A Super-Atomic Norm Minimization Approach to Identifying Sparse Dynamical Graphical Models

Y. Wang M. Sznaier O. Camps

***Abstract*— This paper considers the problem of identifying sparse dynamical graphical models from input/output data. Our main result shows that this problem can be recast into an expanded atomic-norm minimization framework that allows for enforcing block-sparsity. This approach leads to efficient algorithms capable of handling large data sets, unknown inputs and fragmented data records. These results are illustrated with several examples.**

I. INTRODUCTION

Recently, considerable attention has been devoted to the problem of identifying dynamical graphical models, represented by a graph structure $G = \{V, E\}$, where the vertices are associated with time series and the edges relate the values of these series at different time instants. These models appear in fields ranging from systems biology and chemistry to economics and video-analytics.

In general, the problem is ill posed, since an infinite number of topologies can explain a given set of finite, noisy observations. Thus, typically a "sparsity'" prior is added to regularize the problem, encapsulating the fact that usually the solution with the fewest number of links is the correct one. An example of this situation is when graphs are used to encapsulate causal relationships between agents and predict future behavior, exploiting the concept of Granger causality [1]. Several approaches have been proposed to solve the resulting problem. A cycling descent algorithm was adopted in [2] that directly attempted to enforce sparsity and used causal Laguerre basis functions to model the connections between two time series. Since attempting to directly enforce sparsity leads to non-convex problems, a large portion of the existing literature uses the ℓ_1 norm as a convex surrogate for sparsity, leading to a number of convex optimization based algorithms [3], [4], [5], [6], [7].

ℓ_1 regularization based approaches tend to enforce sparsity in terms of the total number of non-zero coefficients involved in representing the graph. Thus, they do not, typically, lead to sparse topologies, since the latter requires enforcing *block-sparsity*, that is, all coefficients of the model associated with a given edge should be zero simultaneously. This observation motivated the use of group lasso based approaches [8], [9], [10]. While these methods usually work well, in some cases they fail to produce the sparsest topology, motivating the introduction of re-weighted iterative algorithms [11], [12], [13]. These algorithms work well in practice, but the use of a sum-of-ℓ_2 norms objective function leads to second-order cone programs, whose complexity is larger than n^3.

An alternative to the approaches above is given by orthogonal matching pursuit type algorithms. Cycling Orthogonal Least Squares (COLS)[12] seeks to find sparse solutions using a modified Orthogonal Least Squares algorithm. A Block Orthogonal Matching Pursuit algorithm has also been investigated [14], using a notion of coherence analog to the one proposed in the context of compressive sensing [15]. A further extension to cases where the blocks have different sizes was presented in [16]. At the present time, these approaches cannot handle cases, often arising in practice, where the network is subject to unknown external inputs.

Finally, a Bayesian approach has been proposed to obtain sparse topologies in [17], where the problem was posed as sparse input selection for MISO LTI systems. A sparse plus low rank criterion has been discussed in [18], in which a two layer structure (manifest and latent) was assumed. Typically, Bayesian approaches require strong prior information about the system to be identified and the resulting algorithms have relatively high computational cost.

As an alternative to the approaches above, inspired by recent progress in obtaining sparse representations by exploiting the geometry of the problem, in this paper we propose a new super-atomic norm based approach. This proposed method aims at solving a similar problem as group-lasso does. However, as we show next, due to the introducing of super-atoms, we achieve substantial improvement in computational efficiency. The main contributions of the paper are:

1) Extension of the original atomic norm framework proposed in [19] for obtaining sparse solutions to sets of linear equations to the *block-sparse* case. This is accomplished by introducing the concept of super-atoms and its associated super-atomic norm, and showing that minimizing this norm indeed minimizes the convex envelope of cardinality of a set of vectors.
2) Showing that the approach above leads to efficient algorithms for minimizing functions subject to block-sparsity constraints that only require performing inner products and thus can handle large data sets.
3) Recasting the network identification framework into a constrained super-atomic norm minimization framework, which allows for directly using these algorithms.

These results are illustrated with several examples where the proposed approach compares favorably against existing ones both in terms of recovery of the underlying network and computational complexity.

This work was supported in part by NSF grants IIS–1318145 and ECCS–1404163; AFOSR grant FA9550-15-1-0392; and the Alert DHS Center of Excellence under Award Number 2013-ST-061-ED0001.

The authors are with the Department of Electrical & Computer Engineering, Northeastern University, MA 02115, USA. emails: wang.yin@husky.neu.edu, {msznaier,camps}@coe.neu.edu

II. PRELIMINARIES

A. Notation and Definitions

$\mathbf{x}, \mathbf{M}$	a vector in $\mathbb{R}^n$ (matrix in $\mathbb{R}^{m\times n}$)
$\mathbf{M}(:, \mathbf{j})$	j$^{\text{th}}$ column of matrix $\mathbf{M}$.
$\|\mathbf{x}\|_2$	ℓ_2 norm of a vector: $\|\mathbf{x}\|_2^2 = \sum_i x_i^2$
$\|\mathbf{x}\|_0$	ℓ_0 quasi-norm, number of non-zero elements in $\mathbf{x}$
$\|\mathbf{x}\|_\infty$	ℓ_∞ norm, $\|\mathbf{x}\|_\infty \doteq \max_i \|x_i\|$
conv($\mathcal{A}$)	Convex hull of the set $\mathcal{A}$.
$\|E\|$	cardinality (e.g. number of elements) of the set E.

