A Randomized Algorithm for Parsimonious Model Identification

Burak Yılmaz, *Member, IEEE*, Korkut Bekiroglu, *Member, IEEE*, Constantino Lagoa, *Member, IEEE*, and Mario Sznaier, *Member, IEEE*

***Abstract*—Identifying parsimonious models is generically a "hard" nonconvex problem. Available approaches typically rely on relaxations such as Group Lasso or nuclear norm minimization. Moreover, incorporating stability and model order constraints into the formalism in such methods entails a substantial increase in computational complexity. Motivated by these challenges, in this paper we present algorithms for parsimonious linear time invariant system identification aimed at identifying low-complexity models which i) incorporate *a priori* knowledge on the system (e.g., stability), ii) allow for data with missing/nonuniform measurements, and iii) are able to use data obtained from several runs of the system with different unknown initial conditions. The randomized algorithms proposed are based on the concept of atomic norm and provide a numerically efficient way to identify sparse models from large amounts of noisy data.**

***Index Terms*—Atomic norm, Frank–Wolfe, Hankel singular values, identification, optimization.**

I. Introduction

Dynamical system based approaches have proven very successful in many areas including some "nonstandard" ones such as design of medical/behavioral treatment and video-analytics. However, when identifying models from data, one can be faced with significant challenges. Namely, i) large data sets, ii) significant amount of measurement noise, and iii) data fragmentation (due e.g., to missing measurements). Successful handling of these scenarios requires an algorithm that uses both *a priori* information and the fragmented data to obtain a "simple" model that explains the behavior observed.

A. Previous Work

Set membership methods [1] generate a model consistent with the experimental data and priors, along with a bound on the worst case identification error that can be directly used by robust control methods. However, they may lead to high-order models, necessitating a model reduction step. Alternatively, subspace-based methods [2] are computationally attractive and can easily be modified to enforce bounds on the order of the resulting model. However, in this context it is hard to enforce consistency with existing *a priori* information such as stability or bounds on the time constants or the identification error. In addition, these methods cannot handle fragmented data records. The latter difficulty can be solved by simply considering the missing data as parameters (see, e.g., [3] for parametric identification and [4], [5] for the nonparametric case), but these approaches still cannot impose consistency.

Recently, considerable effort has been devoted to developing an identification framework capable of handling noisy, fragmented data, while leading to low-order models. In most cases, this is accomplished by reducing the original problem to a sequence of convex semidefinite programs, by using the nuclear norm as a surrogate for rank [6]–[8]. However, incorporating stability constraints into the formalism entails a substantial increase in the computational complexity [8].

To address these issues, we present a new, a computationally efficient framework for parsimonious system identification. The proposed algorithm can easily incorporate *a priori* knowledge on the system, seamlessly handle missing/nonuniform measurements and merge data acquired from multiruns of the system with unknown different initial conditions. This approach is motivated by the recent work in [9] and [10], showing that the problem of obtaining sparse representations of elements contained in the convex hull of a set of suitably chosen points (atoms) can be recast as a convex constrained approximation problem. A potential difficulty in applying these results to identification is that in this case the set of atoms is infinite [all possible impulse responses of stable linear time invarian (LTI) plants], leading to an infinite dimensional (albeit convex) optimization problem. In [9], this difficulty was handled by approximating this set with a finite one, obtained by considering an ϵ-net discretization of the unit disk, combined with an ℓ_1 regularization to enforce sparsity of the resulting representation. While this approach works well for smooth plants, handling lightly damped systems may require considering very fine discretizations, with the entailed increase in computational complexity. Moreover, using first-order plants as the atoms results in complex valued impulse response vectors. This might introduce numerical problems while solving the optimization problem. Finally, the atom normalization factor used in [9] may result in numerical difficulties for systems with slowly decaying modes, especially when the data horizon length is relatively short. In this paper, a randomized algorithm for parsimonious LTI model identification that addresses these issues is developed.

II. Preliminaries

A. Notation

Lower (upper)-case boldface letters denote vectors (matrices). For a complex number $p \in \mathbb{C}$, $\Re(p), \Im(p)$, and $\bar{p}$ denote real and imaginary parts of p, and complex conjugate. $\mathbb{D}_\rho$ denotes the origin centered closed disc in $\mathbb{C}$, with radius ρ. The convex hull of the set

Manuscript received May 6, 2017; accepted July 2, 2017. Date of publication July 6, 2017; date of current version January 26, 2018. This work was supported in part by National Institute on Drug Abuse under Grant P50 DA039838, and in part by NSF under Grant CNS-1329422, Grant ECCS-1201973, Grant ECCS-1404163, and Grant CMMI 1638234. Recommended by Associate Editor M. Verhaegen. *(Corresponding author: Korkut Bekiroglu.)*

B. Yılmaz is with the Medtronic Minimally Invasive Therapies Group, Minneapolis, MN 55432-5604 USA (e-mail: burak.yilmaz@medtronic.com).

K. Bekiroglu is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: korkutbekiroglu@gmail.com).

C. Lagoa is with the Department of Electrical Engineering, Penn State University, State College, PA 16802 USA (e-mail: lagoa@engr.psu.edu).

M. Sznaier is with the Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: msznaier@coe.neu.edu).

Digital Object Identifier 10.1109/TAC.2017.2723959

S is denoted by $\mathrm{conv}(S)$. $\mathrm{mat}(\mathbf{u}, n, m)$ reshapes the elements of u into $n \times m$ matrix, column-first order. $\Upsilon^N\{\cdot\}$ denotes the truncated N-length impulse response vector of a discrete transfer function. Finally, the lower triangular block Toeplitz matrix associated with any finite sequence $\{x_k, k = 0, 1, \ldots, n-1\}$, or any column vector $x = [x_0, x_1, \ldots, x_{n-1}]^T$ is denoted by

$$\mathbf{T}_x \doteq \begin{bmatrix} x_0 & 0 & \ldots & 0 \\ x_1 & x_0 & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 \\ x_{n-1} & x_{n-2} & \ldots & x_0 \end{bmatrix}.$$

B. Problem Statement

In this section, we define a simple system identification scenario for which we develop our core algorithm:

Problem 1: Given

1) an unknown plant $G(z)$, known to be SISO LTI with poles in $\mathbb{D}_\rho$ where $1 > \rho > 0$;
2) an input sequence u_t, applied to $G(z)$, for $t = 1, 2, \ldots, N$; and
3) an output time domain sequence y_t, $t = 1, 2, \ldots, N$, given by

$$y_t = (g * u)_t + \eta_t, \qquad t = 1, \ldots, N. \tag{1}$$

where g represents the impulse response of $G(z)$, $*$ denotes convolution and η_t denotes a noise sequence bounded by a known constant (e.g., $\|\eta\|_2 \leq \eta_{\max}$)

find the most parsimonious LTI model that explains the input–output pair within a given bound on estimation error.

