Infinite-Dimensional Sparse Learning in Linear System Identification

Mingzhou Yin, Mehmet Tolga Akan, Andrea Iannelli, and Roy S. Smith

Abstract— **Regularized methods have been widely applied to system identification problems without known model structures. This paper presents an infinite-dimensional sparse learning algorithm based on atomic norm regularization. Atomic norm regularization decomposes the transfer function into first-order atomic models and solves a group lasso problem that selects a sparse set of poles and identifies the corresponding coefficients. The difficulty in solving the problem lies in the fact that there are an infinite number of possible atomic models. This work proposes a greedy algorithm that generates new candidate atomic models maximizing the violation of the optimality conditions of the existing problem. This algorithm is able to solve the infinite-dimensional group lasso problem with high precision. The algorithm is further extended to reduce the bias and reject false positives in pole location estimation by iteratively reweighted adaptive group lasso and complementary pairs stability selection respectively. Numerical results demonstrate that the proposed algorithm performs better than benchmark parameterized and regularized methods in terms of both impulse response fitting and pole location estimation.**

I. Introduction

System identification investigates the problem of identifying models of dynamical systems from measured input-output data. This problem has been widely studied under the parameter estimation framework, where the system is modeled by a finite-dimensional parametrization [1]. The optimal model parameters can then be estimated by tools in classical statistics. One well-known approach in this category is the prediction error method (PEM) based on maximum likelihood estimation [2].

However, such approaches only work when model structure and complexity are known, and the associated optimization problems are only convex for particular noise models, e.g., ARX models [3]. Alternative approaches have been proposed in the last decade, which identify general high-dimensional models with regularization techniques to encode prior model knowledge [4], [5]. In particular, kernel-based identification [6], [7] has received significant attention, whereby, in its basic form, a truncated impulse response model is identified with a Tikhonov regularization term. The performance of this approach depends heavily on the choice of kernels which need to be carefully designed [8]. This kernel design step poses similar problems as model structure selection in the classical paradigm. In addition, kernel-based identification controls model complexity through the norm of the impulse response induced by an arbitrary reproducing kernel Hilbert space [9]. Such complexity measures do not have clear system theoretic interpretations.

Alternative regularization approaches have also been proposed to directly control the number of poles of the model. This measure has a more concrete meaning for system analysis and control, either when the system is known to have a low-order structure, or when a low-order representation is desired. The Hankel nuclear norm of the impulse response is used as a convex surrogate in [10], [11]. However, this regularizer is known to be prone to stability issues [5]. A different approach consists of modeling the system as a summation of first-order "atoms", which are some predefined basis models. The model complexity can then be controlled by regularizing the l_1-norm of the coefficients. This is known as regularizing the atomic norm with respect to the atomic decomposition [12]. This results in a lasso-type problem that promotes models with a small number of poles [13]. This idea has also been used in periodic system identification [14] and kernel design [15]. Another advantage of the first-order atomic decomposition is that it directly identifies the pole locations of the system. Pole locations are important in classical control design, yet hard to estimate with conventional identification approaches.

Existing work on the atomic norm regularization approach, however, has multiple known drawbacks. First, instead of solving the group lasso problem on an infinite set of stable atoms, only a finite discretization of the atomic set is considered for tractability. This leads to an approximation error which can only be reduced with a very large set of atoms [12]. In addition, a large bias is induced by lasso-type regularization [5], and the pole location estimation contains a possibly large number of false positives due to the "p-value lottery" in high-dimensional regression [16].

In this paper, we propose an infinite-dimensional sparse learning algorithm based on atomic norm regularization, which aims to tackle the above drawbacks. This algorithm directly targets the group lasso problem with an infinite feature set, which has been studied in the machine learning literature [17], [18], [19]. Similar to Algorithm 1 in [17], our proposed algorithm first solves the problem with a small number of randomly generated features. Then, a new atomic model feature is selected to maximize the optimality condition violation for the previous iteration. The algorithm guarantees a decrease in the objective value per iteration and solves the infinite-dimensional problem with an arbitrarily small tolerance.

This work was supported by the Swiss National Science Foundation under Grant 200021_178890.

M. Yin, A. Iannelli, and R. S. Smith are with the Automatic Control Laboratory, Swiss Federal Institute of Technology (ETH Zurich), Physikstrasse 3, 8092 Zurich, Switzerland, {myin, iannelli, rsmith} @control.ee.ethz.ch.

M. T. Akan is with Control System Group, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands, m.t.akan@tue.nl.

Two different strategies are further presented to debias the estimate and reject false positives respectively. Iteratively reweighted adaptive group lasso [20], [21] is applied to reduce the amount of regularization on significant modes of the identified model, and thus reduce the bias. Complementary pairs stability selection (CPSS) [22], [23] solves the problem repeatedly on subsamples of the identification data and estimates the pole location by selecting atoms that are consistently active.

