KYBERNETIKA — VOLUME *62* (2026), NUMBER 2, PAGES 286–304

AN EFFICIENT SPARSE IDENTIFICATION ALGORITHM FOR STOCHASTIC SYSTEMS WITH GENERAL OBSERVATION SEQUENCES

Ziming Wang, Mingkun Li, Yiming Xing, and Xinghua Zhu

This paper studies the sparse identification problem of unknown sparse parameter vectors in stochastic dynamic systems. Firstly, a novel sparse identification algorithm is proposed, which can generate sparse estimates based on least squares estimation by adaptively adjusting the threshold. Secondly, under a possibly weakest non-persistent excited condition, we prove that the proposed algorithm can correctly identify the zero and nonzero elements of the sparse parameter vector using a finite number of observations, and further estimates of the nonzero elements almost surely converge to the true values. Compared with the related works, e. g., LASSO, our method only requires the weakest assumptions and does not require solving additional optimization problems. Thirdly, the number of finite observations that guarantee the convergence of the zero-element set of unknown sparse parameters of the Hammerstein system is derived for the first time. Finally, numerical simulations are provided, demonstrating the effectiveness of the proposed method. Since there is no additional optimization problem, i. e., no additional numerical error, the proposed algorithm performs much better than other related algorithms.

Keywords: stochastic dynamic system, sparse parameter identification, the weakest non-persistent excited condition, feedback control system, strong consistency

Classification: 93A10, 93E12, 93E24

1. INTRODUCTION

Parameter estimation or filtering plays a very important role in system identification and control, signal processing, statistical learning, and other fields ([11, 19]). In recent decades, the classical identification and estimation theory has made great progress, and a relatively complete theoretical system has been formed ([9, 10]).

We note that in many practical application scenarios, such as the selection of effective basis functions in Hamiltonian systems, feature selection and robust inference of biomarkers based on omics data in personalized healthcare, channel estimation in

DOI: 10.14736/kyb-2026-2-0286

ultra-wideband communication systems, since many components in the unknown parameter vector have no or negligible contribution to the system (these elements are zero or close to zero), the unknown parameter vector in a large number of systems is high-dimensional and sparse. Naturally, an interesting research direction is sparse parameter identification, i. e., the precise identification of zero and nonzero elements in the unknown parameter vector, in order to reduce model redundancy and obtain a leaner, better performance, and more reliable prediction model.

It is found that the research on sparse parameter identification has made considerable progress. In the field of signal processing, we find that the research of sparse parameter identification is based on compressed sensing (CS), i. e., solving an L_0(the number of nonzero elements) optimization problem with L_2(Euclidean metric) constraints. Since this optimization problem is NP-hard, researchers have developed some sparse signal estimation algorithms by transforming L_0 optimization into a more solvable L_1 convex optimization problem using the norm equivalence principle. For example, by using compressed sensing methods, Kalouptsidis et al. in [12] developed a sparse identification algorithms based on Kalman filtering and Expectation-Maximization, and verified the efficiency of the proposed algorithm via simulations. Candès and Tao et al. in [2, 3, 4] developed a relatively perfect signal reconstruction theory under some prior assumptions of the sparsity of the model. See [5, 7, 13] for more references. The variable selection problem plays a very important role in the field of statistical learning. The well-known LASSO (the least absolute shrinkage and selection operator) is a linear regression method with L_1 regularization, and sparse identification algorithms represented by it and its variants have been extensively studied. Tibshirani in [17] first proposed the LASSO in 1996. Zhao and Yu in [22] established the model selection consistency under regularity conditions. Zou in [25] first proposed the adaptive lasso and discussed the advantages of the adaptive lasso in detail. Combined with the ubiquitous sparsity in practical systems, we find that CS and adaptive lasso methodology are inherited and developed in the field of system identification and control, and the sparse parameter identification theory of stochastic dynamic systems is established. For example, Satheesh and Arun in [15] proposed a CS-based iterative basis pursuit denoising (BPDN) algorithm to estimate the unknown sparse parameters in the ARMAX model, where the number of zero elements of unknown parameters is known. Roland et al. in [18] studied the consistency of estimation for sparse linear regression models based on CS under the condition that the input signals are independent and identically distributed (i.i.d.). Considering that the regressors in the stochastic dynamic system are often generated by the past input and output signals, so the independence assumption about the observation data is difficult to be satisfied. Based on this, Xie and Guo in [20] developed a compressed consensus normalized least mean squares and provided the stability analysis of the algorithm based on CS theory, where the strong statistical assumptions such as independency are not required, but the prior assumption of the sparsity of the model is. Zhao et al. in [24] proposed a LASSO-based adaptive sparse estimation algorithm with general observation sequences. The zero elements in the unknown parameter are correctly identified with a finite number of observations and a non-persistent excited condition, which is the weakest excited condition in the existing literature about sparse parameter identification. There is also related research in some other fields, such as pruning in deep learning [1, 21], etc.

Summarizing the existing research, we find that there are still some shortcomings in the sparse parameter identification theory. Firstly, the sparse estimation algorithm based on CS theory requires prior knowledge about the sparsity of the unknown parameter and the regression vectors. Secondly, the theoretical analysis of the LASSO-based sparse estimation algorithm requires a regularity condition of the regressor, which is a persistent excited condition actually and difficult or almost impossible to satisfy in stochastic feedback systems. Even the non-persistent excited condition for regressor in [24] is much stronger than the weakest excited condition proposed by Lai and Wei in [14] (See subsection 3.2 and Table 3.2 for detail). Thirdly, for a specific system, such as the Hammerstein system, it is very instructive to give a definite number of finite observations under further assumptions about the regressor, but this cannot be done in [24].

Our work addresses these gaps through three key contributions:

1. **Efficient Algorithm Design**: Existing sparse estimation methods (e. g., LASSO) generate sparse solutions by optimizing criterion functions with penalty terms, which introduces computational complexity. To overcome this, we propose a threshold-based algorithm that produces sparse estimates by adaptively adjusting thresholds applied to least squares results. By eliminating the need for optimization, our approach achieves greater conciseness and efficiency, making it more suitable for real-time applications.

2. **Improved Convergence Guarantees**: Classical least squares theory focuses on asymptotic properties [14], while existing sparse identification methods (e. g., [24]) require strong persistent excitation conditions. Under much weaker non-persistent excitation, we prove our algorithm correctly identifies zero/nonzero elements in finite observations , a result parallel to [24] but under far less restrictive conditions. Furthermore, estimates of the nonzero parameters will converge to the true values almost surely with a convergence rate of $O\left(\sqrt{\frac{\log R_N}{\lambda_{\min}^N}}\right)$, i. e., parameter convergence, which, to the best of the authors' knowledge, has never been achieved in sparse parameter identification algorithms for stochastic regression models.

3. **Finite-Observation Bounds for Hammerstein Systems**: While strong convergence results exist for Hammerstein systems [23], no prior work on sparse parameter identification in stochastic dynamic systems has derived explicit finite-observation bounds. Under comparable conditions, we provide such bounds, quantifying the number of observations required for reliable identification—an important advancement for practical system implementation.

The rest of this paper is organized as follows. In Section 2, we give the problem formulation. Section 3 presents the main results of this paper, including the parameter convergence, the set convergence, and the comparison of conditions for consistency of other algorithm and the proposed algorithm in this paper. Section 4 applies the algorithm to the sparse parameter estimation of the Hammerstein system. A simulation example is given in Section 5, and the concluding remarks are made in Section 6.

