Gaussian Graphical Model Selection from Size Constrained Measurements

Gautam Dasarathy
School of Electrical, Computer, and Energy Engineering
Arizona State University
Tempe, AZ, USA
gautamd@asu.edu

***Abstract*—In this paper, we introduce the problem of learning graphical models from size constrained measurements. This is inspired by a wide range of problems where one is unable to measure all the variables involved simultaneously. We propose notions of data requirement for this setting and then begin by considering an extreme case where one is allowed to only measure pairs of variables. For this setting we propose a simple algorithm and provide guarantees on its behavior. We then generalize to the case where one is allowed to measure up to r variables simultaneously, and draw connections to the field of combinatorial designs. Finally, we propose an interactive version of the proposed algorithm that is guaranteed to have significantly better data requirement on a wide range of realistic settings.**

***Index Terms*—Gaussian graphical models, active learning, sample complexity, combinatorial designs**

I. Introduction

Probabilistic graphical models provide a powerful and flexible framework for expressing the relationships between a large number of entities in a complex system. It is unsurprising therefore that they are finding applications in a large number of complex domains from statistical physics and computational biology to computer vision and natural language processing to computational biology and statistical physics. In this paper, we consider the problem of learning the structure of graphical models from data observed from the underlying system. This is useful as it not only may reveal some fundamental relationships between the underlying variables, but also because it provides a computationally efficient representation of a complex system that could be invaluable for downstream processing.

Unfortunately, in several scenarios where there is a need to perform structure learning, the number of observations is typically much smaller than the total number of variables – the so called *high-dimensional regime*. It is known that many natural sufficient statistics such as the sample covariance matrix are poorly behaved (see e.g., [1], [2]). A recent exciting line of work has explored many conditions under which this problem becomes tractable (e.g., [3]–[6]). Various authors have discovered that by constraining both the structure of the graph and the parameters of the probabilistic model, there is a wide range of interesting situations where given $\mathcal{O}(\log p)$ samples from the underlying distribution, one can learn the structure and sometimes even the parameters of the underlying graphical model.

This paper considers the structure learning problem from a different point of view. We are motivated by a wide variety of applications where it might be impossible to obtain measurements from all the variables involved in the system simultaneously. For instance, in two-photon calcium imaging (see e.g., [7]) which provides state-of-the-art spatial and temporal resolution for measuring neuronal activity, there are hard constraints on the number of neurons that can be measured simultaneously. In a sensor network, obtaining a sample across all the sensors requires intensive synchronization procedures (see e.g., [8]) that may be infeasible in a power-starved system. Similarly in many other problem domains like proteomics [9] and neuroscience [10], [11], it might be much easier to obtain (marginalized) samples from small subsets of variables as opposed to full snapshots.

[12] proposed to handle this problem using an interactive procedure that sequentially and adaptively reduces the number of variables it needs measurements from. While such algorithms are known to compare favorably with traditional algorithms and are known to have near optimal data requirements [13], they still require several measurements from the entire system. In this paper, we consider the setting where this is not just undesirable, but rather impossible. We pose the problem of graphical model selection from size constrained measurements, and propose notions of data requirement for this setting. We then begin by considering an extreme case where one is allowed to only measure pairs of variables, and we propose a simple algorithm and provide guarantees on its behavior. We then generalize to the case where one is allowed to measure no more than r variables simultaneously, and draw connections to the field of combinatorial designs. Finally, we propose an interactive version of the proposed algorithm that is guaranteed to have significantly better data requirement on a wide range of realistic settings.

II. Problem Setup and Notation

Let $G = ([p], E)$ denote an undirected graph on the vertex set $[p] = \{1, 2 \ldots, p\}$ with an edge set $E \subseteq \binom{p}{2}$. To each vertex i of the graph, we associate the $i-$th component of a $0-$mean Gaussian random vector $X \in \mathbb{R}^p$. X is distributed according to the multivariate normal distribution with covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$; the density of X is of course given by $f_X(x) = \frac{1}{\sqrt{(2\pi)^p \det(\Omega)}} \exp\left\{-\frac{1}{2} x^\intercal \Omega x\right\}$, where $\Omega = \Sigma^{-1}$

is the $p-$dimensional inverse covariance (or *concentration*) matrix of the distribution. We will abbreviate this density as $\mathcal{N}(0,\Sigma)$ in the sequel.

X is said to be Markov with respect to the graph G if for any pair of vertices i and j, $\{i,j\} \notin E$ implies that X_i and X_j are conditionally independent given the values of $X_{[p]\setminus\{i,j\}}$[1]. By the Hammersly - Clifford theorem [14], we know that for all $\{i,j\} \notin E$, $\Omega_{ij} = \Omega_{ji} = 0$. We will also use the fact that X_i and X_j are conditionally independent given a set of variables S is equivalent to saying $\rho_{ij|S} = 0$ for the Gaussian distribution. We will refer to the pair (G,Σ) as the (Gaussian) graphical model. For a vertex $i \in [p]$, we will let $N(i)$ denote the neighborhood of the vertex in the graph and let $d_i \triangleq |N(i)|$ denote the degree. We will assume that $d_i \leq d$, for some constant $d \in \mathbb{N}$ for the graphs considered in the sequel.