Numerical results demonstrate that the proposed algorithm performs better than PEM with an ARX model, kernel-based identification with tuned/correlated (TC) kernel design, and the existing atomic norm regularization algorithm in terms of impulse response fitting on a benchmark system. In addition, adaptive group lasso is able to reduce the bias of the algorithm and CPSS obtains more accurate pole location estimation than PEM with fewer false positives.

II. Atomic Norm Regularization in System Identification

In this work, we consider a strictly causal and stable linear time-invariant single-input single-output discrete-time system $y(t) = G_0(q)u(t) + v(t)$, where $u(t)$, $y(t)$, $v(t)$ are the inputs, outputs and additive noise respectively, and q is the shift operator. The transfer function $G_0(q)$ is assumed to have a low number of poles. The additive noise is assumed to be zero-mean i.i.d. Gaussian with a variance of σ^2. An input-output sequence of the system

$$\mathbf{u} = [u(1)\ u(2)\ \dots\ u(N)]^\top,\ \mathbf{y} = [y(1)\ y(2)\ \dots\ y(N)]^\top \quad (1)$$

has been collected. We are interested in identifying the transfer function $G_0(q)$ from the data sequence $(\mathbf{u}, \mathbf{y})$.

In regularized system identification, the transfer function $G_0(q)$ is expressed with a general high-dimensional parametrization $G_0(q) = \sum_{k\in K} c_k A_k(q)$, where $A_k(q)$ are the basis transfer functions known as atoms [12], c_k are the corresponding coefficients, and K denotes the set of indices. Denote the set of coefficients as $C = \{c_k | k \in K\}$. The following regularized optimization problem is solved:

$$\underset{C}{\text{minimize}} \quad V\left(\mathbf{y} - \sum_{k\in K} c_k\, \phi(A_k(q), \mathbf{u})\right) + \lambda J(C), \quad (2)$$

where $\phi(A(q), \mathbf{u})$ denotes the length-N output response of the system $A(q)$ to the inputs $\mathbf{u}$, $V(\cdot)$ is the loss function that penalizes the output residuals, $J(\cdot)$ is the regularization term that encodes prior knowledge of the coefficients, and λ is the regularization parameter to tune the amount of regularization. For the rest of the paper, the loss function is selected as $V(x) = \|x\|_2^2$, which is related to the maximum likelihood estimator when the noise $v(t)$ is i.i.d. Gaussian.

In this paper, the atomic decomposition of the transfer function in [12] is employed, where $A_k(q) = \dfrac{1-|k|^2}{q-k}$, and the corresponding coefficients c_k are complex numbers. Unlike conventional parametrizations, here k is a stable pole within the open unit disk. The set of indices is thus

$$K = \{k = \alpha \exp(j\beta) \,|\, \alpha \in [0,1), \beta \in [0, 2\pi)\}, \quad (3)$$

which has infinite elements. The atoms $A_k(q)$ are normalized to have a Hankel nuclear norm of 1. Define the pole locations of the system as $S = \{k \,|\, |c_k| > 0\}$, which is also known as the active atomic set. Since the system is known to have a small number of poles, a sparsity-promoting regularization term $J(C)$ is desired. In particular, an l_1-norm regularizer

$$J(C) = \sum_{k\in K} |c_k| \quad (4)$$

is used and defined as the atomic norm of the model [24]. Observe that for real-rational systems, the pole locations should be in conjugate pairs and the corresponding atomic responses are also complex conjugates of one another, i.e., $\phi(A_{\bar{k}}(q), \mathbf{u}) = \bar{\phi}(A_k(q), \mathbf{u})$, where the overbar denotes the complex conjugate. This means that coefficients for a conjugate pole pair should also be complex conjugates, i.e., $c_{\bar{k}} = \bar{c}_k$. Adding this constraint on the coefficients of (2), the problem can be reformulated as

$$\underset{\{c_k\}_{k\in\hat{K}}}{\text{minimize}} \quad \left\| \mathbf{y} - \sum_{k\in\hat{K}} \left(c_k\, \phi_k + \bar{c}_k\, \bar{\phi}_k\right) \right\|_2^2 + 2\lambda \sum_{k\in\hat{K}} |c_k|, \quad (5)$$

where $\phi_k := \phi(A_k(q), \mathbf{u})$ and

$$\hat{K} = \{k = \alpha \exp(j\beta) \,|\, \alpha \in [0,1), \beta \in [0, \pi]\} \quad (6)$$

denotes the upper half of the open unit disk.

