid Syst., vol. 5, no. 2, pp. 242–253, May 2011.

JIA JANG received the B.Sc. and M.Sc. degrees from Jiangnan University, Wuxi, China, in 2003 and 2006, respectively. He is currently an Associate Professor with the Wuxi Vocational College of Science and Technology, Wuxi. He has been elected on the Ten Outstanding Young Persons of Wuxi, China. His current research interests include machine learning, process control, and artificial intelligence.