B. Atomic Norms and Sparsity

Let $\mathcal{A}$ be a centrally symmetric collection of atoms. Its atomic norm, $\|\mathbf{x}\|_\mathcal{A}$ is defined as [19]:

$$\|\mathbf{x}\|_\mathcal{A} = \inf\{t > 0 \ : \ \mathbf{x} \in t \ \text{conv}(\mathcal{A})\} \tag{1}$$

It can be easily shown that an equivalent definition is

$$\|\mathbf{x}\|_\mathcal{A} = \inf\left\{ \sum_{a\in\mathcal{A}} |c_a| \ : \mathbf{x} = \sum_{a\in\mathcal{A}} c_a a \right\} \tag{2}$$

As shown in [19], atomic norms play a key role when searching for sparse solutions to problems of the form:

$$\min_x f(x) \text{ subject to } \|x\|_\mathcal{A} \le \tau \tag{3}$$

where the atomic norm constraint is used to encourage sparsity. Further, as shown in [20], this problem can be efficiently solved using the following Frank-Wolfe type algorithm, which has a convergence rate of $\mathcal{O}(\frac{1}{t})$.

Algorithm 1 Generic Frank-Wolfe algorithm to minimize a convex function over the τ-scaled atomic norm ball

1: $x_0 \leftarrow \tau a_0$ for arbitrary $a_0 \in \mathcal{A}$ ▷ Initialization
2: **for** t = 0,1,2,3,... **do**
3: $a_t \leftarrow \text{argmin}_{a\in\mathcal{A}} \langle \partial f(x_t), a\rangle$
4: $\alpha_t \leftarrow \text{argmin}_{\alpha\in[0,1]} f(x_t + \alpha[\tau a_t - x_t])$
5: $x_{t+1} \leftarrow x_t + \alpha_t[\tau a_t - x_t]$
6: **end for**

C. Problem Statement

As indicated in the introduction, in this paper we consider models represented by a directed graph $G = \{V, E\}$ structure, where each node V corresponds to a given time series, and the edges E connecting these are linear shift invariant operators. The corresponding equations are given by

$$x_j(t) = \sum_{i=1}^n \sum_{k=1}^r c_{j,i}(k) x_i(t-k) + \eta_j(t), \quad t \in [r+1, T], \ j = 1, \dots, n \tag{4}$$

where $x_j(.)$ denotes the time series at the j^{th} node, $c_{j,i}(.)$ are the coefficients of an ARX model relating the present value of the time series at node j to the past values measured at node i, and $\eta_j(t)$ represents measurement noise. For ease of notation, define

$$\begin{aligned}
\mathbf{x}_j &\doteq \big[x_j(T), \dots, x_j(r+1)\big]^T \\
\boldsymbol{\eta}_j &\doteq \big[\eta_j(T), \dots, \eta_j(r+1)\big]^T \\
\mathbf{c}_{j,i} &\doteq \big[c_{j,i}(1), \dots, c_{j,i}(r)\big]^T \\
\mathbf{c}_j &\doteq \big[\mathbf{c}_{j,1}^T \dots, \mathbf{c}_{j,n}^T\big]^T \\
\mathbf{C} &\doteq \big[\mathbf{c}_1, \dots, \mathbf{c}_n\big] \\
\mathbf{X} &\doteq \big[\mathbf{x}_1, \dots, \mathbf{x}_n\big] \\
\mathbf{H}_i &\doteq \begin{bmatrix} x_i(T-1) & x_i(T-2) & \dots & x_i(T-r) \\ x_i(T-2) & x_i(T-3) & \dots & x_i(T-r-1) \\ \vdots & \dots & \dots & \vdots \\ x_i(r) & \dots & \dots & x_i(1) \end{bmatrix} \\
\mathbf{H} &\doteq \big[\mathbf{H}_1 \ \dots \ \mathbf{H}_n\big] \\
\boldsymbol{\Xi} &\doteq \big[\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n\big]
\end{aligned}$$

With this notation, the equations describing the complete model can be written in compact form as:

$$\mathbf{X} = \mathbf{H}\mathbf{C} + \boldsymbol{\Xi} \tag{5}$$

Our goal is to identify models of the form (5) from experimental data and a-priori information about the noise and the order of the edge systems. As noted in the introduction, due to the finite data record and the presence of noise, this problem is ill–posed, admitting infinite solutions. However, in the absence of additional information, amongst all these solutions, the sparsest one, in the sense of having the smallest number of edges, is often the most desirable. Thus, we will add a "sparsity" prior leading to the following regularized problem:

Problem 1: Given T measurements of n time series $x_i(t),\ i = 1, \dots, n,\ t \in [1, T]$, and upper bounds ϵ and r on the noise level and edge model order, respectively, solve:

$$\begin{aligned} \min \textstyle\sum_j \sum_i \|\{\mathbf{c}_{j,i}\}\|_0 \text{ s. t. (5) and } \|\boldsymbol{\eta}_j\|_2 \le \epsilon, \\ \forall j = 1, \dots, n \end{aligned} \tag{6}$$

where, $\mathbf{c}_{j,i} \in \mathbb{R}^r$.