Remark 1: After the tool set to tackle Problem 1 is established, we will extend the base problem to encompass more complex scenarios, e.g., data with missing samples and multiple runs with different unknown initial conditions.

C. Parsimonious System Identification

To construct our base sparse system identification problem, we use the following fact: Every strictly proper transfer function with poles in $\mathbb{D}_\rho$ can be approximated to arbitrary precision by a linear combination of *first-order* strictly proper transfer functions; i.e., for any proper nth order $G(z)$, one has

$$G(z) = \frac{B(z)}{A(z)} = \sum_{i=1}^{n} \frac{c_i}{z - p_i} \ : \ p_i \in \mathbb{D}_\rho \text{ and } c_i \in \mathbb{C}. \tag{2}$$

Given this, the parsimonious system identification problem can be formulated as follows: Let $\mathbf{g}_p$ be the truncated impulse response vector of the first-order system $G_p(z) = 1/(z-p)$. Given input u and (noisy) measurements of the output y, solve

$$\min_c \ \text{cardinality}\{c : c_p \neq 0\}$$
$$\text{s.t.} \sum_{t=0}^{N} \left[\sum_{p \in \mathbb{D}_\rho} c_p (\mathbf{T}_u \mathbf{g}_p)_t - y_t \right]^2 \leq \eta_{\max}^2. \tag{3}$$

Here, the objective function enforces parsimony of the transfer function $G(z)$, and the constraint enforces fidelity to collected data. However, there are several challenges associated with (3) that will be addressed in this paper. (C.1) The complex coefficients in c_p can lead to numerical difficulties that prevent finding the sparsest representation; (C.2) minimizing cardinality subject to constraints is an NP-hard problem; and (C.3) there are (uncountably) infinite poles in $\mathbb{D}_\rho$.

III. Convex Relaxation

In this section, we propose a new set of atoms and a convex relaxation of (3) to address the challenges C.1 and C.2.

A. Atoms for LTI System Identification

To address the challenge (C1), we propose a set of atoms that always assures a real impulse response. Define an operator $\mathcal{A}(\cdot)$ that takes in a set of complex poles S as an argument and produces a set of transfer functions as

$$\mathcal{A}\{S\} = \mathcal{A}_1\{S\} \cup \mathcal{A}_2\{S\} \cup \mathcal{A}_3\{S\} \cup \mathcal{A}_4\{S\}$$

where each suboperator is defined as follows:

$$\begin{aligned} \mathcal{A}_1\{S\} &= \left\{ \pm\alpha_p^1 \left(\frac{1}{z-p} + \frac{1}{z-\bar{p}} \right) \ : \ p \in S \right\} \\ \mathcal{A}_2\{S\} &= \left\{ \pm\alpha_p^2 \left(\frac{-j}{z-p} + \frac{j}{z-\bar{p}} \right) \ : \ p \in S \right\} \\ \mathcal{A}_3\{S\} &= \left\{ \pm\frac{\alpha_p}{z-p} \ : \ p \in S, p \text{ real} \right\} \\ \mathcal{A}_4\{S\} &= \{+1, -1\}. \end{aligned} \tag{4}$$

Here, α_p are scaling factors defined by

$$\begin{aligned} \alpha_p^1 &= \sqrt{2\big(\Re(\varphi_p^2) + \varphi_a^2\big) + 2\sqrt{2\Gamma\big(|\varphi_p|^2 - \varphi_a^2\big)}}^{-1} \\ \alpha_p^2 &= \sqrt{2\big(\varphi_a^2 - \Re(\varphi_p^2)\big) + 2\sqrt{2\Gamma\big(|\varphi_p|^2 - \varphi_a^2\big)}}^{-1} \\ \alpha_p &= (1-p^2)/(1-p^{2N+2}) \end{aligned} \tag{5}$$

$$\varphi_p = \frac{1-p^{2N}}{1-p^2} \text{ and } \varphi_a = \frac{1-|p|^{2N}}{1-|p|^2}$$

$$\Gamma = \frac{\Re(\varphi_p) - \varphi_a - \Re\big(p^2(\bar{p})^{2N}\varphi_p\big) + |p|^{2N+2}\varphi_a}{1-|p|^2} \tag{6}$$

where N is the length of the measurement vector $\mathbf{y} \in \mathbb{R}^N$. The α_p above are chosen so that the Hankel matrix of size N associated with the impulse response of each atom has nuclear norm equal to 1. Details of computation of these can be found in [11, Appendix A]. The objective of such a choice is to make the weight of the atoms in the objective function as "uniform" as possible. In practice, we have seen that this choice of weights produces good results.

The set of atoms used in this paper for system identification is $\mathcal{A}\{\mathbb{D}_\rho\}$, which enjoys the following properties.

1) Every proper rational transfer function with poles in $\mathbb{D}_\rho$ can be approximated as a real linear combination of atoms in the set [9].
2) Each atom in the set has a transfer function with real coefficients; hence, it has a purely real impulse response.

To relax the nonconvex problem (3) to a convex optimization [addressing (C.2)], one can consider fixed length impulse responses of the atoms described in (4). The fact that the impulse response of the linear combination of multiple LTI plants is the linear combination of the impulse responses of the individual plants forming the sum leads to the optimization problem described next.

B. Formulation as a Convex Optimization Problem

To be able to derive a convex approximation of problem (3), we start by defining the associated atomic norm. Let $\mathbf{g}$ be the first N terms of

Algorithm 1: Randomized algorithm to minimize a convex function f over the τ-scaled atomic norm ball.

1: $\mathbf{x_0} \leftarrow \tau \Upsilon^N \{a_0(z)\}$ for arbitrary $a_0(z) \in \mathcal{A}$ ▷Init.
2: **for** $k = 0, 1, 2, 3, \ldots, k_{\max}$ **do**
3: Select N_k poles uniformly distributed over $\mathbb{D}_\rho$, denote the set of these poles S_k
4: $\mathbf{a_k} \leftarrow \Upsilon^N \{\text{argmin}_{a(z)\in\mathcal{A}\{S_k\}} \langle \nabla f(x_k), \Upsilon^N \{a(z)\}\rangle\}$
5: $\alpha_k \leftarrow \text{argmin}_{\alpha\in[0,1]} f(\mathbf{x_k} + \alpha[\tau \mathbf{a_k} - \mathbf{x_k}])$
6: $\mathbf{x_{k+1}} \leftarrow \mathbf{x_k} + \alpha_k [\tau \mathbf{a_k} - \mathbf{x_k}]$
7: **end for**

the impulse response of a system $G(z)$. Define

$$\|\mathbf{g}\|_{\mathcal{A}} \doteq \left\{ \inf \sum_{a\in\mathcal{A}(\mathbb{D}_\rho)} |c_a| : \mathbf{g} = \sum_{a\in\mathcal{A}(\mathbb{D}_\rho)} c_a \Upsilon^N\{a\} \right\}.$$