Using $\Re$ and $\Im$ to denote real and imaginary parts, let

$$\gamma_k = \begin{bmatrix}\Re(c_k) & \Im(c_k)\end{bmatrix}^\top,\ \zeta_k = \begin{bmatrix}2\Re(\phi_k) & -2\Im(\phi_k)\end{bmatrix}. \quad (7)$$

Substituting (7) into (5), (5) can be expressed as a real-valued problem,

$$\Gamma^\star := \{\gamma_k^\star\}_{k\in\hat{K}} = \underset{\{\gamma_k\}_{k\in\hat{K}}}{\operatorname{argmin}} \underbrace{\left\| \mathbf{y} - \sum_{k\in\hat{K}} \zeta_k \gamma_k \right\|_2^2 + 2\lambda \sum_{k\in\hat{K}} \|\gamma_k\|_2}_{J(\Gamma)}, \quad (8)$$

where $\Gamma := \left\{\gamma_k \,|\, k \in \hat{K}\right\}$. Note that (8) is a standard group lasso problem [13]. The identified transfer function can be recovered by

$$\hat{G}(q) = \sum_{k\in\hat{K}} [1 \quad j]\, \gamma_k^\star A_k(q) + [1 \quad -j]\, \gamma_k^\star A_{\bar{k}}(q), \quad (9)$$

and the estimated pole locations are

$$\hat{S} = \left\{k \,\middle|\, \|\gamma_k^\star\|_2 > 0\right\} \cup \left\{\bar{k} \,\middle|\, \|\gamma_k^\star\|_2 > 0\right\}. \quad (10)$$

However, problem (8) cannot be directly solved since it is an infinite-dimensional problem. Existing algorithms relax this problem by approximating $\hat{K}$ with a discrete grid [12]. As shown in Proposition 4.1 of [12], the discretization induces a relative error in the atomic norm that is inversely proportional to the square root of the number of elements in the discretized $\hat{K}$.

III. Algorithm for Infinite-Dimensional Atomic Norm Regularization Problems

In this section, an algorithm is proposed to directly solve the infinite-dimensional problem (8). This algorithm is inspired by the feature generation algorithm in [17].

Problem (8) is a non-differentiable convex program, whose optimality conditions are given by $0 \in \partial J(\Gamma)$, where ∂ denotes the subdifferential. In detail, the optimality conditions of (8) are

$$\begin{cases} \left\|\zeta_k^\top R\right\|_2 \leq \lambda, & \text{if } \left\|\gamma_k^\star\right\|_2 = 0, \\ \zeta_k^\top R + \lambda \gamma_k^\star / \left\|\gamma_k^\star\right\|_2 = 0, & \text{if } \left\|\gamma_k^\star\right\|_2 > 0, \end{cases} \tag{11}$$

for all $k \in \hat{K}$, where $R := \mathbf{y} - \sum_{k \in \hat{K}} \zeta_k \gamma_k^\star$ is the vector of output residuals. The derivation makes use of the property

$$\partial \left\|\gamma_k^\star\right\|_2 = \begin{cases} \{w \mid \|w\|_2 \leq 1\}, & \left\|\gamma_k^\star\right\|_2 = 0, \\ \gamma_k^\star / \left\|\gamma_k^\star\right\|_2, & \left\|\gamma_k^\star\right\|_2 > 0. \end{cases} \tag{12}$$

Let $\hat{K}_d = \{k_1, k_2, \ldots, k_p\}$ be a finite subset of $\hat{K}$ with p elements. Then, with an abuse of notation, by replacing $\hat{K}$ with $\hat{K}_d$ in (8), a discretized optimal solution, denoted by $\Gamma^\star(\hat{K}_d) := \{\gamma_i^\star(\hat{K}_d)\}_{i=1}^p$, can be obtained, which satisfies

$$\begin{cases} \left\|\zeta_i(\hat{K}_d)^\top R(\hat{K}_d)\right\|_2 \leq \lambda, & \text{if } \left\|\gamma_i^\star(\hat{K}_d)\right\|_2 = 0, \\ \zeta_i(\hat{K}_d)^\top R(\hat{K}_d) + \lambda \dfrac{\gamma_i^\star(\hat{K}_d)}{\left\|\gamma_i^\star(\hat{K}_d)\right\|_2} = 0, & \text{if } \left\|\gamma_i^\star(\hat{K}_d)\right\|_2 > 0, \end{cases} \tag{13}$$

for $i = 1, \ldots, p$, where $R(\hat{K}_d) := \mathbf{y} - \sum_{i=1}^p \zeta_i(\hat{K}_d)\gamma_i^\star(\hat{K}_d)$ and $\zeta_i(\hat{K}_d) := \zeta_{k_i}$.

Suppose we want to add a new element k_{p+1} to $\hat{K}_d$. Then the optimal solution with respect to $\hat{K}_d^+ := \hat{K}_d \cup \{k_{p+1}\}$ is

$$\gamma_i^\star(\hat{K}_d^+) = \begin{cases} \gamma_i^\star(\hat{K}_d), & i = 1, \ldots, p, \\ \mathbf{0}, & i = p+1, \end{cases} \tag{14}$$

iff $\left\|\zeta_{p+1}(\hat{K}_d^+)^\top R(\hat{K}_d)\right\|_2 \leq \lambda$. In other words, adding such new elements does not improve the optimal objective function value, or change the transfer function estimate $\hat{G}(q)$. So the new element only reduces the objective function value when $\left\|\zeta_{k_{p+1}}^\top R(\hat{K}_d)\right\|_2 > \lambda$. This also guarantees $k_{p+1} \notin \hat{K}_d$ since $\left\|\zeta_{k_i}^\top R(\hat{K}_d)\right\|_2 \leq \lambda$ for $i = 1, \ldots, p$.