Thus, the objective function in this problem is precisely $|E|$, the number of edges in the graph.

Note that due to its structure, the problem above decouples into n subproblems of the form:

Subproblem 1:

$$\begin{aligned} &\min \|\{\mathbf{c}_i\}\|_0 \text{ s. t. } \|\boldsymbol{\eta}_j\|_2 \le \epsilon \text{ and} \\ &\mathbf{x}_j = \textstyle\sum_i \mathbf{H}_i \mathbf{c}_i + \boldsymbol{\eta}_j \end{aligned} \tag{7}$$

where, by a slight abuse of notation, we have defined $\|\{\mathbf{c}_i\}\|_0$ as the number of non-zero vectors in the set $\{\mathbf{c}_{j,i}\}$ given j.

III. SUPER ATOMS AND BLOCK SPARSITY

The class of problems considered in this paper require enforcing *block-sparsity*, rather than sparsity. This will be accomplished by considering *super-atoms* and the associated super-atomic norm, rather than the traditional atomic norms. Assume that the set $\mathcal{A}$ can be partitioned into N centrally symmetric subsets $\mathcal{A}_i$ such that $\mathcal{A} = \cup_i \mathcal{A}_i$ and $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset,\ i \ne j$. In the sequel, we will refer to the sets $\mathcal{A}_i$ as *super-atoms*. Further, to each super-atom $\mathcal{A}_i = \{a_{i,1}, ..a_{i,n_i}\}$ we

will associate the matrix $\mathbf{A}_i$ having as its j$^{\text{th}}$ column $\mathbf{a}_{i,j}$, the coordinates of the atom $a_{i,j}$ in a suitable basis in X.

Definition 1: Given a set of super-atoms $\{\mathcal{A}_i\}$ and a point $\mathbf{x} \in X$, its super-atomic norm is defined as:

$$\|\mathbf{x}\|_{s\mathcal{A}} \doteq \inf \left\{ \tau > 0 \colon \mathbf{x} = \sum_i (\tau \mathbf{A}_i)\mathbf{c}_i \text{ and } \sum_i \|\mathbf{c}_i\|_\infty = 1 \right\} \tag{8}$$

Note that the definition above reduces to the usual atomic norm definition when $\mathcal{A}_i = \{a_i\}$. This connection and the connection with block-sparsity is highlighted by the following easily shown result:

Lemma 1:

$$\|\mathbf{x}\|_{s\mathcal{A}} = \begin{array}{l} \min_{\mathbf{c}} \sum_{i=1}^N \|\mathbf{c}_i\|_\infty \text{ s.t} \\ \mathbf{x} = \sum_i \mathbf{A}_i \mathbf{c}_i \end{array} \tag{9}$$

Remark 1: Recall that, given a vector sequence $\{\mathbf{c}\}$, $\|\mathbf{c}_i\|_\infty \leq 1$, the convex envelope (e.g. the tightest convex relaxation) of its cardinality is given by [21]:

$$\|\{\mathbf{c}\}\|_{0,env} = \sum_i \|\mathbf{c}_i\|_\infty$$

Thus, from Lemma 1, it follows that, minimizing the super-atomic norm is a good surrogate leading to *block-sparse* representations, a key property that we will exploit in this paper.

IV. Sparse Network Identification via Super-Atomic Norm Minimization

From the results in the previous section, it follows that Problem 1 can be recast into a collection of n super-atomic norm minimizations of the form

$$\min \|\mathbf{z}\|_{s\mathcal{A}} \text{ subject to } \|\mathbf{x}_j - \mathbf{z}\|_2 \leq \epsilon \tag{10}$$

by simply defining each super-atom as the collection of columns from the matrices $\mathbf{H}_i$, (e.g a collection of vectors, each containing delayed measurement of the respective time-series):

$$\mathcal{A}_i = \{\mathbf{H}_i(:,t)\},\ t = 1, \dots r$$

The problem above is convex and thus can be solved for instance using interior point methods. However, while these methods work well for moderate size problems, their poor scaling properties render them impractical as the size of the data grows. Thus, in this paper, rather than solving (10), we will solve the related problem

$$\min \|\mathbf{x}_j - \mathbf{z}\|_2 \text{ subject to } \|\mathbf{z}\|_{s\mathcal{A}} \leq \tau \tag{11}$$

that is, we will impose soft, rather than hard constraints on the fitting error. The advantage of the formulation (11) is that it is amenable to be solved by the following extension of Algorithm 1:

Algorithm 2 convex minimization subject to super-atomic norm constraints

1: Data: set of super-atoms $\mathcal{A} = \{\mathcal{A}_1, \dots, \mathcal{A}_i, \dots\}$
2: Initialize $\mathbf{z}^{(0)} \leftarrow \tau \mathbf{a}$ for some arbitrary $\mathbf{a} \in \mathcal{A}$
3: **for** k = 0,1,2,3,..., k_{max} **do**
4: $L \leftarrow \arg\min_m \left\{ \min_{\|\mathbf{c}\|_\infty \leq 1} \langle \partial f(\mathbf{z}^{(k)}), \sum \mathbf{a}_{i,m} c_i \rangle \right.$ s.t. $\left. \mathbf{a}_{i,m} \in \mathcal{A}_m \right\}$
5: $\mathbf{c} \leftarrow \arg \min_{\|\mathbf{c}\|_\infty \leq 1} \langle \partial f(\mathbf{z}^{(k)}), \sum \mathbf{a}_{i,L} c_i \rangle$ s.t. $\mathbf{a}_{i,L} \in \mathcal{A}_L$.
6: $\mathbf{a} \leftarrow \sum_i \mathbf{a}_{i,L} c_i$
7: $\alpha_k \leftarrow \text{argmin}_{\alpha \in [0,1]} f(\mathbf{z}^{(k)} + \alpha[\tau \mathbf{a} - \mathbf{z}^{(k)}])$
8: $\mathbf{z}^{(k+1)} \leftarrow \mathbf{z}^{(k)} + \alpha_k[\tau \mathbf{a} - \mathbf{z}^{(k)}]$
9: **end for**

Steps 4–6 in the algorithm above correspond to step 3 in Algorithm 1. The first step selects the super-atom whose elements yield the largest decrease in the cost function and Steps 5 and 6 select the best linear combination of elements in this super-atom. Thus, the combination of steps 4-6 guarantees that at each step, both the objective function will improve, unless already at the optimum, that only elements from a single super-atom will be added to the solution, and that, at all times, $\|\mathbf{z}^{(k)}\|_{s\mathcal{A}} \leq \tau$. Proceeding as in [20] it can be shown that, as long as the objective function $f(.)$ is convex and smooth, the algorithm above is guaranteed to converge to the optimum, with a convergence rate of $\mathcal{O}(\frac{1}{k})$. Further, as shown below, all the steps in Algorithm 2 admit an explicit solution once being applied to solve problem (11).

Lemma 2: Let $\mathbf{A}_m$ denote the matrix having as columns the coordinates of $\mathbf{a}_{i,m}$, the elements of the super-atom $\mathcal{A}_m$, and assume that the super-atoms are centrally symmetric, that is, $\mathbf{a} \in \mathcal{A}_m \Rightarrow -\mathbf{a} \in \mathcal{A}_m$. Then, explicit solutions to steps 4-6 of Algorithm 2 are given by

(i) Step 4: $L \leftarrow \arg\max_m \{\|[\partial f(\mathbf{z}^{(k)})]^T \mathbf{A}_m\|_1\}$
(ii) Step 5: $\mathbf{c} = -\operatorname{sign}([\partial f(\mathbf{z}^{(k)})]^T \mathbf{A}_L)$
(iii) Step 6: $\mathbf{a} \leftarrow \mathcal{A}_L \mathbf{c}$

Further, for the case where $f(z) = \frac{1}{2}\|\mathbf{x}_j - \mathbf{z}\|_2^2$, the explicit solution to Step 7 is given by $\alpha_k = \max\{\min\{\alpha_o, 1\}, 0\}$ where

$$\alpha_o \doteq \frac{[\tau \mathbf{a} - \mathbf{z}^{(k)}]^T [\mathbf{x}_j - \mathbf{z}^{(k)}]}{\|\tau \mathbf{a} - \mathbf{z}^{(k)}\|_2^2} \tag{12}$$

Proof: (only a sketch given due to space constraints). Follows from using the fact that the sets $\mathcal{A}_i$ are centrally symmetric to show that the optimum in Step 4 is achieved by the linear combination of atoms given by $\mathbf{a} = \mathbf{A}_L \mathbf{c}$ with $c_i = sign\langle \partial f(\mathbf{z}^{(k)}), \mathbf{a}_{i,L} \rangle$, and using the explicit expression for $f(.)$ to compute $\partial f / \partial \alpha$. ■

The results above lead to the following Frank-Wolfe type algorithm for the specific case of network identification:

Note that this algorithm requires computing only inner products and sorting a vector and thus can comfortably handle very large data sets.

V. Extensions

In this section we cover several extensions of the basic algorithm needed to handle practical scenarios.