This leads to the following convex relaxation of the parsimonious system identification problem (3):

$$\min_{\mathbf{g}} \quad \|\mathbf{T_u g} - \mathbf{y}\|_{\ell_2}^2$$
$$\text{subject to} \quad \|\mathbf{g}\|_{\mathcal{A}} \le \tau. \tag{7}$$

Low complexity is promoted by constraining the optimal solution to be inside the τ-scaled atomic norm ball ($\|\mathbf{g}\|_{\mathcal{A}} \le \tau$) [10]. Note that the *a priori* information about the stability margin of the unknown system, or other information about the poles of the identified system, is implicitly incorporated in the choice of the atomic set. Since the noise sequence is assumed to be bounded, the system to be identified can be approximated to arbitrary precision as described above with a *finite* $\|\mathbf{g}\|_{\mathcal{A}}$.

IV. Identification Algorithm

The optimization problem (7) is a special case of the class of problems studied in [12] where the authors consider problems involving atomic norms of the form

$$\min_{x} \quad f(x)$$
$$\text{subject to} \quad \|x\|_{\mathcal{A}} \le \tau \tag{8}$$

where $f(x)$ is a convex and smooth function. In this paper, we propose to solve the problem above using the following randomized version of the well-known Frank–Wolfe algorithm [addressing (C.3)]:

Next, we provide a step by step explanation of Algorithm 1. In step 1, a random atom is picked and scaled by τ to serve as the initial solution. Note that the initial solution belongs to the boundary of the feasible set. Then in step 3, a fixed number (N_k) of random atoms are selected from the atomic set. For step 4, the gradient is calculated as $\nabla f(x_k) = \mathbf{T_u}^T (\mathbf{T_u} x_k - y)$ for the objective function in problem (7). The random atomic responses selected in step 3 are checked exhaustively with the gradient vector and the optimum atom is found very efficiently. Finding the optimum α_k at step k, given the best atom, is a second-order polynomial minimization problem in $\alpha \in [0, 1]$ where the optimal α^* has the following closed form:

$$\alpha^* = \max(0, \min(\alpha_u, 1)) \text{ where}$$
$$\alpha_u = \frac{(\mathbf{T_u x_k} - \mathbf{y})^T (\mathbf{T_u}(\tau \mathbf{a_k} - \mathbf{x_k}))}{(\mathbf{T_u}(\tau \mathbf{a_k} - \mathbf{x_k}))^T (\mathbf{T_u}(\tau \mathbf{a_k} - \mathbf{x_k}))}. \tag{9}$$

Thus, calculations of the optimum atom and α are computationally easy problems, involving only inner products.

Lemma 1: Algorithm 1 converges in expectation and almost surely for any sequence N_k satisfying $N_k \ge 1$ for all k; i.e., let f^* be the optimum of problem (8), then

$$\lim_{k\to\infty} f(x_k) - f^* = 0, \text{ a.s.}$$

and

$$\lim_{k\to\infty} E[f(x_k)] - f^* = 0.$$

The proof of Lemma is presented in Appendix A-A.

Note that the cost function in Algorithm 1 is nonincreasing with probability 1 at each step, although the rate of descent is typically not optimal. However, as noted above, the iterations are extremely fast since they only entail computing inner products. Algorithm 1 enjoys a linear rate of convergence $\mathcal{O}(1/k)$ in expected value. More precisely, we state the following theorem for the convergence rate of Algorithm 1:

Theorem 2: Let $C_2 > 0$, $L > 0$ and $0 < s < 1$ be given constants. Let the number of samples N_k be large enough so that for any $\mathbf{x}$ with $\|\mathbf{x}\| \le \tau$ the following holds:

$$\text{Prob}\left\{ \min_{a\in S_k} \langle \nabla f(\mathbf{x}), \Upsilon^N\{a(z)\}\rangle - d^* \le 0.25 C_2/(k+L) \right\} \ge s$$

where

$$d^* = \min_{a^*\in\mathcal{A}} \langle \nabla f(\mathbf{x}), \Upsilon^N\{a^*(z)\}\rangle.$$

Then, there exists a constant C_1 so that

$$\mathbf{E}[f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*)] \le \frac{C_1}{k+L} + \frac{C_2}{k+L+1}$$

where $\mathbf{x}^*$ is an optimal solution of problem (7).

Proof: The proof and details are presented in Appendix A-C.

We note that, although Theorem 2 assumes that a specific adaptive N_k is used, drawing a fixed number of atoms at each iteration of Algorithm 1 performs well in practice.

Remark 3: It can be shown that the number N_k is finite. The proof relies on the fact that given the *a priori* information about the LTI system, the gradient vector can be bounded at each iteration in terms of the ℓ_∞ norm (or uniformly bounded for all steps by a single bound) for bounded inputs and finite data horizons. Details on the computation of N_k are given in Appendix A-B.

V. LTI Model From Multiruns With Missing Data

As mentioned in Section I, many practical problems require identifying a system from multiple runs, with unknown initial conditions, nonuniform sampling, and possibly missing data. In this section, we show that this scenario can be easily accommodated by the proposed framework.

Consider $\mathcal{P}$ different runs of an LTI system. More precisely, given $\mathcal{P}$ input/output vector pairs, we assume that, for each experiment $i = 1, 2, \ldots, \mathcal{P}$, the following noise-corrupted output is available:

$$\mathbf{y}_i = \mathbf{g}_i^{ic} + \mathbf{T_{u_i}} \mathbf{g} + \eta_i \tag{10}$$

where $\mathbf{g}_i^{ic}$ is the initial condition response for the ith run, $\mathbf{u}_i$ is the input applied at ith run, and $\mathbf{g}$ is the impulse response of the system to be identified. For simplicity of presentation and without loss of generality, we assume that all the runs have the same data horizon length N.