Motivated by the above observation, Algorithm 1 is proposed to solve the infinite-dimensional group lasso problem (8), where a greedy strategy is applied that chooses the new element by maximizing $\left\|\zeta_{k_{p+1}}^\top R(\hat{K}_d)\right\|_2$. Note that $\hat{K}_d^l$ denotes the set $\hat{K}_d$ at the l-th iteration in Algorithm 1. The transfer function and the pole location estimates $\hat{G}(q)$ and $\hat{S}$ can be calculated by (9) and (10) respectively with discretized atomic set $\hat{K}_d^l$, which is the output of Algorithm 1.

Let

$$\hat{\Gamma}^\star = \left\{ \gamma_k^\star \;\middle|\; \gamma_k^\star = \begin{cases} \gamma_i^\star(\hat{K}_d^l), & k = k_i \in \hat{K}_d^l \\ \mathbf{0}, & k \in \hat{K} \setminus \hat{K}_d^l \end{cases} \right\}. \tag{16}$$

Algorithm 1 guarantees the following property.

Algorithm 1 A greedy algorithm for the infinite-dimensional group lasso problem (8)

```
Input: identification data $(\mathbf{u}, \mathbf{y})$, $\varepsilon > 0$, $l_{\max}$
Initialize $\hat{K}_d^0 = \{k_1, k_2, \ldots, k_{p_0}\}$.
Calculate $\Gamma^\star(\hat{K}_d^0)$.
$l \leftarrow 0$
repeat
    Construct a candidate new atom
        $k^+ \leftarrow \operatorname{argmax}_{k \in \hat{K}} \left\|\zeta_k^\top R(\hat{K}_d^l)\right\|_2.$    (15)
    if $\left\|\zeta_{k^+}^\top R(\hat{K}_d^l)\right\|_2 \geq \lambda + \varepsilon$ then
        begin
        $k_{p_0+l+1} \leftarrow k^+$, $\hat{K}_d^{l+1} \leftarrow \hat{K}_d^l \cup \{k_{p_0+l+1}\}$
        Calculate $\Gamma^\star(\hat{K}_d^{l+1})$ via program (8).
        end
    else
        Break
    $l \leftarrow l + 1$
until $l \geq l_{\max}$
Output: $\hat{K}_d^l$, $\Gamma^\star(\hat{K}_d^l)$
```

Proposition 1: If Algorithm 1 terminates without reaching the maximum number of iterations ($l < l_{\max}$), $\hat{\Gamma}^\star$ satisfies the approximate optimality conditions

$$\begin{cases} \left\|\zeta_k^\top R\right\|_2 < \lambda + \varepsilon, & \text{if } \left\|\gamma_k^\star\right\|_2 = 0, \\ \zeta_k^\top R + \lambda \gamma_k^\star / \left\|\gamma_k^\star\right\|_2 = 0, & \text{if } \left\|\gamma_k^\star\right\|_2 > 0, \end{cases} \tag{17}$$

for all $k \in \hat{K}$.

Proof: Since $\gamma_k^\star = 0$ for $k \notin \hat{K}_d^l$ in $\hat{\Gamma}^\star$, we have $R = R(\hat{K}_d^l)$. For $k \in \hat{K}_d^l$, the discretized optimality conditions (13) guarantee the satisfaction of (17). According to Algorithm 1, $\left\|\zeta_k^\top R(\hat{K}_d^l)\right\|_2 = \left\|\zeta_k^\top R\right\|_2 < \lambda + \varepsilon$. So for $k \notin \hat{K}_d^l$, (17) is satisfied since $\left\|\gamma_k^\star\right\|_2 = 0$. ■

Proposition 1 shows that the infinite-dimensional problem (8) is approximately equivalent to the finite-dimensional problem with $(p_0 + l)$ atoms

$$\underset{\{\gamma_i\}_{i=1}^{p_0+l}}{\operatorname{argmin}} \left\|\mathbf{y} - \sum_{i=1}^{p_0+l} \zeta_{k_i}\gamma_i\right\|_2^2 + 2\lambda \sum_{i=1}^{p_0+l} \|\gamma_i\|_2. \tag{18}$$

For the rest of the paper, define $p = p_0 + l$.

The main difficulty in Algorithm 1 is solving the non-convex problem (15). However, even if (15) is not solved exactly, Algorithm 1 still guarantees a decrease in the objective function value at each iteration as long as $\left\|\zeta_{k^+}^\top R(\hat{K}_d^l)\right\|_2 \geq \lambda + \varepsilon$ is satisfied for the candidate atom k^+.

IV. Debiasing and Stability Selection

Algorithm 1 provides a method to solve the group lasso problem (8). However, solutions to lasso-type regularized problems are known to have a large bias and a large number of false positives in feature selection [22]. To mitigate these problems, the following tools in high-dimensional statistics are applied to debias the estimate and reject false positives in pole location estimation from Algorithm 1.