2. PROBLEM FORMULATION

2.1. Some preliminaries

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be the probability space, ω be an element in Ω, and $\mathbb{E}(\cdot)$ be the expectation operator. Denote $\|\cdot\|$ as the 2-norm of vectors or matrices in this paper. For two positive sequences $\{a_k\}_{k\geq 1}$ and $\{b_k\}_{k\geq 1}$, $a_k = O(b_k)$ means $a_k \leq cb_k$ for $k \geq 1$ and some $c > 0$, while $a_k = o(b_k)$ means $a_k/b_k \to 0$ as $k \to \infty$.

2.2. Sparse identification algorithm

Consider the parameter identification problem of the following discrete-time stochastic regression model,

$$y_{k+1} = \varphi_k^T \theta + w_{k+1}, \quad k \geq 0 \tag{1}$$

where y_{k+1} is the scalar observation or output at time k, $\varphi_k \in \mathbb{R}^r$ is the r-dimensional stochastic regression vector which may be the function of current and past inputs and outputs, $\theta \in \mathbb{R}^r$ is an unknown r-dimensional parameter to be estimated, and w_{k+1} is the stochastic noise sequence.

The above model (1) includes many parameterized systems, such as ARX system and Hammerstein system. We further denote the parameter vector θ and the index set of its zero elements by

$$\begin{aligned} \theta &\triangleq (\theta(1), \ldots, \theta(r))^T \\ H^* &\triangleq \{l \in \{1, \ldots, r\} \mid \theta(l) = 0\}. \end{aligned} \tag{2}$$

The classical identification algorithms, such as the LS algorithm, can generate consistent estimates for the unknown parameters as the number of data goes to infinity. However, due to the existence of system noises and the finite number of observations in practice, it is hard to deduce from such estimates which parameters are exactly zero or not. In the existing literature, L_1 penalty term is added to the criterion function to be optimized to generate sparse estimates so the zero elements in the unknown parameter θ can be identified. However, there are two issues worth thinking about. Firstly, the existence of the penalty term destroys the optimality of LS and makes the estimation inevitably biased, so a stronger excited condition is needed to ensure the consistency of the algorithm. Secondly, solving the optimization problem with the penalty term brings great difficulty. Our problem is to design a new sparse adaptive estimation algorithm that does not require penalty terms to infer the set H^* in a finite number of steps and identify the unknown parameter θ by using stochastic regression vectors and the observation signals $\{\varphi_k, y_{k+1}\}_{k=1}^n$.

To address the limitations of penalty-based methods discussed above, we propose a new sparse adaptive estimation algorithm that avoids L_1 penalty terms while enabling finite-sample identification of the support set H^*.

Our algorithm operates in an iterative framework using the stochastic regression vectors $\{\varphi_k\}$ and observation signals $\{y_{k+1}\}$ from the system. The key idea is to adaptively refine a threshold that distinguishes zero from nonzero parameters in the least squares estimate, eliminating the need for optimization with penalty terms.

The algorithm is presented in Algorithm 1 below.

Algorithm 1 Sparse Identification Algorithm

Step 0:(Initialization) Choose a positive sequence $\{\alpha_n\}$ satisfying $\alpha_n \to 0$ as $n \to \infty$ and

$$\sqrt{\frac{\log R_n}{\lambda_{\min}^n}} = o(\alpha_n),$$

where $R_n = 1 + \sum_{k=0}^n \|\varphi_k\|^2$, $\lambda_{\min}^n = \lambda_{\min}\{P_0^{-1} + \sum_{k=0}^n \varphi_k\varphi_k^T\}$.

Step 1: Based on $\{\varphi_k, y_{k+1}\}_{k=1}^n$, begin with an initial vector θ_0 and an initial matrix $P_0 > 0$, compute the matrix P_{n+1}^{-1} and the estimate θ_{n+1} of θ by the following recursive LS algorithm,

$$P_{k+1} = P_k - a_k P_k \varphi_k \varphi_k^T P_k, \tag{3}$$

$$a_k = \frac{1}{1 + \varphi_k^T P_k \varphi_k}, \tag{4}$$

$$\theta_{k+1} = \theta_k + a_k P_k \varphi_k \big\{y_{k+1} - \varphi_k^T \theta_k\big\}. \tag{5}$$

Step 2: Obtain $\beta_{n+1}(l)$ by the following mechanism,

$$\beta_{n+1}(l) = \begin{cases} 0 & if\ |\theta_{n+1}(l)| < \alpha_{n+1} \\ \theta_{n+1}(l) & else \end{cases}, \tag{6}$$

then obtain

$$\beta_{n+1} = (\beta_{n+1}(1), \ldots, \beta_{n+1}(r))^T \tag{7}$$

$$H_{n+1} \triangleq \{l \in \{1, \ldots, r\} \mid \beta_{n+1}(l) = 0\}. \tag{8}$$

3. THEORETICAL PROPERTIES OF THE SPARSE IDENTIFICATION ALGORITHM

In this section, we will investigate the asymptotic analysis of the unknown sparse parameter vector and the convergence of the sets of zero elements with a finite number of observations, To this end, we introduce the following assumptions to be used for the theoretical analysis.

Assumption 1. The noise $\{w_k, \mathcal{F}_k\}_{k\geq 1}$ is a martingale difference sequence, i. e., $\mathbb{E}[w_{k+1} \mid \mathcal{F}_k] = 0, k \geq 1$, and there exists some $\gamma > 2$ such that $\sup_k \mathbb{E}[|w_{k+1}|^\gamma \mid \mathcal{F}_k] < \infty$ a.s.

Assumption 2. (Non-Excited Condition) The growth rate of $\log\left(\lambda_{\max}\left(P_n^{-1}\right)\right)$ is slower than $\lambda_{\min}\left(P_n^{-1}\right)$.

Recall the previously defined $R_n = 1 + \sum_{k=0}^{n} \|\varphi_k\|^2$. By the *Gram matrix property* (or *construction of* P_n), $\lambda_{\max}\left(P_n^{-1}\right)$ and R_n are of the same order, i. e., there exist constants $0 < c < C$ such that

$$cR_n \leq \lambda_{\max}\left(P_n^{-1}\right) \leq CR_n.$$

Thus, $\log\left(\lambda_{\max}\left(P_n^{-1}\right)\right)$ and $\log R_n$ have equivalent growth rates.

For brevity, let $\lambda_{\min}^n = \lambda_{\min}\left(P_n^{-1}\right)$; the condition is equivalent to:

$$\lim_{n\to\infty} \frac{\log R_n}{\lambda_{\min}^n} = 0 \quad \text{a.s.}$$

Remark 1. A further explanation of this non-excited condition is provided in subsection 3.2.

3.1. Set and parameter convergence of estimates

Assume that there are d (d is unknown) nonzero elements in vector θ. Without loss of generality, we assume $\theta = [\theta(1), \ldots, \theta(d), \theta(d+1), \ldots, \theta(r)]^T$ and $\theta(i) \neq 0, i = 1, \ldots, d, \theta(j) = 0, j = d+1, \ldots, r$. Before giving the main results, we first state a classical result in stochastic adaptive control.