First of all, the initial condition response of any transfer function can be approximated with arbitrary precision by the same set of atoms defined for impulse response identification, only shifted up one sampling

time:

$$\mathbf{g}_i^{ic} \approx \sum_{a(z)\in\mathcal{A}} c_{i,a}^{ic}\,\Upsilon^N\{za(z)\} \quad \forall c_{i,a}^{ic}\in\mathbb{R}, i=1,2,\ldots,\mathcal{P}. \tag{11}$$

Hence, the problem of identifying the most parsimonious model from multirun experiments with different initial conditions can be posed as maximizing the "right" block sparsity measure. Define

$$\mathbf{c}_a \doteq \left[c_{1,a}^{ic}\ c_{2,a}^{ic}\ \ldots\ c_{1,\mathcal{P}}^{ic}\ c_a\right]^T.$$

Then, the parsimonious identification problem can be formulated as

$$\begin{aligned}
&\min_{\mathbf{c}}\ \text{cardinality}\{\mathbf{c}:\mathbf{c}_a\neq 0\}\\
&\text{s.t.}\sum_{i=1}^{i=\mathcal{P}}\|\mathbf{T}_{\mathbf{u}_i}\mathbf{g}+\mathbf{g}_i^{ic}-\mathbf{y}_i\|_{\ell_2}^2\le\eta_{\max}^2\\
&\mathbf{g}_i^{ic}=\sum_{a\in\mathcal{A}}c_{i,a}^{ic}\,\Upsilon^N\{za(z)\};\ i=1,2,\ldots,\mathcal{P}\\
&\mathbf{g}=\sum_{a\in\mathcal{A}}c_a\,\Upsilon^N\{a(z)\}.
\end{aligned}\tag{12}$$

As before, in order to develop a convex relaxation of the sparsity problem above, we define a suitable atomic norm. Let

$$\tilde{\mathbf{g}}\doteq[\mathbf{g}_1^{ic\,T}\ \mathbf{g}_2^{ic\,T}\ \ldots\ \mathbf{g}_{\mathcal{P}}^{ic\,T}\mathbf{g}^T]^T.$$

In this context, the atomic norm of interest is

$$\begin{aligned}
\|\tilde{\mathbf{g}}\|_{\tilde{\mathcal{A}}}=\inf_{\mathbf{c}}\ &\sum_{a\in\mathcal{A}}\|\mathbf{c}_a\|_\infty\\
\text{s.t. }\mathbf{g}_i^{ic}=&\sum_{a(z)\in\mathcal{A}}c_{i,a}^{ic}\,\Upsilon^N\{za(z)\};\ i=1,2,\ldots,\mathcal{P}\\
\mathbf{g}=&\sum_{a(z)\in\mathcal{A}}c_a\,\Upsilon^N\{a(z)\}
\end{aligned}$$

and a convex relaxation of the parsimonious system identification problem from multiple runs is given by

$$\begin{aligned}
&\min_{\tilde{\mathbf{g}}}\ \frac{1}{2}\sum_{i=1}^{i=\mathcal{P}}\|\left(\mathbf{T}_{\mathbf{u}_i}\mathbf{g}+\mathbf{g}_i^{ic}-\mathbf{y}_i\right)\|_{\ell_2}^2\\
&\text{subject to }\ \|\tilde{\mathbf{g}}\|_{\tilde{\mathcal{A}}}\le\tau.
\end{aligned}\tag{13}$$

Next, consider the identification problem described in here but where one has nonuniform sampling/missing samples. Formally, assume that for each of the responses, noisy measurements of $y_i(t)$ are collected at (commensurate) times

$$0\le t_{i,1}<t_{i,2}<\cdots<t_{i,m_i}=N; m_i\le N+1.$$

Without loss of generality, we assume that the sampling instants $t_{i,j}$ are integers. Define an $m_i\times n_i$ measurement matrix $\mathbf{\Omega}_i$ whose (j,k) entry is 1 if $k=t_{i,j}$ and 0 otherwise, which "extracts" the output at the measured times from the overall response for $i=1,2,\ldots,\mathcal{P}$. Then, we propose the following formulation:

$$\begin{aligned}
&\min_{\tilde{\mathbf{g}}}\ \frac{1}{2}\sum_{i=1}^{i=\mathcal{P}}\|\mathbf{\Omega}_i\left(\mathbf{T}_{\mathbf{u}_i}\mathbf{g}+\mathbf{g}_i^{ic}-\mathbf{y}_i\right)\|_{\ell_2}^2\\
&\text{subject to }\ \|\tilde{\mathbf{g}}\|_{\tilde{\mathcal{A}}}\le\tau
\end{aligned}\tag{14}$$

where $\tilde{\mathbf{g}}\doteq[\mathbf{g}_1^{ic\,T}\ \mathbf{g}_2^{ic\,T}\ \ldots\ \mathbf{g}_{\mathcal{P}}^{ic\,T}\mathbf{g}^T]^T$. Note that the proposed algorithm estimates an impulse response of the system from this multirun data.

TABLE I
CONVERGENCE RATE VERSUS NUMBER OF ATOMS CHECKED

Atoms checked per iteration	1	2	4	8	16	32	64
Approximate iterations	19 k	9 k	4 k	2 k	1 k	500	300

VI. NUMERICAL EXAMPLES

In this section, first an analysis of the effect of number of atoms used at each step on the convergence rate of the algorithm is given. Next, a selection of well-known identification methods in the literature are compared against the proposed method on a number of random synthetic examples. Due to space limitations, multiple runs examples are not provided (these can be found in [13]). Finally, the performance of the proposed method is illustrated on real data for different τ.

A. Number of Atoms Checked at Each Step Versus Convergence

To illustrate the effect of "number of random atoms checked per iteration" on the convergence rate, a sample case is presented in Table I. Algorithm 1 is run to solve the minimization problem, where the atoms are unit norm unit cardinality vectors, $f(\mathbf{x})=\|\mathbf{T}\mathbf{x}-\mathbf{y}\|_2^2$ with $\mathbf{T}\in\mathbb{R}^{30\times 64}$, and $\mathbf{y}$ is generated from a normal distribution with zero mean and unit standard deviation. The ground truth is obtained using matlab software for disciplined convex programming (CVX) [14]. For each trial, the algorithm was started at zero initial conditions and terminated when the relative error with respect to the ground truth fell below 1%. Thirty random experiments were conducted for each case. The average number of iterations is presented in Table I.

Given that in most cases, calculation of the gradient vector at a particular step is considerably more expensive than checking a random atom with the calculated gradient (a dot product), a reasonable approach is to check multiple atoms with the gradient at a particular iteration, as suggested by Table I.