Algorithm 3 Topology Identification via Super-Atomic Norm Minimization

Define $\mathcal{A} = \{\mathbf{A}_1, \dots, \mathbf{A}_i, \dots\}$ and $\boldsymbol{\phi} = \{\mathbf{c}_1, \dots, \mathbf{c}_i, \dots\}$. Denote the l-th element as $\mathbf{A}_l$ and $\boldsymbol{\phi}_l$, respectively.
Initialize $\mathbf{z}^{(0)} = 0$ and $\boldsymbol{\phi}_l = 0$, $\forall l = 1, 2, \dots$
for k = 0,1,2,3,..., k_{max} **do**
  $L \leftarrow \arg\max_l \{\|[\partial f(\mathbf{z}^{(k)})]^T \mathbf{A}_l\|_1\}$
  $\mathbf{c} = -\operatorname{sign}([\partial f(\mathbf{z}^{(k)})]^T \mathbf{A}_L)$
  $\mathbf{a} \leftarrow \mathbf{A}_L \mathbf{c}$
  $\alpha_k \leftarrow \max\{\min\{\alpha_o, 1\}, 0\}$ where α_o is defined in (12)
  $\mathbf{z}^{(k+1)} \leftarrow \mathbf{z}^{(k)} + \alpha_k[\tau \mathbf{a} - \mathbf{z}^{(k)}]$
  $\boldsymbol{\phi}_l = (1 - \alpha_k)\boldsymbol{\phi}_l$, $\forall l$
  $\boldsymbol{\phi}_L = \boldsymbol{\phi}_L + (\alpha_k \tau)\mathbf{c}$
end for

A. External inputs

Many practical situations require taking into account relatively rate external events. Following [11], we will model these interactions by adding at each node, a piecewise constant signal $u_j(\cdot)$, with a sparse derivative.

$$x_j(t) = \sum_{i=1}^{n} \sum_{k=1}^{r} c_{j,i}(k) x_i(t-k) + u_j(t) + \eta_j(t), \quad t \in [r+1, T],\ j = 1, \dots, n \tag{13}$$

This extension fits naturally the proposed framework by modifying the objective in (6) to

$$\min \|\{\mathbf{c}_i\}\|_0 + \lambda \|\{\boldsymbol{\Delta} u_j\}\|_0 \tag{14}$$

where $\boldsymbol{\Delta} u_j \doteq \big[u_j(2) - u_j(1) \dots u_j(t) - u_j(t-1) \dots\big]$ and the parameter λ allows for trading-off graph versus input sparsity. The problem above can be reformulated in terms of super-atomic norm minimization, by simply adding the following super-atoms to the set $\mathcal{A}$:

$$\mathcal{A}_u = \frac{1}{\lambda}\{\mathbf{u}_1, \dots, \mathbf{u}_T\} \tag{15}$$

where $\mathbf{u}_t$ is defined as the t-th column of a lower triangular matrix with $\{0, 1\}$ elements. Thus, Algorithm 3 can be easily extended to handle external inputs by simply adding $\mathcal{A}_u$ into $\mathcal{A}$ and the corresponding coefficients to $\boldsymbol{\phi}$.

B. Missing data

Consider now a situation where some of the data is missing, due for instance to sensor outages, or in the case of video-based applications, occlusion. This scenario can be handled by noticing that from (5) it follows that

$$\mathbf{M} \doteq \begin{bmatrix} \mathbf{X} - \boldsymbol{\Xi} & \mathbf{H} \end{bmatrix}$$

does not have full column rank. Thus, missing data can be recovered by minimizing the rank of $\mathbf{M}$ with respect to the missing data and noise sequence, for instance by solving a regularized nuclear norm minimization problem of the form:

$$\min_{\hat{\mathbf{m}}, \hat{\mathbf{X}}, \hat{\mathbf{H}}} \| \begin{bmatrix} \hat{\mathbf{X}} & \hat{\mathbf{H}} \end{bmatrix} \|_* + \mu \|\hat{\mathbf{m}} - \mathbf{m}\|_\infty \tag{16}$$

where, $\begin{bmatrix} \hat{\mathbf{X}} & \hat{\mathbf{H}} \end{bmatrix}$ denotes the low rank estimation of $\mathbf{M}$, $\hat{\mathbf{m}}$ and $\mathbf{m}$ denote the elements of $\begin{bmatrix} \hat{\mathbf{X}} & \hat{\mathbf{H}} \end{bmatrix}$ and $\begin{bmatrix} \mathbf{X} & \mathbf{H} \end{bmatrix}$, respectively, at the positions where data is available.

C. Further sparsity enhancement via re-weighting

While often successful, in some scenarios Algorithm 3 may fail to find the sparsest solutions, specially in cases where the super-atoms have a large variation in norm. To address this issue, in this section we propose an iterative reweighted heuristic variant of Algorithm 3, where at each iteration the weights assigned to each super-atom are adjusted, in order to promote sparsity. This idea was first introduced in [22]. Note that an approach in the same spirit was used, in the context of network identification, in [13], in that case involving a re-weighted group lasso type penalty. At each iteration the modifies algorithm solves:

$$\begin{aligned} \min_{\mathbf{c}_i, p_t} \quad & \|\mathbf{x}_j - \mathbf{z}\|_2 \\ \text{s.t.} \quad & \|\mathbf{z}\|_{s\mathcal{A}} \leq \tau \end{aligned} \tag{17}$$

where $\mathbf{z}$ consists of two type of super atoms, $\mathcal{A}_a$ obtained from the data measured at each node, and $\mathcal{A}_u$ that accounts for (unknown) piecewise constant inputs, as outlined in section V-A, with the corresponding coefficients denoted as $\mathbf{c}_i$ and p_t, respectively. The algorithm proceeds by using the solution to the optimization at the k iteration to scale the super-atoms, with the initial weights given by their 2 norm. The complete procedure is shown in Algorithm 4.