Lemma 1. (Guo [8]) Assume that Assumptions 1 holds. Then as $n \to \infty$, the estimation error of the recursive LS algorithm (3) – (5) is bounded by

$$\|\theta_{n+1} - \theta\|^2 \leq C_0 \frac{\log R_n}{\lambda_{\min}^n} \text{ a.s.},$$

where C_0 is a finite constant.

Remark 2. The specific value of C_0 can be referred to [8].

Remark 3. We remark that the recursive LS algorithm has strong consistency under Assumption 2, i. e., $|\theta_{n+1}(l) - \theta(l)| \to 0$ *a.s.* for all $l = 1, \ldots, r$.

For the estimate β_{n+1} generated by equations (3) – (7), we have the following main results.

Theorem 1. (Parameter Convergence) Assume that Assumptions 1 – 2 hold. Then

$$\beta_{n+1}(l) \underset{n\to\infty}{\longrightarrow} \theta(l), \quad l = 1, \ldots, r \text{ a.s.}$$

P r o o f. By the definition of β_{n+1}, we have for $\forall l \in \{1, \ldots, r\}$

$$\begin{aligned} |\beta_{n+1}(l) - \theta(l)| &\leq |\beta_{n+1}(l) - \theta_{n+1}(l)| + |\theta_{n+1}(l) - \theta(l)| \\ &\leq \alpha_{n+1} + |\theta_{n+1}(l) - \theta(l)|. \end{aligned}$$

Thus the proof is completed by the fact that $\alpha_n \to 0$ as $n \to \infty$ and the convergence results $|\theta_{n+1}(l) - \theta(l)| \to 0$ *a.s.* □

Theorem 2. (Set Convergence) Assume that Assumptions 1 – 2 hold. Then there exists an ω-space Ω_0 with $\mathcal{P}\{\Omega_0\} = 1$ such that for any $\omega \in \Omega_0$, there exists an integer $N_0(\omega)$ such that

$$\beta_{N+1}(d+1) = \ldots = \beta_{N+1}(r) = 0, \quad N \geq N_0(\omega),$$

i. e., $H_{N+1} = H^*$, where H_{N+1} and H^* are defined by (8) and (2) respectively.

P r o o f. Since Assumption 2 holds almost surely, there exists a sample space Ω_0 with $\mathcal{P}\{\Omega_0\} = 1$ such that the assumption is satisfied for every $\omega \in \Omega_0$. In what follows, we analyze the estimate sequence $\{\beta_{n+1}\}$ along a fixed sample path $\omega \in \Omega_0$.

Step 1: Lower bound for nonzero parameters. For any $i \in \{1, \ldots, d\}$ (where $|\theta(i)| > 0$), we decompose the estimation error as:

$$\begin{aligned} |\beta_{n+1}(i)| &\geq |\theta(i)| - |\beta_{n+1}(i) - \theta(i)| \\ &\geq |\theta(i)| - |\beta_{n+1}(i) - \theta_{n+1}(i)| - |\theta_{n+1}(i) - \theta(i)| \\ &\geq |\theta(i)| - \alpha_{n+1} - |\theta_{n+1}(i) - \theta(i)|. \end{aligned}$$

The last inequality follows from the thresholding operation: $|\beta_{n+1}(i) - \theta_{n+1}(i)| \leq \alpha_{n+1}$ by construction.

Since $\alpha_n \to 0$ and $|\theta_{n+1}(i) - \theta(i)| \to 0$ almost surely, there exists $N_1(\omega)$ such that for all $n > N_1$:

$$\alpha_{n+1} + |\theta_{n+1}(i) - \theta(i)| \leq \frac{|\theta(i)|}{2}.$$

Therefore, for all $n > N_1$ and $i \in \{1, \ldots, d\}$:

$$|\beta_{n+1}(i)| \geq \frac{|\theta(i)|}{2} > 0.$$

This ensures that nonzero parameters do not converge to zero.

Step 2: Exact sparsity for zero parameters. Now consider $i \in \{d+1, \ldots, r\}$ (where $\theta(i) = 0$). By Lemma 1, the least-squares estimate satisfies:

$$|\theta_{n+1}(i)| = O\left(\sqrt{\frac{\log R_n}{\lambda_{\min}^n}}\right) \quad \text{a.s.}$$

Using the key property that $\sqrt{\frac{\log R_n}{\lambda_{\min}^n}} = o(\alpha_n)$ (which follows from Assumption 2), there exists $N_2(\omega)$ such that for all $n > N_2$:

$$|\theta_{n+1}(i)| < \alpha_{n+1}.$$

By the thresholding rule in Algorithm 1, this implies $\beta_{n+1}(i) = 0$ for all $n > N_2$ and $i \in \{d+1, \ldots, r\}$.

Conclusion. Taking $N_0(\omega) = \max(N_1, N_2)$, we conclude that for all $n > N_0$:

- $|\beta_{n+1}(i)| > 0$ for $i = 1, \ldots, d$ (nonzero elements remain identified),
- $\beta_{n+1}(i) = 0$ for $i = d+1, \ldots, r$ (zero elements are correctly set to zero).

Hence, the support of β_{n+1} matches that of θ exactly for all sufficiently large n, completing the proof. □

Remark 4. Theorem 2 shows that the index set of the zero elements in θ can be correctly identified with a finite number of observations, and estimates for the nonzero elements will also be nonzero in a finite number of observations.

The following corollary follows immediately from Theorem 2 and Lemma 1.

Corollary 1. Under assumptions of Theorem 2, Algorithm 1 has the following convergence rate as $N \to \infty$:

$$\|\beta_{N+1} - \theta\| = O\left(\sqrt{\frac{\log R_N}{\lambda_{\min}^N}}\right) \quad a.s.$$

To the best of the authors' knowledge, this convergence rate is parallel to that of [8, 14] and has never been achieved in sparse parameter identification algorithms for stochastic regression models. Next, we compare the conditions of observation data that guarantee the convergence of Algorithm 1, LASSO, and its variants respectively.

3.2. Comparison of conditions for consistency of LASSO, adaptive LASSO, the algorithm proposed in [24] and Assumption 2

Method	Conditions on system
LASSO [22]	· Regularity condition: $D_n \to D$ as $n \to \infty$ · Strong irrepresentable condition: for some $\eta > 0$, $\lvert D_n^{21}(D_n^{11})^{-1}\mathrm{sgn}(\theta_1)\rvert \le 1 - \eta$
Adaptive LASSO [25]	· Regularity condition: $D_n \to D$ as $n \to \infty$
Method in [24]	· $\frac{R_n}{\lambda_{\min}^n}\sqrt{\frac{\log R_n}{\lambda_{\min}^n}} \to 0$ as $n \to \infty$
Algorithm 1	· $\frac{\log R_n}{\lambda_{\min}^n} \to 0$ as $n \to \infty$

Tab. 1. Conditions for consistency of LASSO and its variations and Algorithm 1.