B. Comparison of ID Methods on Random Synthetic Examples

In this section, the atomic norm minimization approach with gridding, i.e., discretized atomic soft thresholding (DAST) [9], subspace ID [15], prediction error estimate (PEM) [16], and the proposed method are compared on random synthetic examples. DAST is implemented using a uniformly spaced discrete net of the unit disk consisting of approximately 2000 poles. In order to present a fair comparison, ground truth bounds on the atomic norms are supplied to DAST (the true bound on the ℓ_1 norm in the problem [9] can be calculated based on the partial fraction of the true transfer function) and to the proposed algorithm, and the correct model order is fed to subspace ID and PEM. Additionally, the model identified by subspace ID is used as an initial model for the $\text{PEM}_{\text{initialzied}}$ method. One hundred experiments were run with randomly picked stable LTI systems and orders ranging from 1 to 10 as a base comparison. The data horizon was chosen randomly between 50 and 150. Pseudorandom binary signals (PRBS) input was applied, and the output was corrupted by additive noise bounded by 5% of the peak absolute value the output. The statistics are summarized in Table II. For each experiment, the estimated impulse response is placed in a Hankel matrix and the tail sum of normalized singular values vector, i.e., $\sum_{i>n}\sigma(i)/\sigma(1)$, is given as sparsity measure, where n is the LTI system order. Note that this measure is identically 0 for the ground truth, subspace ID, $\text{PEM}_{\text{initialized}}$, and PEM. Next, the identification is carried out on lightly damped systems and/or systems with slowly decaying modes. Keeping everything else the same as in

TABLE II
BASE COMPARISON STATISTICS

	Estimation Error			Sparsity Measure		
	Min	Avr	Max	Min	Avr	Max
Proposed method	0.11	6.39	28.81	≈ 0	0.19	0.86
DAST	0.15	7.30	50.73	≈ 0	0.22	2.34
Subspace ID	0.41	7.82	54.92	0.00	0.00	0.00
$\text{PEM}_{\text{initialized}}$	0.42	6.58	38.65	0.00	0.00	0.00
PEM	0.42	1.185 e+9	8.64 e+10	0.00	0.00	0.00

TABLE III
COMPARISON STATISTICS FOR LIGHTLY DAMPED SYSTEMS

	Estimation Error			Sparsity Measure		
	Min	Avr	Max	Min	Avr	Max
Proposed method	1.55	9.57	34.78	≈ 0	0.07	0.4
DAST	2.28	29.44	65.35	≈ 0	0.07	0.53
Subspace ID	2.82	29.61	110.97	0.00	0.00	0.00
$\text{PEM}_{\text{initialized}}$	0.20	12.20	105.91	0.00	0.00	0.00
PEM	0.26	2.9 e+39	2.9 e+41	0.00	0.00	0.00

the base comparison, for this set the plants are randomly chosen to have at least one pole with $0.98 \leq |p| \leq 1$. Results are given in Table III.

For the base case (Table II), our method has significantly better performance than DAST, Subspace ID, $\text{PEM}_{\text{initialized}}$, and PEM. The advantages of the proposed algorithm are even more evident in the case where the system to be identified has slow decaying modes (see Table III). A possible explanation for the better performance with respect to DAST are the facts that our algorithm is not constrained to have the poles in a specific net and that its scaling is much better suited for short runs. As for the comparison with subspace ID and PEM, our method had both the advantage of using *a priori* information on the allowable set for the poles and being able to better deal with measurements with a significant amount of noise. Regarding PEM, note that, in this set of experiments, it outperforms the proposed method only in the best case scenario, even when provided with the model identified by subspace ID.

Before closing this section, we would like to provide a few remarks comparing the proposed algorithm against augmented Lagrangian method (ALM) with Hankel nuclear norm minimization [17]. In this comparison, we highlight the run-time and the computational complexity of the proposed algorithm and its ability to handle very large data sets. In this example, the data horizon length is increased from 100 to 1000 by steps of 50, and at each data horizon 10 random experiments using impulse response data are conducted. The LTI systems are chosen randomly, with maximum order 10, and 10% noise is added. The ALM formulation is chosen as

$$\min_{\mathbf{h}} \|(S\mathbf{h})\|_* + \frac{\lambda}{2}\|\mathbf{h} - \mathbf{d}\|_{\ell_2}^2$$

where $S\mathbf{h}$ is the Hankel matrix associated with vector $\mathbf{h}$, λ is the parameter penalizing misfit to measurement, and $\mathbf{d}$ is the measured impulse response. Due to the lack of an intuitive way to choose a "ground truth" λ in the above formulation, the following procedure is employed: The ground truth $\mathbf{h}$ is inserted into the objective function and the values of the nuclear norm term and the data fidelity term are observed. The parameter λ is chosen such that, for the ground truth $\mathbf{h}$, both terms are very close to each other.

The results shown in Fig. 1 are expected. The complexity of the ALM algorithm is dominated, eventually, by the singular value decomposition (SVD) operation, making it practical only for medium size identification problems. On the other hand, the proposed algorithm scales gracefully with the data size.

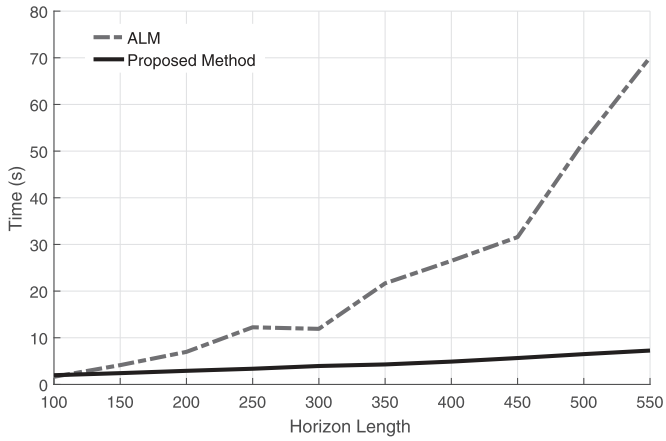

Fig. 1. Run time versus data size comparison.

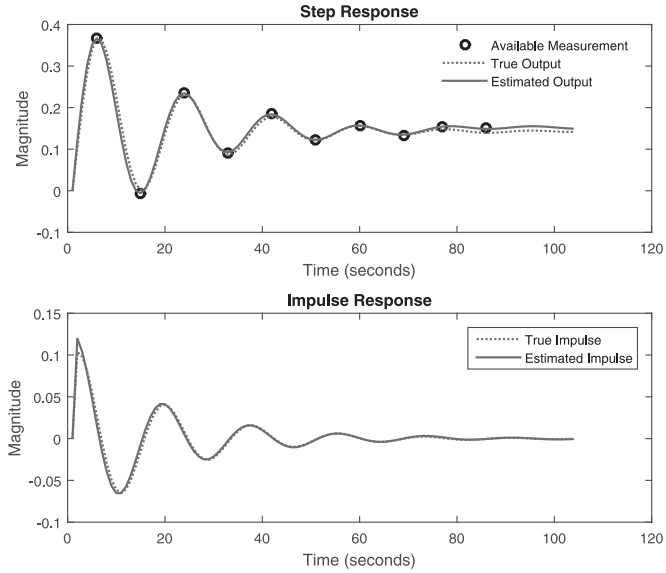

Fig. 2. Proposed method with missing data.