Algorithm 4 Reweighted Network Topology Identification

Initialize $w_i^a = \|\mathcal{A}_a(i)\|_2 / mean(\{\|\mathcal{A}_a(i)\|_2, \forall i\})$, $\forall i$; $w_t^u = 1$, $\forall t$; $s_i = 1$, $\forall i \neq j$ and $s_j > 1$ (self-loop penalty)
while not converge **do**
  $w_i^a \leftarrow s_i w_i^a$, $w_t^u \leftarrow \lambda w_t^u$
  $\mathcal{A}_a \leftarrow \frac{1}{w^a} \circ \mathcal{A}_a$, $\mathcal{A}_u \leftarrow \frac{1}{w^u} \circ \mathcal{A}_u$
  Solve (17) using Algorithm 3
  $\mathbf{c}_i \leftarrow \frac{1}{w_i^a}\mathbf{c}_i$, $p_t \leftarrow \frac{1}{w_t^u} p_t$
  $w_i^a \leftarrow 1/(\|\mathbf{c}_i\|_\infty + \delta)$
  $w_t^u \leftarrow 1/(|p_t| + \delta)$
end while

VI. Examples

In this section, we illustrate the advantages of the proposed approach using two examples. In both cases, we compare the proposed method against Cycling Orthogonal Least Squares (COLS)[12], Group Lasso (GL)[9] and its variant modified to handle external inputs (GLEI) [13].

A. Synthetic Example

In this example, we first generated N nodes, and for each node a signal of length 1000 was drawn from a Normal distribution $\mathcal{N}(0, \mathbf{I})$. Then, in order to generate the time series observed at the output node, we randomly chose n_a nodes and generated random ARX models of order r , with coefficients uniformly distributed on $(0, 1)$. Finally, the

r	2	6	10
Proposed	**0.8232**	**5.4491**	**9.6660**
Group Lasso	3.0599	76.0957	352.1598
COLS	3.8695	11.1641	19.7986

TABLE I

EXAMPLE 1: MEAN COMPUTING TIME AS A FUNCTION OF SYSTEM ORDER ($N = 500$, $n_a = 10$)

n_a	5	10	20	50
Proposed	5.5009	**5.4970**	**5.4772**	**5.6117**
Group Lasso	52.3320	61.0487	53.2604	55.9573
COLS	**2.7994**	11.2746	47.1583	389.2545

TABLE II

EXAMPLE 1: MEAN COMPUTING TIME AS A FUNCTION OF NUMBER OF LINKS ($N = 500$, $r = 6$)

output was then corrupted using Gaussian noise drawn from $\mathcal{N}(0, 0.05\mathbf{I})$, achieving a Signal-to-Noise ratio around $25dB$.

We considered three different scenarios obtained by fixing two parameters from the set r, n_a and N and varying the third. For each parameter setting, we ran 10 experiments and compared the results against those obtained using Group Lasso and COLS. All three approaches successfully identified the underlying system. However, as shown in Tables I, II and III, Algorithm 3 outperforms the others in terms of computational complexity as the size of the problem grows.

B. Human Interaction

The goal here is to identify causal relationships from video data. For this experiment, we took the first example from [13], which considers two video sequences from the UT Human Interaction Data Set [24]. In both video sequences, we use as data the position of each agent's head in image coordinates, normalized to the interval $[-1, 1]$. The proposed reweighted algorithm (Algorithm 4), reweighted GLEI and COLS were run using system order $r = 2$. For the proposed method, we set the self-loop penalty equal to 10, $\lambda = 0.05$ and $\tau = 5$. For the reweighted GLEI, we adopt the parameter setting reported in [13]. Since COLS needs information on the number of links, we used 1 for each agent which is intuitively consistent with the ground truth, that is the fact that the video consists mostly of pair-wise interactions.

For both sequence 6 and 16, a clip of around 100 frames was used. The causal correlations between agents identified by all three methods are shown in Figure 1 and 2, respectively. Note that COLS fails to identify the correct relationships, since it does not take into account external inputs[1]. On the other hand, both the proposed method and reweighted GLEI correctly identified the causally interacting people. However, the proposed method required less computing time per iteration than reweighted GLEI. For sequences 6 and 16, the proposed method took 0.0643 and 0.1115 seconds per iteration, respectively, while reweighted GLEI required 0.4867 and 0.3583 seconds per iteration. In this experiment, both the proposed method and reweighted GLEI converged in about the same number of iterations.