Algorithm 2 Iteratively reweighted adaptive group lasso

```
Input: identification data $(\mathbf{u},\mathbf{y})$, $\varepsilon' > 0$, $m_s$
Find $\hat{K}_d^l = \left\{k_1,\ldots,k_p\right\}$, $\Gamma^\star(\hat{K}_d^l) := \left\{\gamma_1^{\star,0},\ldots,\gamma_p^{\star,0}\right\}$ from Algorithm 1.
for $m = 1$ to $m_s$ do
    begin
    Find $\left\{\gamma_i^{\star,m}\right\}_{i=1}^p$ by solving
        $\underset{\{\gamma_i\}_{i=1}^p}{\operatorname{argmin}} \left\| \mathbf{y} - \sum_{i=1}^p \zeta_{k_i}\gamma_i \right\|_2^2 + 2\lambda \sum_{i=1}^p \frac{\|\gamma_i\|_2}{\left\|\gamma_i^{\star,m-1}\right\|_2 + \varepsilon'}.$   (21)
    end
Calculate $\hat{G}(q)$ by (9) with discretized atomic set $\hat{K}_d^l$ and coefficients $\left\{\gamma_i^{\star,m_s}\right\}_{i=1}^p$.
Output: $\hat{G}(q)$
```

Algorithm 3 Complementary pairs stability selection

```
Input: identification data $(\mathbf{u},\mathbf{y})$, $\tau \in (0.5,1]$, $n_s$
Find $\hat{K}_d^l$ from Algorithm 1.
for $i = 1$ to $n_s$ do
    begin
    Generate a random subsample $B_i \subset \{1,2,\ldots,N\}$ with $\lfloor N/2 \rfloor$ elements.
    $\bar{B}_i \leftarrow \{1,2,\ldots,N\} \setminus B_i$
    Calculate $\hat{S}_{B_i}$, $\hat{S}_{\bar{B}_i}$ by solving (18) with the loss function $V(\cdot)$ replaced by $V_{B_i}(\cdot)$, $V_{\bar{B}_i}(\cdot)$ respectively.
    end
$\hat{S} \leftarrow \left\{k \,\middle|\, \frac{1}{2n_s}\sum_{i=1}^{n_s}\left(\mathbb{1}_{\hat{S}_{B_i}}(k) + \mathbb{1}_{\hat{S}_{\bar{B}_i}}(k)\right) \geq \tau\right\}$, where $\mathbb{1}$ denotes the indicator function.
Output: $\hat{S}$
```

A. Iteratively Reweighted Adaptive Group Lasso

The l_1-norm regularizer (4) is a convex relaxation of the ideal sparsity promoting function $J^*(C) = n(S)$, where $n(\cdot)$ denotes the cardinality of the set, which counts the number of poles in the model. Compared to the ideal regularizer which penalizes all the active atoms with a fixed value of 1, the l_1-norm regularizer penalizes them with the magnitude of the corresponding coefficients. This induces a negative bias, especially for the atoms with larger coefficients, i.e., the dominant modes. This bias is a large source of error in atomic norm regularization [5].

To reduce such bias, adaptive lasso [25] has been proposed which adds a second step that applies a reweighted version of the l_1-norm regularizer

$$J_{\mathrm{a}}(C) = \sum_{k\in K} \frac{|c_k|}{\left|c_k^{\star,0}\right| + \varepsilon'}, \tag{19}$$

where $c_k^{\star,0}$ is the solution to the original problem, and $\varepsilon' > 0$ is a small constant to avoid singularity. This regularizer reduces the amount of regularization for atoms estimated with large coefficients in the original problem, and is close to $J^*(C)$ when $c_k \approx c_k^{\star,0}$. This approach is extended to apply this reweighting iteratively (Section 2.8.5 in [22]), which is sometimes known as iteratively reweighted lasso. It is pointed out in [21] that the iteratively reweighted lasso can be interpreted as a difference of convex programming algorithm to solve the regularized problem with a non-convex log regularizer

$$J_{\log}(C) = \sum_{k\in K} \frac{\log\left(|c_k| + \varepsilon'\right)}{\log \varepsilon'}. \tag{20}$$

This iteratively reweighted adaptive approach is applied to the group lasso problem (8) in Algorithm 2. It is easy to see that the cardinality of the active atomic set $J^*(C)$ is non-increasing at each iteration.

B. Complementary Pairs Stability Selection

Lasso-type regularized problems are known to have favorable consistency properties in terms of prediction under mild conditions. However, in terms of estimating the active atomic set S, they can only guarantee that the non-active atoms are not in the true model with high probability under practical assumptions (Chapter 2 in [22]). This means that the number of false positives in the estimated pole locations is not controlled. In fact, there are usually many more estimated poles than the true poles, with many occurring at "random" locations depending on the noise realization. This will be shown in Section V. This phenomenon is known as "p-value lottery" [16].