C. Comparison of ID Methods on Missing Measurement Example

In this section, the available methods for missing measurements with the proposed method is compared on an example. MATLAB's system identification toolbox provides two different methods for such data sets: the function "misdata" that uses either a known model or default order state space model to estimate the missing part of the data by minimizing output prediction errors, and the function "merge" that combines small isolated clusters. To compare the proposed method and these functions, a second-order system ($g(z) = (0.1037z - 0.08657)/(z^2 - 1.78z + 0.9)$) is employed. The step response of the system is contaminated with $\mathcal{N}(0, 0.02)$ and measured at the local extrema. The proposed method results are shown in Fig. 2. The function "merge" cannot be used with this data since we do not have isolated clusters. The function "misdata" requires more data points to closely interpolate the missing measurements. Note that as the number of available measurements are gradually increased, the misdata function correctly interpolates the whole data.

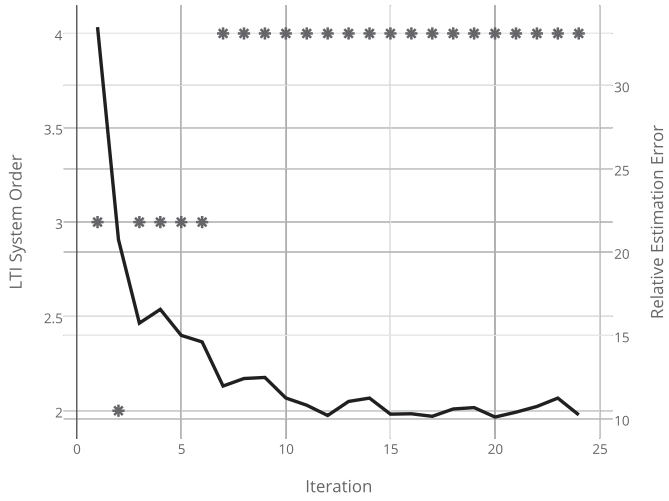

Fig. 3. Error and order for different τ.

D. Example Using Experimental Data

The proposed algorithm is implemented on a benchmark problem from DaISy [18]. The experimental data for a hair dryer setup is provided as an input/output vector of data horizon 1000. The first half of the data is used for identification and the second half is used for validation. Results are compared with Subspace ID. Fig. 3 shows the order of the system identified versus output estimation error as τ is changed. The initial value for τ is chosen as 0.5, and the heuristic given in [11, Sec. 4.5.4] is run for 25 steps for illustration purposes, at the end of which τ reaches the value 1.1. For the model-order calculation, the estimated impulse response was used to form a Hankel matrix and the vector of singular values was computed. The first index for which the cumulative sum of this vector exceeded 99% of its total was chosen as the system order. Subspace ID achieved 1.9% identification error and 2.4% validation error. The proposed method achieved 1.8% identification error and 1.75% validation error.

VII. Conclusion and Future Work

In this paper, we consider the problem of identifying LTI models from corrupted input/output measurements. By using a sparsity inducing norm, i.e., atomic norm, we propose a method that is robust to noise and missing information and that promotes low-order models as solution to the identification problem. Comparing against competing system identification methods of similar nature, we showed improved performance over a statistically meaningful set of random trials. Possible future research directions include extending the results to more complex settings, e.g., multi-input multi-output (MIMO) LTI system identification and Wiener system identification.

Appendix A
Convergence Results and Complexity

A. Supermartingale:

Definition 4: [19, Ch. 12.1] Let Y be a sequence of a random variable. Then, a discrete time *supermartingale* satisfies $E[Y_{k+1}|Y_k,\dots,Y_0] \le Y_k$ and $E[Y_{k+1}^-] \le \infty$ where $Y^- = -\min(0, Y)$.

Proof of Lemma 1: Define the random variable $Y_k = f(x_k) - f(x^*)$. Assume that x^* is the optimum, i.e., $f(x)$ attains its global minimum at x^*. Since the objective function is convex, at each iteration, the inequality $Y_k \ge 0$ holds for all k. Algorithm 2 also satisfies, for any k—$Y_{k+1} \le Y_k$. Taking the conditional expectation yields:

$$E[Y_{k+1}|Y_k,\dots,Y_1,Y_0] \le Y_k.$$

Therefore, we have a *supermartingale* and this proves the convergence in expectation and almost surely.

B. Bounding Deviation From Optimal Descent:

Given the gradient vector $\nabla f_k = [df_{-1}, df_0, df_1, \dots, df_{N-2}]^T$ at iteration k of the randomized algorithm, the corresponding optimal pole p_k and a pole p^* satisfying $|p_k - p^*| \le \epsilon$, we want to find a bound of the form

$$|p_k - p^*| \le \epsilon \Rightarrow$$
$$d = |\alpha_{p_k} \langle \nabla f_k, \Upsilon^N\{a_{p_k}(z)\}\rangle - \alpha_p^* \langle \nabla f_k, \Upsilon^N\{a_{p^*}(z)\}\rangle| \le \gamma$$

where $\gamma = \frac{C_1}{k+L} + \frac{C_2}{k+L+1}$. Here, $\Upsilon^N\{a_{p^*}(z)\}$ represents the unscaled N-length impulse response of the atom generated by the pole p^* and $\alpha_p^* > 0$ is the scaling associated with pole p^* (similar definitions for p_k).

We consider the atoms in $\mathcal{A}_1$. The other sets can be treated in a similar fashion. The impulse response $\Upsilon^N\{a_{p^*}(z)\}$ is given by $\Upsilon^N\{a_{p^*}(z)\} = [0, 1, \Re(p^*), \Re(p^{*2}), \dots, \Re(p^{*N-2})]^T$. We consider two cases in writing the deviation d from the optimum descent.

$$\text{Case 1: } \alpha_{p_k} < \alpha_p^* = \alpha$$
$$d = \alpha_p^* \langle \nabla f_k, \Upsilon^N\{a_{p^*}(z)\}\rangle - \alpha_{p_k} \langle \nabla f_k, \Upsilon^N\{a_{p_k}(z)\}\rangle$$
$$\le \alpha \left(\langle \nabla f_k, \Upsilon^N\{a_{p^*}(z)\}\rangle - \langle \nabla f_k, \Upsilon^N\{a_{p_k}(z)\}\rangle \right).$$
$$\text{Case 2: } \alpha_p^* \le \alpha_{p_k} = \alpha$$
$$d = \alpha_{p_k} \langle \nabla f_k, \Upsilon^N\{a_{p_k}(z)\}\rangle - \alpha_p^* \langle \nabla f_k, \Upsilon^N\{a_{p^*}(z)\}\rangle$$
$$\le \alpha \left(\langle \nabla f_k, \Upsilon^N\{a_{p_k}(z)\}\rangle - \langle \nabla f_k, \Upsilon^N\{a_{p^*}(z)\}\rangle \right).$$