For simplicity of notations, we still assume that the parameter vector $\theta = \begin{bmatrix} \theta_1^T & \theta_2^T \end{bmatrix}^T$, $\theta_1 = [\theta(1) \ldots \theta(d)]^T, \theta_2 = [\theta(d+1) \ldots \theta(r)]^T$ such that $\theta(i) \neq 0, i = 1, \ldots, d$ and $\theta(j) = 0, j = d+1, \ldots, r$. Denote

$$D_n \triangleq \frac{1}{n}\sum_{k=1}^{n} \varphi_k \varphi_k^T = \begin{bmatrix} D_n^{11} & D_n^{12} \\ D_n^{21} & D_n^{22} \end{bmatrix}$$

where $D_n^{11} \in \mathbb{R}^{d\times d}$ and D_n^{12}, D_n^{21} and D_n^{22} are with compatible dimensions. The comparison on conditions for consistency of LASSO and its variations as well as Algorithm 1 is made in Table 1. The strong irrepresentable condition given in Table 1 is actually prior structural information on the sparsity of the parameter vector, which is not required in our paper. Furthermore, from Table 1, we can directly verify that Assumption 2 includes the regularity condition as its special case. Besides, The algorithm proposed by [24] requires an excited condition that $R_n\sqrt{\log R_n} = o\left((\lambda_{\min}^n)^{3/2}\right)$, which fails for $R_n = O\left((\lambda_{\min}^n)^{\delta}\right)$, $\forall\delta > 3/2$. However, from Theorem 2 and Corollary 1 we know that Algorithm 1 can achieve the consistency for $R_n = O\left((\lambda_{\min}^n)^{\delta}\right)$, $\forall\delta > 0$, which greatly improves the applicability of the algorithm. As far as the authors know, Assumption 2 is the weakest excited condition compared with the one required for sparse parameter identification in the existing literature.

4. PRACTICAL APPLICATIONS

Many applications of the sparse identification algorithm have been shown in [24], such as the identification of Hammerstein systems [6] and linear stochastic systems with self-tuning regulation control [8]. However, [24] only gave the existence of the constant $N_0(\omega)$ as in Theorem 2, but did not give its specific range, which is not conducive to the development of practical applications. In this section, we take the identification of Hammerstein systems as an example, and provide the specific range of the constant $N_0(\omega)$.

Hammerstein system is a modular nonlinear dynamic system with a static nonlinear function followed by a linear dynamic subsystem. Due to its simple structure, as well as a good simulation of nonlinear, Hammerstein system has been widely used in many practical engineering scenarios, such as power amplifiers, manipulators, as well as the neutralization reaction in chemical processes, lithium-ion battery heating systems and heat exchange system, (cf., [6, 16]), etc.

Considering a Hammerstein system whose linear subsystem is an ARX system and whose nonlinear function is a combination of basis functions:

$$y_{k+1} = a_1 z y_{k+1} + \ldots + a_p z^p y_{k+1} + b_1 f(u_k) + \ldots + b_q f\left(z^{q-1}u_k\right) + w_{k+1}, \tag{9}$$

$$f(u_k) = \sum_{j=1}^{m} c_j g_j(u_k) \tag{10}$$

where $\{g_j(\cdot)\}_{j=1}^m$ are the basis functions, z is the backward-shift operator (i. e., $zy_{k+1} = y_k$), y_{k+1} and u_k are the output and input, w_{k+1} is the system noise, $a_1, \ldots, a_p, b_1, \ldots, b_q, c_1, \ldots, c_m$ are unknown coefficients that need to be estimated. Denote

$$\begin{aligned}\theta &\triangleq (a_1, \ldots, a_p, (b_1c_1), \ldots, (b_1c_m), \ldots, (b_qc_1), \ldots, (b_qc_m))^T\\ \varphi_k &\triangleq (y_k, \ldots, y_{k+1-p}, g_1(u_k), \ldots, g_m(u_k), \ldots,\\ &\qquad g_1(u_{k+1-q}), \ldots, g_m(u_{k+1-q}))^T,\end{aligned}$$

the Hammerstein system can be written in the form of (1). Specifically, if $c_l = 0$ (i. e., the lth basis function $g_l(\cdot)$ is non-effective), then all components $(b_1c_l), (b_2c_l), \ldots, (b_qc_l)$ corresponding to c_l in the parameter vector θ are also zero.

In order to obtain a good approximation of nonlinear functions $f(\cdot)$, a large number of basis functions (which leads to a very large m) are often used in practice, which will lead to the unknown parameter vector θ is probably to be high dimensional and sparse in practice. To obtain a simpler but more precise model of the system, we find that it is crucial to determine the number of the effective basis functions in $\{g_j(\cdot)\}_{j=1}^m$, or sparse identification of the unknown parameter vector θ. Thus Algorithm 1 can be applied to infer the zero elements in θ. Denote

$$M = \left[\ M(1), \ldots, M(m)\ \right]$$

with $M(l) = [b_1 c_l, \ldots, b_q c_l]^T$, $l = 1, \ldots, m$. Thus we can find that the noneffective basis functions in $\{g_j(\cdot)\}_{j=1}^m$ correspond to zero columns in matrix M. Consequently, executing algorithms (3) – (5) by using the inputs and outputs $\{u_k, y_{k+1}\}_{k=1}^n$ of Hammerstein system (9), we can obtain θ_{n+1}, followed by executing (6) – (8), we have

$$\begin{aligned}
\beta_{n+1} = &\left[a_{1,n+1}, \ldots, a_{p,n+1}, (b_1 c_1)_{n+1}, \ldots, (b_1 c_m)_{n+1}, \right. \\
&\left. \ldots, (b_q c_1)_{n+1}, \ldots, (b_q c_m)_{n+1}\right]^T, \qquad (11)\\
M_{n+1} = &\left[M_{n+1}(1), \ldots, M_{n+1}(m)\right], \qquad (12)\\
\text{with } M_{n+1}(l) = &\left[(b_1 c_l)_{n+1}, \ldots, (b_q c_l)_{n+1}\right]^T, l = 1, \ldots, m,\\
H^* = &\left\{l \in \{1, \ldots, m\} \mid c_l = 0\right\}, \qquad (13)\\
H_{n+1} = &\left\{l = \{1, \ldots, m\} \mid M_{n+1}(l) = 0\right\}. \qquad (14)
\end{aligned}$$

Since $M_{n+1}(l)$ in (12) is the subvector of β_{n+1} corresponding to c_l, we have $M_{n+1}(l) = 0 \iff (c_l)_{n+1} = 0$. Before presenting the main results, we need the following assumptions.

Assumption 3. $\{1, g_1(x), \ldots, g_m(x)\}$ is linearly independent over some interval $[a, b]$.

Assumption 4. Polynomial $A(z) = 1 - a_1 z - \ldots - a_p z^p$ is stable, i. e., $|A(z)| \neq 0, \forall |z| \leq 1$ and $b_1^2 + \ldots + b_q^2 > 0$.

Assumption 5. $\{u_k\}_{k \geq 1}$ is an i.i.d. sequence with density $p(x)$ which is positive and continuous on $[a, b]$ and $0 < \mathbb{E} g_j^2(u_k) < \infty, j = 1, \ldots, m$. Further, $\{u_k\}_{k \geq 1}$ and $\{w_k\}_{k \geq 1}$ are mutually independent.

Remark 5. Assumptions 3 – 5 are essential to ensure convergence of parameter estimates in the Hammerstein system, as described in detail in [23]. Assumption 5 is typically satisfied in feedforward identification setups. In feedback systems, where $u(t)$ may depend on past noise values, this assumption may not hold.