Subsampling techniques have been used to increase the stability of the active atomic set estimation. In particular, the complementary pairs stability selection (CPSS) is applied in this work [23]. This method generates complementary pairs of subsamples from the identification data, and repeats the baseline variable selection procedure (group lasso problem (8) here) on each subsample. For our problem, this corresponds to replacing the loss function with

$$V_B(\cdot) = \left\| \mathbf{y}(B) - \sum_{i=1}^p \zeta_{k_i}(B,:)\gamma_i \right\|_2^2, \tag{22}$$

where $B \subset \{1,2,\ldots,N\}$ defines a random subsample of data. Define the estimated pole locations on the subsample as $\hat{S}_B$. Then the so-called stable solution of the problem is defined as the atoms that have higher empirical probabilities of being included in $\hat{S}_B$ than a predefined threshold τ. The algorithm has favorable false-positive rejection properties when $\tau > 0.5$ [23]. The method is summarized in Algorithm 3. The transfer function can also be estimated by least squares on the stable solution of the atomic set.

V. Numerical Results

The performances of the proposed algorithms are assessed by numerical simulation on a benchmark fourth-order system previously analyzed in [26]:

$$G(q) = \frac{0.10884q + 0.19513}{q^4 - 1.41833q^3 + 1.58939q^2 - 1.31608q + 0.88642}. \tag{23}$$

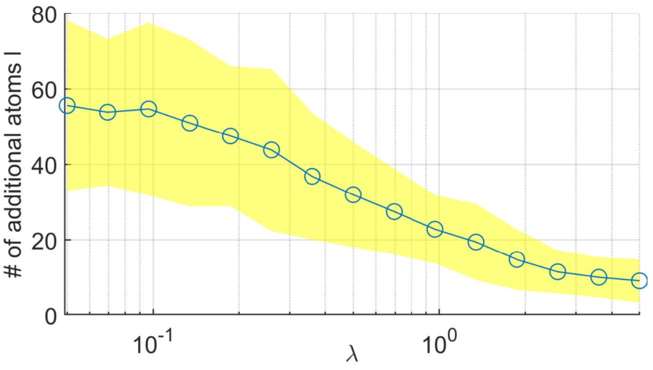

Fig. 1. The number of additional atoms l in Algorithm 1 for $\sigma^2 = 0.1$. Blue: mean values, yellow: ranges within one standard deviation.

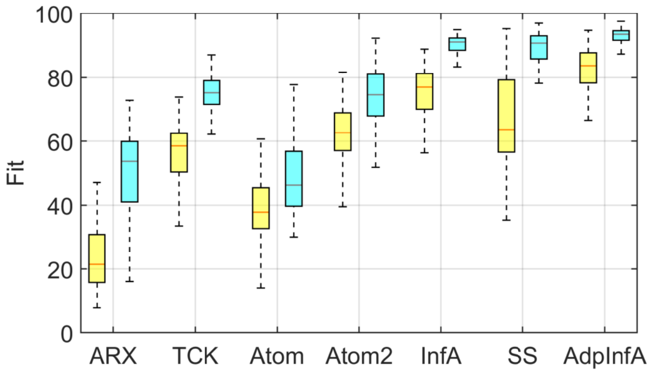

Fig. 2. Boxplot of impulse response fitting. Yellow: $\sigma^2 = 0.1$, cyan: $\sigma^2 = 0.01$.

TABLE I

BIAS-VARIANCE ANALYSIS OF IMPULSE RESPONSE ESTIMATION

	TCK	**Atom**	**Atom2**	**InfA**	**SS**	**AdpInfA**
$\sigma^2 = 0.1$						
Bias^2 $[\times 10^{-2}]$	6.76	23.42	6.34	2.63	8.28	0.91
Var $[\times 10^{-2}]$	13.04	13.59	8.52	3.80	15.68	2.70
MSE $[\times 10^{-2}]$	19.80	37.01	14.86	6.44	23.96	3.60
$\sigma^2 = 0.01$						
Bias^2 $[\times 10^{-2}]$	1.78	15.92	2.22	0.43	0.47	0.07
Var $[\times 10^{-2}]$	5.45	11.68	5.26	0.76	3.12	0.52
MSE $[\times 10^{-2}]$	7.23	27.60	7.48	1.18	3.59	0.59

The system has been normalized to have an $\mathscr{H}_2$-norm of 1. In what follows, results obtained with Algorithms 1, 2, and 3 are labelled by *InfA*, *AdpInfA*, and *SS* respectively.

Identification data of length $N = 100$ are generated with zero-mean i.i.d. unit Gaussian inputs from a zero initial condition. Two noise levels $\sigma^2 = 0.1$ and 0.01 are considered. The atomic responses ϕ_k are also generated from a zero initial condition. 100 Monte Carlo simulations are conducted for each noise level. The initial discretized atomic set $\hat{\mathcal{K}}_d^0$ contains $p_0 = 50$ randomly generated atoms with $k_i = \alpha_i \exp(j\beta_i)$, where α_i and β_i are subject to uniform distributions in $[0,1)$ and $[0,\pi]$ respectively. Finite-dimensional group lasso problems are solved by MOSEK. The candidate atom generation problem (15) is solved by the particle swarm solver in MATLAB. The hyperparameter λ is selected by cross-validation from a 15-point log-space grid between 0.05 and 5 for $\sigma^2 = 0.1$, and between 0.005 and 0.5 for $\sigma^2 = 0.01$, except for *SS* where λ is fixed to 0.5 for $\sigma^2 = 0.1$ and 0.05 for $\sigma^2 = 0.01$. The following parameters are used in simulation: $\varepsilon = \varepsilon' = 10^{-5}$, $\tau = 0.9$, $n_s = 50$, $m_s = 2$.