Both cases follow the same steps; hence, we only present one of them. Since the norm of the gradient is uniformly bounded, i.e., $\|\nabla f_k\|_\infty \le M$ for all k, we proceed as follows:

$$|d| \le \alpha \left| {\nabla f_k}^T \Upsilon^N\{a_{p_k}(z)\} - {\nabla f_k}^T \Upsilon^N\{a_{p^*}(z)\} \right|$$
$$\le \alpha \left| {\nabla f_k}^T \left(\Upsilon^N\{a_{p_k}(z)\} - \Upsilon^N\{a_{p^*}(z)\} \right) \right|$$
$$\le \alpha \left| \sum_{r=1}^{N-2} df_r \left(\Re(p^r) - \Re(p^{*r}) \right) \right|$$
$$\le \alpha \sum_{r=1}^{N-2} \left| df_r \left(\Re(p^r) - \Re(p^{*r}) \right) \right|$$
$$\le \alpha M \sum_{r=1}^{N-2} |\Re(p^r) - \Re(p^{*r})| \le \alpha M \sum_{r=1}^{N-2} |p^r - p^{*r}|.$$

Looking at a generic term above, i.e., $|p^r - p^{*r}|$, we can write

$$|p^r - p^{*r}| = |p - p^*| \left| p^{r-1} + p^{r-2}p + \dots pp^{*r-2} + p^{*r-1} \right| \le \epsilon r \rho^{r-1}.$$

Replacing each $|p^r - p^{*r}|$ term in the previous inequality with its upper bound found above yields

$$\alpha M \sum_{r=1}^{N-2} |p^r - p^{*r}| \le \alpha M \epsilon \left\{ \frac{1 - (N-1)\rho^{N-2}}{1-\rho} + \frac{\rho - \rho^{N-1}}{(1-\rho)^2} \right\}.$$

The final step stems from the fact that $\sum_{r=1}^{N-2} r\rho^{r-1}$ can be written as the derivative of the geometric series $\sum_{r=1}^{N-2} \rho^r$. Hence, for a finite

horizon N and radius ρ, the upper bound on the deviation from optimum descent is

$$|d| \leq \alpha_{\max} M \epsilon \sum_{r=1}^{N-2} r\rho^{r-1} \tag{15}$$

where $\alpha_{\max}$ is the maximum attained by the scaling function.

Second, we will show how many poles N_k need to be drawn in each iteration to ensure that the inequality $d \leq \gamma$ holds with probability $\geq s$, or, equivalently, $|p_k - p^*| \leq \epsilon$ holds with probability $\geq s$.

Given a fixed pole p^*, define an event Ω as follows: out of N_k poles uniformly drawn from $\mathbb{D}_\rho$, at least one pole satisfies $|p - p^*| \leq \epsilon$. For the case $\inf_{|p|=\rho} |p^* - p| \geq \epsilon$, the probability of this event, s, is equal to the the ratio of area of $\epsilon-$ ball to the area of $\mathbb{D}_\rho$:

$$s = 1 - \overline{s} = 1 - \left(1 - \frac{\epsilon^2}{\rho^2}\right)^{N_k}$$

where $\overline{s} = 1 - s$. The worst case probability is smaller though, and is realized when p^* is on the boundary of $\mathbb{D}_\rho$. In this case, the probability s is the ratio of area of $\epsilon-$ ball in $\mathbb{D}_\rho$ to the area of $\mathbb{D}_\rho$. The analytic expression for this ratio is given by [20, circle–circle intersection]:

$$F = \frac{1}{\pi}\left[\zeta^2 \cos^{-1}\left(\frac{\zeta}{2}\right) + \cos^{-1}\left(1 - \frac{\zeta}{2}\right)\right] - \frac{\zeta\sqrt{4-\zeta^2}}{2} \tag{16}$$

where $\zeta = \epsilon/\rho$ for $\epsilon < \rho$. With this definition, the worst case probability is given by $s = 1 - F^{N_k}$. Therefore, for any given ϵ and probability s, the finite number of required poles to be drawn is explicitly given by $N_k(\epsilon, s) = \ln(1-s) - \ln F$.

C. Proof of Theorem 2:

Definition 5: Curvature Constant [21]: The curvature constant C_f of a convex and differentiable objective function $f : \mathbb{R}^n \to \mathbb{R}$ in the set $\|x\|_{\mathcal{A}} \leq \tau$ is defined as follows:

$$C_f = \sup_x \frac{f(x') - f(x) - \langle \nabla f(x), x' - x\rangle}{\alpha^2}$$
$$\text{s.t. } \|x\|_{\mathcal{A}} \leq \tau, \ \ \|\tau a\|_{\mathcal{A}} \leq \tau, \ \ x' = x + \alpha(\tau a - x), \ \alpha \in [0,1]. \tag{17}$$

Definition 5 readily implies the following for $f(x)$:

$$f(x_k + \alpha(\tau a_k - x_k)) \leq f(x_k) + \alpha\langle \nabla f(x_k), (\tau a_k - x_k)\rangle + \alpha^2 C_f$$

which is true for all k. Suppose that number of atoms drawn are such that at each iteration, event Ω is realized with probability s, guaranteeing $\langle \nabla f(x_k), a_k\rangle - \langle \nabla f(x_k), a_k^*\rangle \leq \gamma_k$ for $k \geq 0$. Define the primal error as $h(x_k) = f(x_k) - f(x^*) \geq 0$ and subtract $f(x^*)$ from both sides of the above inequality:

$$\begin{aligned} h(x_{k+1}) &\leq h(x_k) + \alpha\langle \nabla f(x_k), (\tau a_k^* - x_k)\rangle \\ &\quad + \alpha\gamma_k + \alpha^2 C_f \ \rightarrow \ \text{Prob } s \\ h(x_{k+1}) &\leq h(x_k) \rightarrow \text{Prob } > 1 - s. \end{aligned}$$

Hence, we can write

$$\begin{aligned} \mathbf{E}\left[h(x_{k+1}) | \, x_k\right] &\leq s(\alpha\langle \nabla f(x_k), (\tau a_k^* - x_k)\rangle + \alpha\gamma_k + \alpha^2 C_f) \\ &\quad + h(x_k). \end{aligned}$$

Note that $\langle \nabla f(x_k), (\tau a_k^* - x_k)\rangle \leq -h(x_k)$, since any linear approximation to a convex function at any point is a lower bound to the function over the domain (weak duality), which implies

$$\mathbf{E}\left[h(x_{k+1}) | \, x_k\right] \leq h(x_k) + s(-\alpha h(x_k) + \alpha\gamma_k + \alpha^2 C_f).$$

Choosing $\gamma_k = \alpha_k C_f$ and taking expectation on both sides of above inequality yields

$$\mathbf{E}\left[h(x_{k+1})\right] \leq (1 - s\alpha_k)\mathbf{E}\left[h(x_k)\right] + 2s\alpha_k^2 C_f.$$