Lemma 2. (Zhao [23]) If Assumptions 1 and 3 – 5 hold, then for the maximal and minimal eigenvalues of $\sum_{k=1}^n \varphi_k \varphi_k^T$, the following inequalities

$$C_1 n \leq R_n \leq C_2 n, \text{ a.s.} \qquad (15)$$

$$C_3 n \leq \lambda_{\min}^n \leq C_4 n, \text{ a.s.} \tag{16}$$

hold for some $0 < C_1 < C_2, 0 < C_3 < C_4$. Further, the LS estimate θ_{n+1} has the following convergence speed,

$$\|\theta_{n+1} - \theta\| = O\left(\sqrt{\frac{\log n}{n}}\right) \text{ a.s.}$$

Proposition 1. Set $\alpha_n = \{\frac{\log n}{n}\}^\epsilon$ for any fixed $\epsilon \in \left(0, \frac{1}{2}\right)$. If Assumptions 1 and 3 – 5 hold for Hammerstein system (9) – (10), then there exists an ω-set Ω_0 with $\mathcal{P}\{\Omega_0\} = 1$ such that for any $\omega \in \Omega_0$,

$$H_{N+1} = H^* \quad \forall N \geq N_0(\omega)$$

holds for $N_0(\omega)$ being an positive integer, i. e., the effective basis functions in $\{g_j(\cdot)\}_{j=1}^m$ can be correctly identified.

Proof. By (15) – (16) and noticing $\alpha_n = \{\frac{\log n}{n}\}^\epsilon, \epsilon \in \left(0, \frac{1}{2}\right)$, we can verify that Assumptions 1 – 2 hold for the regression model composed of (9) – (10), thus, the desired results followed by Theorem 2 directly. □

To give a specific selection method of the positive integer $N_0(\omega)$, we need the following assumption.

Assumption 6. $|\theta(i)| > C_5 > 0$ for $i = 1, \ldots, d$.

Remark 6. This assumption is necessary and natural, and if $|\theta(i)| > C_5$ $i = 1, \ldots, d$ is very close to 0, the required constant $N_0(\omega)$ is inevitably going to be very large.

Without loss of generality, set the threshold value α_n in the form of $\alpha_n = M\left(\frac{R_n}{\lambda_{\min}^n}\right)^\epsilon$ with $M > 0$ and $0 < \epsilon < \frac{1}{2}$.

Theorem 3. Under Assumptions 1 – 6, there exists an ω-set Ω_0 with $\mathcal{P}\{\Omega_0\} = 1$ such that for any $\omega \in \Omega_0$,

$$H_{N+1} = H^* \quad \forall N > N_0 \tag{17}$$

holds, where

$$\begin{aligned} N_0 &= \max\left\{\frac{47}{C_2}, \frac{2k_1}{C_3}\log\left(\frac{C_2 k_1}{C_3}\right), \frac{2k_2}{C_3}\log\left(\frac{C_2 k_2}{C_3}\right)\right\}, \\ k_1 &= \left(\frac{\sqrt{C_0}}{M}\right)^{\frac{2}{1-2\epsilon}}, \quad k_2 = \left(\frac{2M}{C_5}\right)^{\frac{1}{\epsilon}}. \end{aligned} \tag{18}$$

The proof of this theorem is technical and lengthy; to maintain the flow of the presentation, we have deferred it to Appendix.

In Theorem 3, the constant N_0 is corresponding to k_1 and k_2. To quickly identify the position of the zero elements, we need to make $\max\{k_1, k_2\}$ as small as possible.

Clearly, for fixed ϵ, k_1 decreases as M increases, while k_2 increases as M increases. A straightforward calculation shows that

$$\begin{aligned} &\max\{k_1, k_2\} \geq k_1 \geq \frac{4C_0}{C_5^2}, \quad \text{if } M \leq 2^{2\epsilon-1} C_0^{\epsilon} C_5^{1-2\epsilon}, \\ &\max\{k_1, k_2\} \geq k_2 \geq \frac{4C_0}{C_5^2}, \quad \text{if } M \geq 2^{2\epsilon-1} C_0^{\epsilon} C_5^{1-2\epsilon}. \end{aligned} \tag{19}$$

Thus the lower bound of $\max\{k_1, k_2\}$ is $4C_0C_5^{-2}$, which motivates the following corollary.

Corollary 2. Under conditions of Theorem 3, with $M = 2^{2\epsilon-1} C_0^{\epsilon} C_5^{1-2\epsilon}$, we have

$$H_{N+1} = H^*, \ \forall N > \max\left\{ \frac{47}{C_2}, \frac{8C_0}{C_3C_5^2} \log\left(\frac{4C_2C_0}{C_3C_5^2} \right) \right\}. \tag{20}$$

P r o o f. The proof is a combination of (19) and Theorem 3. □

Remark 7. Noticing that $M \triangleq [b_1 \dots b_q]^T \begin{bmatrix} c_1 & \dots & c_m \end{bmatrix}$, we can further infer the estimates for the nonzero elements in $\{b_i, i = 1, \dots, q\}$ and $\{d_l, l = 1, \dots, m\}$ by using a singular value decomposition (SVD) algorithm to M_{N+1} defined by (12), see Chaoui et al. (2005) for more details.

5. SIMULATION RESULTS

In this section, we provide an example to illustrate the performance of the novel sparse identification algorithm (i. e., Algorithm 1) proposed in this paper.

Example 1. Consider a discrete-time stochastic regression model (1) with the dimension $r = 10$. The noise sequence $\{w_{k+1}\}, k \geq 0$ in (1) is independent and identically distributed with $w_{k+1} \sim \mathcal{N}(0, 0.1)$ (Gaussian distribution with zero mean and variance 0.1). Let the regression vector $\varphi_k \in \mathbb{R}^r (k \geq 1)$ be generated by the following state space model,

$$\begin{aligned} x_k &= Ax_{k-1} + \varepsilon_k, \\ \varphi_k &= B_k x_k, \end{aligned}$$

where $x_k \in \mathbb{R}^r$ is the state of the above system with $x_0 = [\underbrace{1, \dots, 1}_{r}]^T$, the matrices A, B_k and vector ε_k are chosen according to the following way,

$$\begin{aligned} A_k &= \text{diag}\{\underbrace{1.01, \dots, 1.01}_{r}\}, \\ B_k &= \{b_{k,ij}\} \in \mathbb{R}^{r \times r} \text{ with } b_{k,ij} \sim \mathcal{N}(0, 1), \\ \varepsilon_k &= \{\varepsilon_{k,i}\} \in \mathbb{R}^r \text{ with } \varepsilon_{k,i} \sim \mathcal{N}(0, 1). \end{aligned}$$

The true value of the unknown parameter is

$$\begin{aligned} \theta &= [\theta(1), \theta(2), \theta(3), \theta(4), \theta(5), \theta(6), \theta(7), \theta(8), \theta(9), \theta(10)]^T \\ &= [0.8, 1.6, -0.3, 0.05, 0, 0, 0, 0, 0, 0]^T. \end{aligned}$$

We take $\alpha_n = 0.1\left(\frac{\log R_n}{\lambda_{\min}^n}\right)^{\frac{1}{4}}$ for Algorithm 1, $\lambda_n = n^{0.75}$ for LASSO (see [22]), and $\lambda_n = (\lambda_{\min}^n)^{0.75}$ for the algorithm proposed by [24]. Ten simulations are performed. Figure 1 shows the estimate sequences

$$\{\theta_n(1), \theta_n(2), \ldots, \theta_n(10)\}_{n=1}^{600}$$

generated by least squares, LASSO, the algorithm proposed by [24] and Algorithm 1 from one of the simulations. Table 2 compares the averaged estimates from the ten simulations generated by these methods for $\theta(5), \theta(6), \theta(7), \theta(8), \theta(9), \theta(10)$, with different data length n. We adopt the Python CVX tools (http://cvxopt.org/) to solve the convex optimization in LASSO ([22]) and the algorithm proposed by [24].