First, the number of additional atoms l required in Algorithm 1 is plotted against the λ values in Figure 1. The maximum l in all Monte Carlo simulations is 118, which is below the $l_{\max}$ setting. Results show that the proposed greedy atom generation approach is able to converge within a reasonable number of iterations, and the required number of additional atoms decreases with λ.

To demonstrate the performance of the proposed algorithms, they are compared to three benchmark algorithms: 1) least-squares estimation with an ARX model and a known model order (*ARX*); 2) kernel-based identification with a TC kernel design (*TCK*) [9]; 3) discretized atomic norm regularization in [12] with 50 (*Atom*) and 500 (*Atom2*) random atoms. Note that *Atom2* uses a significantly larger atomic set compared to Algorithm 1, as shown in Figure 1.

Figure 2 compares the identification accuracy of all algorithms in terms of the impulse response fitting, defined as

$$W = 100 \cdot \left(1 - \left[\frac{\sum_{i=1}^{N-1}(g_i - \hat{g}_i)^2}{\sum_{i=1}^{N-1}(g_i - \bar{g})^2}\right]^{1/2}\right), \qquad (24)$$

where g_i are the true impulse responses, $\hat{g}_i$ are the estimated impulse responses, and $\bar{g}$ is the mean of g_i. It can be seen that the three proposed algorithms all perform better than the benchmark algorithms at both noise levels. In particular, *InfA* obtains better fitting compared to *Atom2* which uses a much larger atomic set. This demonstrates the effectiveness of the proposed atom generation approach. *AdpInfA* further improves on the identification accuracy of *InfA* with iterative reweighting.

To further investigate the sources of the estimation errors, Table I shows the bias-variance analysis of impulse response estimation. As an algorithm proposed to debias the estimate, *AdpInfA* indeed produces a much smaller bias compared to all other algorithms. This is also the main contributor to the reduction of MSE compared to the baseline *InfA* algorithm, on which *AdpInfA* is based.

Finally, the capability of estimating the poles of the system is demonstrated in Figures 3 and 4. It is illustrated in Figure 3 that all the algorithms that directly solve group lasso problems estimate a much larger number of poles compared to the true one. *AdpInfA* mitigates the over-estimation since the active atomic set shrinks at each iteration, whereas SS obtains a very accurate estimation of the model order.

To assess the accuracy of pole location estimation, Figure 4 further compares the distributions of estimated pole locations in all 100 Monte Carlo simulations. Despite knowing the true model order, *ARX* fails to give accurate estimations of the pole locations. Although the estimated model order is close to the true one, *AdpInfA* estimates a significant number

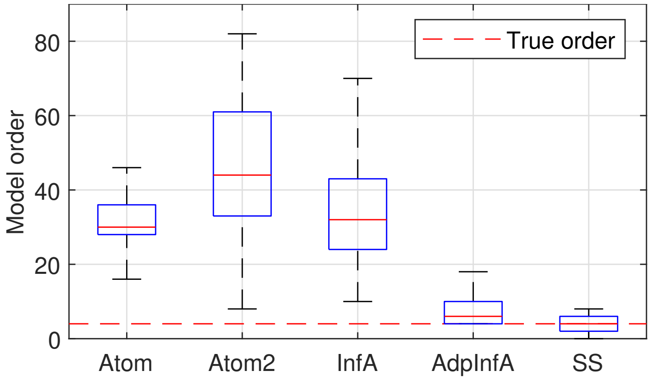

Fig. 3. Comparison of estimated model orders for $\sigma^2 = 0.1$.

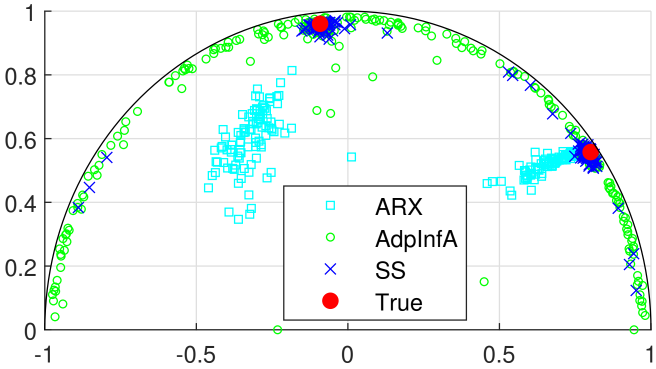

Fig. 4. Comparison of pole location estimation distributions in all 100 Monte Carlo simulations for $\sigma^2 = 0.1$.

of false positives in terms of the actual pole locations. Among all the algorithms, only *SS* is able to obtain accurate pole location estimations with few false positives, which proves the effectiveness of the CPSS method.