N	100	200	300	400	500
proposed	**0.4373**	**1.2237**	**2.6294**	**4.0679**	**5.4672**
Group Lasso	0.5539	3.7758	12.7853	29.3695	51.5379
COLS	2.0871	4.2519	6.4819	8.7866	11.1526

TABLE III

EXAMPLE 1: MEAN COMPUTING TIME AS A FUNCTION OF THE TOTAL NUMBER OF NODES ($r = 6$, $n_a = 10$)

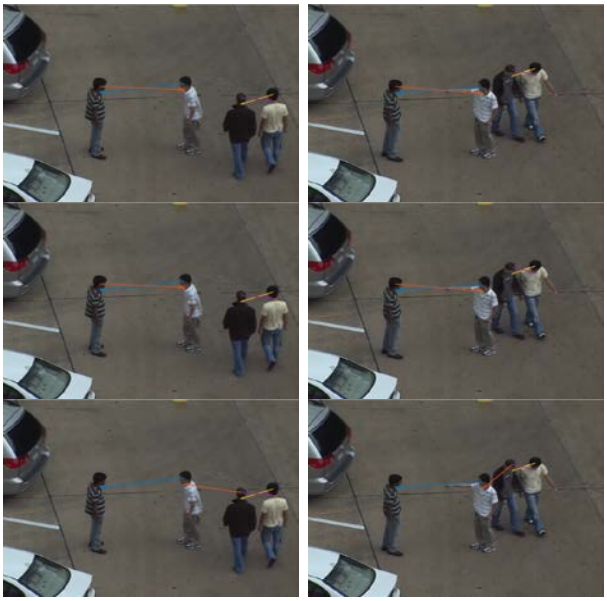

Fig. 1. Sample Frames of the UT Sequence 6 showing the causally interacting groups identified using different methods. Top: Proposed Method. Center: reweighted GLEI. Bottom: COLS.

C. Youtube Tennis Game

In this example, the goal is to identify causally interacting agents in a clip of 210 frames taken from a mixed double tennis match at the London 2012 Olympics. The position of the centroid of each player was recorded, and normalized to interval $[-1, 1]$. In this experiment, we also set the system order for each edge to $r = 2$. For both the proposed method and reweighted GLEI, we set the self-loop penalty to 10 and $\lambda = 0.5$. In the proposed method, we used $\tau = 20$. For reweighted GLEI, we chose $\epsilon = 0.1$. For COLS, we set the link number on each player to 2 which is consistent to the fact that each team consisted of two people. The causal correlations identified by three methods are shown in Fig. 3.

In this case, intuitively it is expected that players will react primarily to their opponents. Therefore, we would expect links between player and his/her opponents. The graph identified by COLS doesn't match this intuition, while the ones obtained using the proposed method and reweighted GLEI do. For this experiment, Algorithm 4 and reweighted GLEI also took about the same number of iterations to

[1]In this applications, these inputs account for interactions between an agent and its environment.

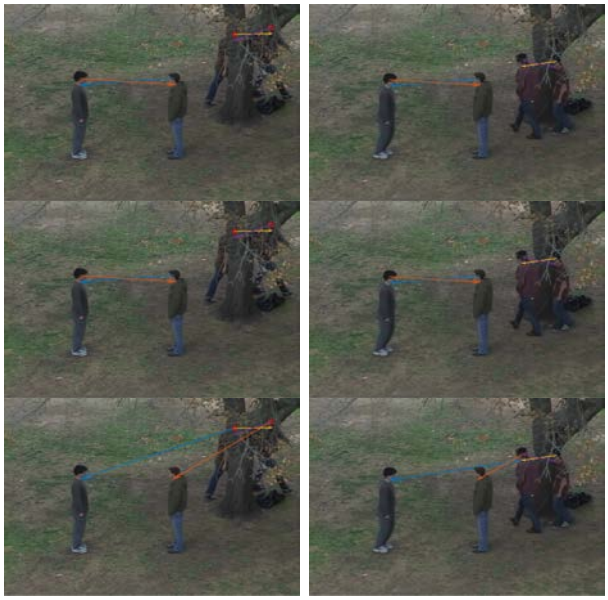

Fig. 2. Sample Frames of the UT Sequence 16 showing the causally interacting groups identified using different methods. Top: Proposed Method. Center: reweighted GLEI. Bottom: COLS. The red circle denotes the agent position recovered by solving (16)

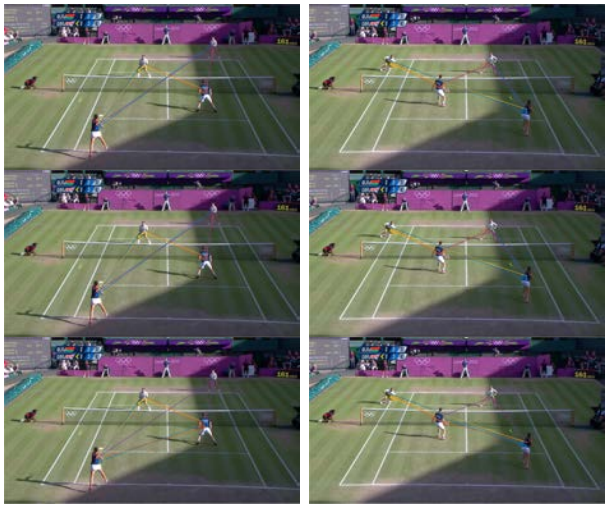

Fig. 3. Causally interacting groups in Double Tennis. Top: Proposed Method. Center: reweighted GLEI. Bottom: COLS.

converge. However. each iteration of the proposed method required 0.3802 seconds versus 2.13 for GLEI.