Choose $\alpha_k = \frac{2}{s(k+k_0+2)}$, where k_0 is the smallest integer such that $k_0 \geq \frac{2}{s} - 2$ holds true for given probability s. We claim that

$$\mathbf{E}\left[h(x_{k+1})\right] \leq \frac{k_0 \mathbf{E}\left[h(x_0)\right]}{k + k_0 + 2} + \frac{8C_f}{s(k+k_0+3)}.$$

Base case with $k = 0$ is straightforward using weak duality as follows:

$$\begin{aligned} \mathbf{E}\left[h(x_1)\right] &\leq \left(1 - s\frac{2}{s(k_0+2)}\right)\mathbf{E}\left[h(x_0)\right] + 2s\left(\frac{2}{s(k_0+2)}\right)^2 C_f \\ &\leq \frac{k_0}{k_0+2}\mathbf{E}\left[h(x_0)\right] + \frac{8C_f}{s}\frac{1}{(k_0+2)(k_0+2)} \\ &\leq \frac{k_0}{k_0+2}\mathbf{E}\left[h(x_0)\right] + \frac{8C_f}{s}\frac{1}{(k_0+3)}. \end{aligned}$$

The last inequality follows from the fact that $(k_0+2)^2 \geq (k_0+3)$ for $k_0 \geq 0$. Next, we show that if the inequality holds for k, it also holds for $k+1$, thus proving the theorem by induction. Weak duality for $\mathbf{E}\left[h(x_{k+1})\right]$ implies

$$\begin{aligned} \mathbf{E}\left[h(x_{k+1})\right] &\leq \left(1 - s\frac{2}{s(k+k_0+2)}\right)\mathbf{E}\left[h(x_k)\right] \\ &\quad + 2s\left(\frac{2}{s(k+k_0+2)}\right)^2 C_f \\ &\leq \left(\frac{k+k_0}{k+k_0+2}\right)\mathbf{E}\left[h(x_k)\right] + \frac{8C_f}{s}\left(\frac{1}{k+k_0+2}\right)^2 \\ &\leq \left(\frac{k+k_0}{k+k_0+2}\right)\left(\frac{k_0\mathbf{E}\left[h(x_0)\right]}{k+k_0+1} + \frac{8C_f}{s(k+k_0+2)}\right) \\ &\quad + \frac{8C_f}{s}\left(\frac{1}{k+k_0+2}\right)^2 \\ &\leq \frac{k_0\mathbf{E}\left[h(x_0)\right]}{k+k_0+2} + \left(\frac{k+k_0+1}{(k+k_0+2)^2}\right)\frac{8C_f}{s} \\ &\leq \frac{k_0\mathbf{E}\left[h(x_0)\right]}{k+k_0+2} + \left(\frac{1}{k+k_0+3}\right)\frac{8C_f}{s} \quad . \end{aligned}$$

The last step in the above derivation stems from the fact that $(n-1)(n+1) < n^2$, applied for $n = k + k_0 + 2$. Note that the derived inequality is in the form

$$\mathbf{E}\left[f(x_{k+1}) - f(x^*)\right] \leq \frac{C_1}{k+L} + \frac{C_2}{k+L+1}$$

with $C_1 = k_0\mathbf{E}\left[h(x_0)\right]$, $C_2 = 8C_f/s$, $L = k_0 + 2$, $x^* \in \|x\|_{\mathcal{A}} \leq \tau$ an optimal solution to the optimization problem.

REFERENCES

[1] R. S. Sánchez-Peña and M. Sznaier, *Robust Systems Theory and Applications*. New York, NY, USA: Wiley, 1998.

[2] P. V. Overschee and B. De Moor, "N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems," *Automatica*, vol. 30, no. 1, pp. 75–93, 1994.

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

[4] T. Ding, M. Sznaier, and O. I. Camps, "A rank minimization approach to video inpainting," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[5] Z. Liu and L. Vandenberghe, "Semidefinite programming methods for system realization and identification," in *Proc. IEEE 48h Conf. Decis. Control Held Jointly 28th Chin.*, Dec. 2009, pp. 4676–4681.
[6] M. Ayazoglu and M. Sznaier, "An algorithm for fast constrained nuclear norm minimization and applications to systems identification," in *Proc. 51st IEEE Conf. Decision Control*, 2012, pp. 3469–3475.
[7] Z. Liu and L. Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," *SIAM J. Matrix Anal. Appl.*, vol. 31, no. 3, pp. 1235–1256, Jan. 2010.
[8] M. Sznaier, M. Ayazoglu, and T. Inanc, "Fast structured nuclear norm minimization with applications to set membership systems identification," *IEEE Trans. Autom. Control*, vol. 59, no. 10, pp. 2837–2842, Oct. 2014.
[9] P. Shah, B. N. Bhaskar, G. Tang, and B. Recht, "Linear system identification via atomic norm regularization," in *Proc. IEEE 51st Annu. Conf. Decision Control*, Dec. 2012, pp. 6265–6270.
[10] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, Oct. 2012.
[11] B. Yılmaz, "Sparsity Based Methods In System Identification," Ph.D. dissertation, Dept. Elect. Comput. Eng., , Northeastern Univ., Boston, MA, USA, 2015.
[12] A. Tewari, P. K. Ravikumar, and I. S. Dhillon, "Greedy algorithms for structurally constrained high dimensional problems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 882–890.
[13] K. Bekiroglu, "From Data to Interventions: Using System Identification and Robust Control Algorithms to Design Effective Treatments," Ph.D. dissertation, Nanyang Technol. Univ., Singapore, 2015.
[14] M. Grant and S Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," 2014. [Online]. Available: http://cvxr.com/cvx
[15] P. van Overschee and B. L. R. de Moor, *Subspace Identification for Linear Systems: Theory, Implementation, Applications.* Norwell, MA, USA: Kluwer, vol. 1, 1996.
[16] L. Ljung, "System Identification Toolbox User's Guide R 2013 b," MathWorks, Natick, MA, USA, 2013.
[17] M. Ayazoglu, M. Sznaier, and O. I. Camps, "Fast algorithms for structured robust principal component analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1704–1711.
[18] B. D. Moor, P. D. Gersem, B. D. Schutter, and W. Favoreel, "Daisy: A database for identification of systems," *J. A*, vol. 38, pp. 4–5, 1997.
[19] G. Grimmett and D. Stirzaker, *Probability and Random Processes*, 3rd ed. New York, NY, USA: Oxford, 2001.
[20] E. W. Weisstein, *CRC Concise Encyclopedia of Mathematics, 2nd ed*. Boca Raton, FL, USA: CRC Press, 2002.
[21] M. Jaggi, "Revisiting Frank-Wolfe: Projection-free sparse convex optimization," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 427–435.