From Figure 1 and Table 2, we can find that, compared with the least squares estimates, the LASSO ([22]), the algorithm proposed by [24] and Algorithm 1 generate sparser and more accurate estimates for the system parameters, naturally, give us valuable information in inferring the zero and nonzero elements in the unknown parameters. Moreover, compared with LASSO ([22]) and the algorithm proposed by [24], Algorithm 1 does not solve the additional optimization problem. Thus it can avoid many unnecessary numerical errors and make the estimates of true zeros exactly zero.

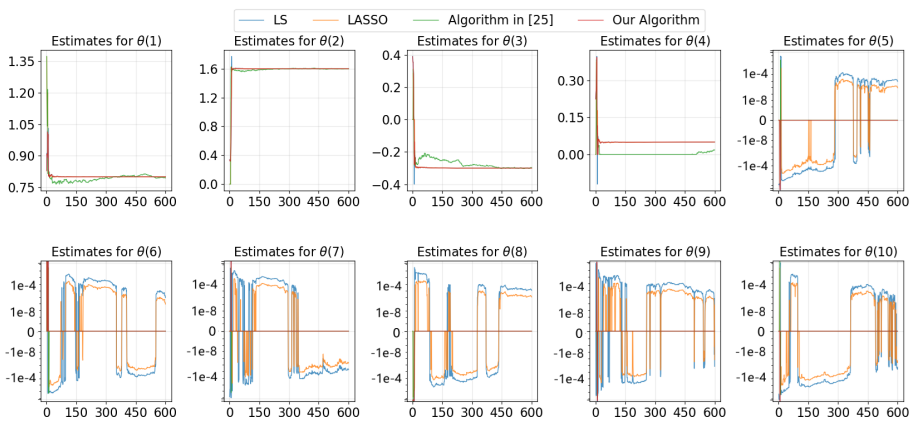

Fig. 1. Estimate sequences in example 1.

Example 2. Consider the Hammerstein system,

$$y_{k+1} + a_1 y_k + a_2 y_{k-1} = b_1 f(u_k) + b_2 f(u_{k-1}) + w_{k+1},$$

where $a_1 = -1.5$, $a_2 = 0.56$, $b_1 = 1$, $b_2 = -2$ and $f(u) = \sum_{j=1}^{4} d_j u^j$ is a 4th polynomial with $d_1 = 1$, $d_3 = 0.2$ and $d_2 = d_4 = 0$. Denote

$$\theta = [-a_1,\ -a_2,\ (b_1 d_1),\ \ldots,\ (b_1 d_4),\ (b_2 d_1),\ \ldots,\ (b_2 d_4)]^T,$$
$$\varphi_k = \left[y_k,\ y_{k-1},\ u_k,\ \ldots,\ u_k^4,\ u_{k-1},\ \ldots,\ u_{k-1}^4\right]^T.$$

	N=100	N=200	N=300	N=400	N=500
Estimates for $\theta(5)$					
By Algorithm 1	0	0	0	0	0
By least squares	2.0754×10^{-4}	1.1548×10^{-4}	2.6044×10^{-5}	7.4872×10^{-6}	5.4098×10^{-6}
By LASSO	1.5651×10^{-5}	1.1055×10^{-5}	2.6356×10^{-6}	7.3534×10^{-7}	5.4774×10^{-7}
By method in [24]	1.1805×10^{-12}	1.2767×10^{-17}	1.0867×10^{-18}	2.1817×10^{-20}	1.2387×10^{-21}
Estimates for $\theta(6)$					
By Algorithm 1	0	0	0	0	0
By least squares	1.7753×10^{-4}	3.6510×10^{-5}	4.7818×10^{-6}	5.8922×10^{-6}	1.2250×10^{-6}
By LASSO	1.2038×10^{-5}	2.9274×10^{-6}	4.0088×10^{-7}	5.9536×10^{-7}	1.2497×10^{-7}
By method in [24]	7.2487×10^{-12}	8.0211×10^{-18}	2.8565×10^{-18}	2.3326×10^{-20}	3.0134×10^{-22}
Estimates for $\theta(7)$					
By Algorithm 1	0	0	0	0	0
By least squares	7.1995×10^{-4}	4.6556×10^{-5}	2.1246×10^{-5}	9.3689×10^{-6}	2.1212×10^{-6}
By LASSO	6.4569×10^{-5}	3.7443×10^{-6}	2.1639×10^{-6}	9.4968×10^{-7}	2.1478×10^{-7}
By method in [24]	7.9009×10^{-12}	2.0456×10^{-18}	7.1981×10^{-19}	2.6872×10^{-20}	1.4004×10^{-21}
Estimates for $\theta(8)$					
By Algorithm 1	0	0	0	0	0
By least squares	6.2161×10^{-4}	3.2298×10^{-5}	1.3463×10^{-5}	1.2123×10^{-5}	1.7761×10^{-6}
By LASSO	5.4481×10^{-5}	2.7948×10^{-6}	1.2180×10^{-6}	1.2338×10^{-6}	1.7343×10^{-7}
By method in [24]	3.3657×10^{-12}	3.4547×10^{-18}	2.5375×10^{-19}	1.6296×10^{-20}	8.1807×10^{-22}
Estimates for $\theta(9)$					
By Algorithm 1	0	0	0	0	0
By least squares	1.6405×10^{-4}	4.8687×10^{-5}	1.7557×10^{-5}	1.2013×10^{-5}	2.1303×10^{-6}
By LASSO	1.5557×10^{-5}	3.9550×10^{-6}	1.7270×10^{-6}	1.1999×10^{-6}	2.1130×10^{-7}
By method in [24]	4.8125×10^{-12}	4.7665×10^{-18}	7.8394×10^{-19}	1.5668×10^{-20}	4.5756×10^{-22}
Estimates for $\theta(10)$					
By Algorithm 1	0	0	0	0	0
By least squares	3.4718×10^{-4}	4.6033×10^{-5}	1.8614×10^{-5}	3.5427×10^{-6}	1.6878×10^{-6}
By LASSO	2.8112×10^{-5}	3.6159×10^{-6}	1.7057×10^{-6}	3.5963×10^{-7}	1.8325×10^{-7}
By method in [24]	3.7684×10^{-12}	4.4883×10^{-18}	9.8752×10^{-19}	5.9050×10^{-20}	2.8450×10^{-21}

Tab. 2. Averaged estimates by our method, least squares and LASSO from ten simulations in example 1.