VII. Conclusions

Many problems of practical interest require identifying a dynamical graphical model from input/output data. As shown in this paper, this can be efficiently done by recasting the problem into an expanded atomic-norm minimization framework that promotes block-sparsity and allows for exploiting computationally efficient Frank-Wolfe type algorithms. Further, the proposed framework can be easily expanded to accommodate unknown exogenous inputs and missing data. These results were illustrated with several examples drawn from video-analytics, showing that the proposed method outperforms existing ones, specially as the size of the problem increases.

References

[1] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, pp. 424–438, 1969.

[2] A. J. Seneviratne and V. Solo, "Topology identification of a sparse dynamic network," in *2012 IEEE CDC,* pp. 1518–1523.

[3] P. A. Valdés-Sosa, *et. al.* "Estimating brain functional connectivity with sparse multivariate autoregression," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1457, pp. 969–981, 2005.

[4] A. Arnold, Y. Liu, and N. Abe, "Temporal causal modeling with graphical granger methods," in *Proc. of the 13th ACM SIGKDD*, 2007, pp. 66–75.

[5] A. Papachristodoulou and B. Recht, "Determining interconnections in chemical reaction networks," in *American Control Conference, 2007. ACC'07*. IEEE, 2007, pp. 4872–4877.

[6] M. M. Zavlanos, S. P. Boyd, G. J. Pappas *et al.*, "Identification of stable genetic networks using convex programming," in *American Control Conference, 2008*. IEEE, 2008, pp. 2755–2760.

[7] D. Hayden, Y. H. Chang, J. Goncalves, and C. Tomlin, "Compressed sensing for network reconstruction," *arXiv:1411.4095*, 2014.

[8] 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.

[9] S. Haufe, G. Nolte, K.-R. Mueller, and N. Krämer, "Sparse causal discovery in multivariate time series," *arXiv:0901.2234*, 2009.

[10] A. Bolstad, B. D. Van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *IEEE Trans. Signal Processing*, vol. 59, no. 6, pp. 2628–2641, 2011.

[11] M. Ayazoglu, M. Sznaier, and N. Ozay, "Blind identification of sparse dynamic networks and applications," in *2011 IEEE CDC*, pp. 2944–2950.

[12] D. Materassi, G. Innocenti, L. Giarré, and M. Salapaka, "Model identification of a network as compressing sensing," *Systems & Control Letters*, vol. 62, no. 8, pp. 664–672, 2013.

[13] M. Ayazoglu, B. Yilmaz, M. Sznaier, and O. Camps, "Finding causal interactions in video sequences," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3575–3582.

[14] B. M. Sanandaji, T. L. Vincent, and M. B. Wakin, "Exact topology identification of large-scale interconnected dynamical systems from compressive observations," in *2011 ACC*, pp. 649–656.

[15] E. Candes and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse problems*, vol. 23, no. 3, p. 969, 2007.

[16] B. M. Sanandaji, T. L. Vincent, and M. B. Wakin, "Compressive topology identification of interconnected dynamic systems via clustered orthogonal matching pursuit," in *2011 IEEE CDC*, pp. 174–180.

[17] A. Chiuso and G. Pillonetto, "A bayesian approach to sparse dynamic network identification," *Automatica*, 48, 8, pp. 1553–1565, 2012.

[18] M. Zorzi and A. Chiuso, "A bayesian approach to sparse plus low rank network identification," *arXiv preprint arXiv:1503.07340*, 2015.

[19] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Foundations of Computational mathematics*, vol. 12, no. 6, pp. 805–849, 2012.

[20] A. Tewari, P. K. Ravikumar, and I. S. Dhillon, "Greedy algorithms for structurally constrained high dimensional problems," in *Advances in Neural Information Processing Systems*, 2011, pp. 882–890.

[21] N. Ozay, M. Sznaier, C. M. Lagoa, O. Camps *et al.*, "A sparsification approach to set membership identification of switched affine systems," *Automatic Control, IEEE Transactions on*, vol. 57, no. 3, pp. 634–648, 2012.

[22] E. J. Candes, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier analysis and applications*, vol. 14, no. 5-6, pp. 877–905, 2008.

[23] I. Gurobi Optimization, "Gurobi optimizer reference manual," 2015. [Online]. Available: http://www.gurobi.com

[24] M. S. Ryoo and J. K. Aggarwal, "UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA)," http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html, 2010.