VI. Conclusions

This work applies advanced techniques studied in high-dimensional statistics to the atomic norm regularization problem in linear system identification. A greedy algorithm is presented to generate new candidate atomic models from infinitely many possible pole locations. Common drawbacks of lasso-type regularization are mitigated by adaptively adjusting the regularization weights for each atom and selecting only repeatedly occurring pole locations from subsamples of data. Results in this paper suggest that sparse learning algorithms are a promising alternative to kernel-based methods with fewer design requirements and direct pole location estimation. Further research directions include improvements in computational efficiency, comparison with model order reduction methods, and extensions to MIMO systems and frequency-domain data.

References

[1] L. Ljung, *System Identification: Theory for the User*. Upper Saddle River, NJ, USA: Prentice-Hall, 1999.

[2] K. Åström, "Maximum likelihood and prediction error methods," *Automatica*, vol. 16, no. 5, pp. 551–574, 1980.

[3] L. Ljung, "On convexification of system identification criteria," *Automation and Remote Control*, vol. 80, no. 9, pp. 1591–1606, 2019.

[4] L. Ljung, T. Chen, and B. Mu, "A shift in paradigm for system identification," *International Journal of Control*, vol. 93, no. 2, pp. 173–180, 2019.

[5] G. Pillonetto, T. Chen, A. Chiuso, G. D. Nicolao, and L. Ljung, "Regularized linear system identification using atomic, nuclear and kernel-based norms: the role of the stability constraint," *Automatica*, vol. 69, pp. 137–149, 2016.

[6] G. Pillonetto and G. De Nicolao, "A new kernel-based approach for linear system identification," *Automatica*, vol. 46, no. 1, pp. 81–93, 2010.

[7] G. Pillonetto, 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, 2014.

[8] T. Chen, "On kernel design for regularized LTI system identification," *Automatica*, vol. 90, pp. 109–122, 2018.

[9] T. Chen, H. Ohlsson, and L. Ljung, "On the estimation of transfer functions, regularizations and gaussian processes—revisited," *Automatica*, vol. 48, no. 8, pp. 1525–1535, 2012.

[10] R. S. Smith, "Frequency domain subspace identification using nuclear norm minimization and Hankel matrix realizations," *IEEE Transactions on Automatic Control*, vol. 59, no. 11, pp. 2886–2896, 2014.

[11] M. Fazel, H. Hindi, and S. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings of the 2001 American Control Conference. (Cat. No.01CH37148)*, vol. 6, 2001, pp. 4734–4739.

[12] P. Shah, B. N. Bhaskar, G. Tang, and B. Recht, "Linear system identification via atomic norm regularization," in *51st IEEE Conference on Decision and Control (CDC)*, 2012, pp. 6265–6270.

[13] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[14] M. Yin, A. Iannelli, M. Khosravi, A. Parsi, and R. S. Smith, "Linear time-periodic system identification with grouped atomic norm regularization," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 1237–1242, 2020, 21st IFAC World Congress.

[15] M. Khosravi, M. Yin, A. Iannelli, A. Parsi, and R. S. Smith, "Low-complexity identification by sparse hyperparameter estimation," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 412–417, 2020, 21st IFAC World Congress.

[16] N. Meinshausen, L. Meier, and P. Bühlmann, "p-values for high-dimensional regression," *Journal of the American Statistical Association*, vol. 104, no. 488, pp. 1671–1681, 2009.

[17] A. Rakotomamonjy, R. Flamary, and F. Yger, "Learning with infinitely many features," *Machine Learning*, vol. 91, no. 1, pp. 43–66, 2012.

[18] S. Rosset, G. Swirszcz, N. Srebro, and J. Zhu, "l_1 regularization in infinite dimensional feature spaces," in *Learning Theory*. Berlin, Heidelberg: Springer, 2007, pp. 544–558.

[19] I. E. H. Yen, T. W. Lin, S. D. Lin, P. K. Ravikumar, and I. S. Dhillon, "Sparse random feature algorithm as coordinate descent in Hilbert space," in *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[20] H. Wang and C. Leng, "A note on adaptive group lasso," *Computational Statistics & Data Analysis*, vol. 52, no. 12, pp. 5277–5286, 2008.

[21] G. Gasso, A. Rakotomamonjy, and S. Canu, "Recovering sparse signals with a certain family of nonconvex penalties and DC programming," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4686–4698, 2009.

[22] P. Bühlmann and S. van de Geer, *Statistics for high-dimensional data: methods, theory and applications*. Berlin, Heidelberg: Springer, 2011.

[23] R. D. Shah and R. J. Samworth, "Variable selection with error control: another look at stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 75, no. 1, pp. 55–80, 2013.

[24] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[25] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[26] I. Landau, D. Rey, A. Karimi, A. Voda, and A. Franco, "A flexible transmission system as a benchmark for robust digital control," *European Journal of Control*, vol. 1, no. 2, pp. 77–96, 1995.