It is to directly verify that the Hammerstein system can be formulated by $y_{k+1} = \theta^T\varphi_k + w_{k+1}$ and the following equality takes place

$$\begin{bmatrix} a_1 & b_1d_1 & b_1d_2 & b_1d_3 & b_1d_4 \\ a_2 & b_2d_1 & b_2d_2 & b_2d_3 & b_2d_4 \end{bmatrix} = \begin{bmatrix} -1.5 & 1 & 0 & 0.2 & 0 \\ 0.56 & -2 & 0 & -0.4 & 0 \end{bmatrix}.$$

For identification of the Hammerstein system, we select the input $\{u_k\}$ as an i.i.d. sequence that is uniformly distributed over $[-5, 5]$. We assume that the noise sequence $\{w_k\}$ is i.i.d. with Gaussian distribution $\mathcal{N}(0, 1)$ and independent of $\{u_k\}$.

Figure 2 shows the estimate sequences

$$\{a_1, a_2, b_1d_1, \ldots, b_1d_4, b_2d_1, \ldots, b_2d_4\}_{n=1}^{600}$$

generated by least squares, LASSO with $\lambda_n = n^{0.75}$, the algorithm proposed by [24] with $\lambda_n = (\lambda_{\min}^n)^{0.75}$ and Algorithm 1 with $\alpha_n = 0.1\left(\frac{\log R_n}{\lambda_{\min}^n}\right)^{\frac{1}{4}}$ from one simulation. The results presented in Figure 2 further validate the superiority of our algorithm.

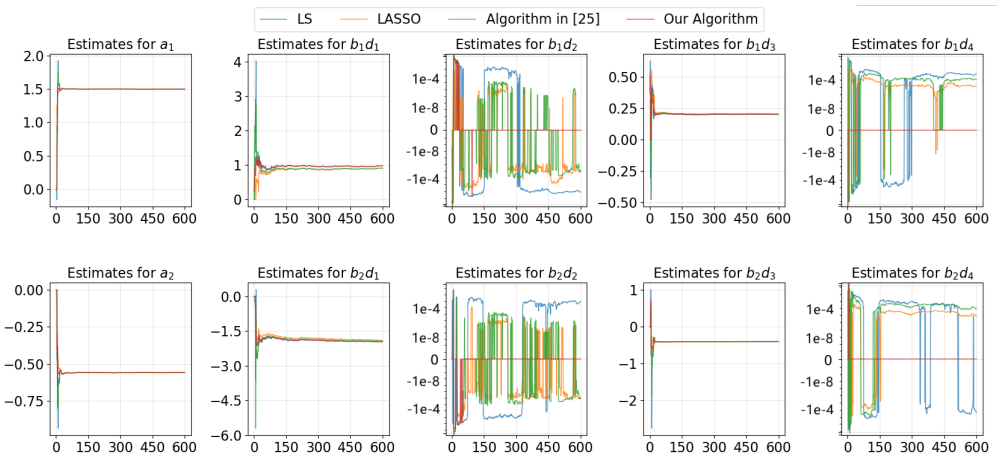

Fig. 2. Estimate sequences in example 2.

To evaluate the efficiency of Algorithm 1 in identifying zero elements within the Hammerstein system, we define the precise identification step, denoted by N. That is, for iterations after N, the estimates for parameters $a_1, a_2, b_1d_1, b_1d_3, b_2d_1, b_2d_3$ are non-zero, while the estimates for $b_1d_2, b_1d_4, b_2d_2, b_2d_4$ are strictly zero. Table 3 presents the average number of iteration step N required for exact identification across 50 independent trials under different values of α_n.

As shown in Table 3, selecting an appropriately sized α_n is crucial for the algorithm to achieve precise identification rapidly. An excessively large α_n may lead to over-sparsification, incorrectly forcing true non-zero parameters to zero. Conversely, if α_n is too small, the sparsity-promoting effect is weakened, resulting in a slower convergence of the true zero parameters to their exact values.

6. CONCLUSION

In this paper, the identification of sparse unknown parameters in stochastic regression models is studied. We proposed an efficient algorithm for generating sparse estimates without any penalty terms. Under the weakest non-persistent excited condition, we proved that the set of zero elements in the unknown sparse parameter vector could be correctly identified in finite observations, and the non-zero elements almost surely converge to the true values. For the Hammerstein system, we give the determined number of finite observations. Some meaningful problems deserve further consideration, e. g., the application of the proposed algorithm to identify time-varying unknown sparse parameters. The distributed algorithm to estimate the unknown parameter using local measurement, the adaptive control by using the estimation algorithm based on the sampled data, and the continuing execution of the algorithm when the set of zero elements is identified in finite observations, i. e., the influence of the initial value on the algorithm.

The value of α_n	Average precise identification step N
$0.1\left(\frac{\log R_n}{\lambda_{\min}^n}\right)^{\frac{1}{4}}$	171.76
$0.2\left(\frac{\log R_n}{\lambda_{\min}^n}\right)^{\frac{1}{4}}$	34.06
$0.3\left(\frac{\log R_n}{\lambda_{\min}^n}\right)^{\frac{1}{4}}$	76.64
$0.5\left(\frac{\log R_n}{\lambda_{\min}^n}\right)^{\frac{1}{4}}$	221.78
$0.5\left(\frac{\log R_n}{\lambda_{\min}^n}\right)^{\frac{1}{4}}$	539.76
$0.1\left(\frac{\log R_n}{\lambda_{\min}^n}\right)^{\frac{1}{3}}$	305.18
$0.2\left(\frac{\log R_n}{\lambda_{\min}^n}\right)^{\frac{1}{3}}$	42.54
$0.3\left(\frac{\log R_n}{\lambda_{\min}^n}\right)^{\frac{1}{3}}$	56.36
$0.4\left(\frac{\log R_n}{\lambda_{\min}^n}\right)^{\frac{1}{3}}$	117.56
$0.5\left(\frac{\log R_n}{\lambda_{\min}^n}\right)^{\frac{1}{3}}$	222.56

Tab. 3. The average precise identification step N for Algorithm 1 from 50 simulations under different α_n.

7. APPENDIX

Proof of Theorem 3.

P r o o f . From the proof of Theorem 2 we know that, the following two formulas are sufficient conditions for $H_{N+1} = H^*$:

$$|\theta_{N+1}(i) - \theta(i)| + \alpha_{N+1} < |\theta(i)|, \quad i \in \{1, \ldots, d\}, \tag{21}$$

$$|\theta_{N+1}(i)| < \alpha_{N+1}, \quad i \in \{d+1, \ldots, r\}. \tag{22}$$

Thus, if we can verify (21) and (22) under the above N, then the desired result (17) is true naturally. We first provide an inequality as follows,

$$\frac{\log N}{N} < \frac{1}{t}, \quad \forall N > \max\{47, 2t \log t\},\ t > 0.$$

In fact, for $0 < t < 10$, we have

$$\frac{\log N}{N} < \frac{\log 47}{47} < \frac{1}{10} < \frac{1}{t}.$$

For $t \geq 10$, we have

$$\frac{\log N}{N} < \frac{\log t + \log 2 + \log\log t}{2t\log t} < \frac{2\log t}{2t\log t} = \frac{1}{t}.$$

On one hand, by (17) and (18) we have

$$C_2 N > \max\left\{47, \frac{2C_2k_1}{C_3}\log\left(\frac{C_2k_1}{C_3}\right)\right\},$$

thus $\frac{\log(C_2N)}{C_2N} < \frac{C_3}{C_2}\left(\frac{M}{\sqrt{C_0}}\right)^{\frac{2}{1-2\epsilon}}$, that is $\frac{\log(C_2N)}{C_3N} < \left(\frac{M}{\sqrt{C_0}}\right)^{\frac{2}{1-2\epsilon}}$.

Now by Lemma 1 and (15) as well as (16) we have

$$\begin{aligned}\frac{|\theta_{N+1}(i)-\theta(i)|}{\alpha_{N+1}} &\leq \frac{\sqrt{C_0}}{M}\left(\frac{\log(C_2N)}{C_3N}\right)^{\frac{1}{2}-\epsilon} \\ &< \frac{\sqrt{C_0}}{M}\left(\frac{M}{\sqrt{C_0}}\right)^{\frac{1-2\epsilon}{1-2\epsilon}} \\ &= 1.\end{aligned} \tag{23}$$

Thus (22) is satisfied by the fact that $\theta(i) = 0$ for $i = d+1, \ldots, r$. On the other hand, by

$$C_2 N > \max\{47, \frac{2C_2k_2}{C_3}\log\left(\frac{C_2k_2}{C_3}\right)\},$$

we have $\frac{\log(C_2N)}{C_3N} < \left(\frac{C_5}{2M}\right)^{\frac{1}{\epsilon}}$. Then by (15) we have,

$$\begin{aligned}2\alpha_{N+1} &\leq 2M\left(\frac{\log(C_2N)}{C_3N}\right)^{\epsilon} \\ &< 2M\left(\frac{C_5}{2M}\right)^{\frac{\epsilon}{\epsilon}} \\ &= C_5.\end{aligned} \tag{24}$$

(21) is satisfied by combining (23) and (24). This completes the proof of the theorem. □

(Received March 24, 2025)

REFERENCES

[1] O. A. Alzubi, J. A. Alzubi, M. Alweshah, I. Qiqieh, S. Al-Shami, and M. Ramachandran: An optimal pruning algorithm of classifier ensembles: dynamic programming approach. Neural Computing Appl. *32* (2020), 20, 16091–16107. DOI:10.1007/s00521-020-04761-6

[2] E. J. Candès et al.: Compressive sampling. In: Proc. International congress of mathematicians, volume 3, Citeseer 2006, pp. 1433–1452.

[3] E. J. Candès, J. K. Romberg, and T. Tao: Stable signal recovery from incomplete and inaccurate measurements. Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences *59* (2006), 8, 1207–1223.

[4] E. J. Candes and T. Tao: Decoding by linear programming. IEEE Trans. Inform. Theory *51* (2005), 12, 4203–4215. DOI:10.1109/TIT.2005.858979

[5] Y. Chen, Y. Gu, and A. O. Hero: Sparse lms for system identification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE 2009, pp. 3125–3128.

[6] E. Eskinat, S. H. Johnson, and W. L. Luyben: Use of hammerstein models in identification of nonlinear systems. AIChE J. *37* (1991), 2, 255–268. DOI:10.1002/aic.690370211

[7] Y. Gu, J. Jin, and S. Mei: l_0 norm constraint LMS algorithm for sparse system identification. IEEE Signal Process. Lett. *16* (2009), 9, 774–777. DOI:10.1109/LSP.2009.2024736

[8] L. Guo: Convergence and logarithm laws of self-tuning regulators. Automatica *31* (1995), 3, 435–450. DOI:10.1016/0005-1098(94)00127-5

[9] L. Guo: Time-varying Stochastic Systems, Stability and Adaptive Theory. Second edition. Science Press, Beijing 2020.

[10] L. Guo and H. F. Chen: Identification and Stochastic Adaptive Control. Springer Science, Boston 1991.

[11] S. Haykin: Radar signal processing. IEEE Signal Process. Magazine *2* (1985), 2, 2–18. DOI:10.1109/MASSP.1985.1163737

[12] N. Kalouptsidis, G. Mileounis, B. Babadi, and V. Tarokh: Adaptive algorithms for sparse system identification. Signal Process. *91* (2011), 8, 1910–1919. DOI:10.1016/j.sigpro.2011.02.013

[13] Y. Kopsinis, K. Slavakis, and S. Theodoridis: Online sparse system identification and signal reconstruction using projections onto weighted l_1 balls. IEEE Trans. Signal Process. *59* (2010), 3, 936–952. DOI:10.1109/TSP.2010.2090874

[14] T. L. Lai and Ch. Z. Wei: Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. Annals Statist. *10* (1982), 1, 154–166.

[15] S. K. Perepu and A. K. Tangirala: Identification of equation error models from small samples using compressed sensing techniques. In: 9th IFAC Symposium on Advanced Control of Chemical Processes ADCHEM 2015, IFAC-PapersOnLine *48* (2015), 8, 795–800. DOI:10.1016/j.ifacol.2015.09.066

[16] J. G. Smith, S. Kamat, and K. P. Madhavan: Modeling of ph process using wavenet based hammerstein model. J. Process Control *17* (2007), 6, 551–561. DOI:10.1016/j.jprocont.2006.11.001

[17] R. Tibshirani: Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc.: Series B: Methodological *58* (1996), 1, 267–288. DOI:10.1111/j.2517-6161.1996.tb02080.x

[18] R. Tóth, B. M. Sanandaji, K. Poolla, and T. L. Vincent: Compressive system identification in the linear time-invariant framework. In: 50th IEEE Conference on Decision and Control and European Control Conference 2011, pp. 783–790.

[19] M. Vaezi and A. Izadian: Piecewise affine system identification of a hydraulic wind power transfer system. IEEE Trans. Control Systems Technol. *23* (2015), 6, 2077–2086. DOI:10.1109/TCST.2015.2398311

[20] S. Xie and L. Guo: Analysis of compressed distributed adaptive filters. Automatica *112* (2020), 108707. DOI:10.1016/j.automatica.2019.108707

[21] X. Xu, B. Pan, Z. Chen, Z. Shi, and T. Li: Simultaneously multiobjective sparse unmixing and library pruning for hyperspectral imagery. IEEE Trans. Geosci. Remote Sensing *59* (2020), 4, 3383–3395. DOI:10.1109/TGRS.2020.3016941

[22] P. Zhao and B. Yu: On model selection consistency of lasso. J. Machine Learn. Res. *7* (2006), 2541–2563.

[23] W.-X. Zhao: Parametric identification of hammerstein systems with consistency results using stochastic inputs. IEEE Trans. Automat. Control *55* (2010), 2, 474–480, 2010. DOI:10.1109/TAC.2009.2036380

[24] W. Zhao, G. Yin, and E.-W. Bai: Sparse system identification for stochastic systems with general observation sequences. Automatica *121* (2020), 109162. DOI:10.1016/j.automatica.2020.109162

[25] H. Zou: The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. *101* (2006), 476, 1418–1429. DOI:10.1198/016214506000000735

Ziming Wang, Beijing Empyrean Technology Co., Ltd., Beijing 100102. P. R. China.
e-mail: wangziming@lsec.cc.ac.cn

Mingkun Li, China Academy of Aerospace Science and Innovation, Beijing, 100176. P. R. China.
e-mail: mingkunli.96@gmail.com

Yiming Xing, School of Mathematical Sciences, Tongji University, Shanghai 200092. P. R. China.
e-mail: yimingx4@tongji.edu.cn

Xinghua Zhu, Corresponding author. China Energy Electrical Power Marketing Center Co., Ltd., Beijing, 100011. P. R. China.
e-mail: zxh@amss.ac.cn