The goal of the structure learning problem is to recover the structure of the graph G given samples from the distribution f. That is, we would like to construct an estimate $\widehat{E}$ of the edge set E. As mentioned in Section I, in this paper we are interested in the setting where our estimator can only access the distribution by obtaining samples of size at most r. That is, the algorithm is allowed to specify a sequence $(S_k, n_k) \in 2^{[p]} \times \mathbb{N}, k = 1,2,\ldots,m$ such that $|S_k| \leq r$. The algorithm then observes n_k i.i.d copies of X_{S_k} for each $k \in [m]$; notice that the distribution of X_{S_k} is given by $\mathcal{N}(0,\Sigma(S_k))$, where by $\Sigma(S_k))$ we mean the submatrix of Σ corresponding to only the rows and columns corresponding to S_k. We will call such an algorithm an *$r-$constrained algorithm*. In order to characterize the performance of such an algorithm, we posit a natural measure of complexity along the lines of [12] – we will measure the total number of scalars that the algorithm accrues to deliver an estimate of a particular accuracy at a required confidence level. Towards this end, notice that an algorithm $\mathcal{A}$ that specifies $(S_k, n_n) \in 2^{[p]} \times \mathbb{N}$ accumulates a total number of scalars that is given by $n_{\text{tot}}(\mathcal{A}) \triangleq \sum_{k=1}^{m} n_k |S_k|$. We will now define the following notion of data requirement for this problem.

Definition 1 ($r-$Sample Complexity)**.** *We will say that a graphical model (G,Σ) can be learnt with an $r-$sample complexity of n_0 if the following holds. Fix $\delta \in (0,1)$. Then, there exists an $r-$constrained algorithm $\mathcal{A}$ and a function $n_0 : (0,1) \to \mathbb{N}$ such that whenever $n_{\text{tot}}(\mathcal{A}) \geq n_0(\delta)$, $\mathcal{A}$ returns an edge set $\widehat{E}$ that satisfies $\mathbb{P}[\widehat{E} \neq E] \leq \delta$.*

In what follows, we will investigate the learnability of Gaussian graphical models by $r-$constrained algorithms and to especially understand the data requirement of these algorithms by understanding their $r-$sample complexities.

III. The case of $r = 2$

In this section, we will first consider an extreme version of this problem where we're only allowed to look at pairs of variables. In what follows, we will establish a bound on the $2-$sample complexity by demonstrating an algorithm and analyzing its data requirement. The algorithm is computationally unsophisticated, and essentially performs an exhaustive search over all possible conditional independence relationships. This can be easily improved by instead using our estimates of the partial correlations in the PC algorithm [15], [16] which organizes the conditional independence tests carefully. Our algorithm will rest on the following *recursive definition* of the partial correlation coefficient. Given $i,j \in [p]$ and $S \subset [p] \setminus \{i,j\}$, the following holds for any $k \in S$:

$$\rho_{ij|S} = \frac{\rho_{ij|S_{-k}} - \rho_{ik|S_{-k}}\rho_{jk|S_{-k}}}{\sqrt{(1-\rho^2_{ik|S_{-k}})(1-\rho^2_{jk|S_{-k}})}}. \tag{1}$$

Therefore, the partial correlation coefficients can be recursively computed from partial correlation coefficients of lower orders. And of course, the unconditional partial correlation is the regular (Pearson) correlation coefficient and is given by $\rho_{XY} = \text{Cov}(X,Y)/\sigma_X\sigma_Y$. We will now characterize the

Algorithm 1

Input: Threshold $\eta > 0$.

```
Initialize Ê = (p choose 2), the complete graph
for Each pair each pair i, j ∈ [p] do
    for Each set S ⊂ [p] such that |S| ≤ d do
        Compute ρ̂_{ij|S} according to (1)
        if |ρ̂_{ij|S}| < η then
            Ê = Ê \ {i, j}; break
        end if
    end for
end for
```

performance of this algorithm and thereby an estimate of the $2-$sample complexity of learning a bounded degree Gaussian graphical model. In order to do this we need some assumptions on the parameters:

(A1) The distribution of X is faithful to G.

(A2) For each $i,j \in [p]$ and $S \subset [p]$ with $|S| \leq d$, we have $\rho_{ij|S} \leq M$. Furthermore, if X_i is not conditionally independent of X_j given S, then we have $\rho_{ij|S} \geq m$

Assumption (A1) is a standard assumption in the graphical model selection literature and is violated only on a set of measure 0 (see e.g., [15], [17]). Assumption (A2) has appeared in the literature (e.g., [16]) as a way of strengthening the faithfulness assumption. While the upper bound in assumption (A2) is a mild regularity condition, the lower bound of (A2) may be hard to verify in practice. However, under certain parametric and structural conditions, one can obtain a handle on m. For example, the authors in [6] show that if the underlying graph has small local separators and if the concentration matrix is *walk-summable*, then m in (A2) can be replaced essentially by the smallest non-zero entry of the concentration matrix.

Theorem 1. *Fix a $\delta \in (0,1)$. There is a constant $C_0 > 1$ that depends only on m and M from Assumptions (A1) and (A2) such that Algorithm 1 succeeds in recovering the edge set of the graph G with probability greater than $1-\delta$, provided*

[1]We will write X_A for $A \subset [p]$ to denote the vector in $\mathbb{R}^{|A|}$ that is a concatenation of $\{X_i : i \in A\}$.

that the total number of scalar samples the algorithm obtains satisfies

$$n_{\text{tot}} \geq C_0^d \binom{p}{2} \times \log\left(\frac{p^2}{\delta}\log\left(\frac{\binom{p}{2}}{\delta}\right)\right). \tag{2}$$

We will now sketch the proof of this theorem.

Proof: (Sketch) The first step to proving this theorem is to notice that the algorithm succeeds in recovering the graph if we have the exact partial correlation coefficients. This follows from the definition of the partial correlations and the faithfulness assumption (A1); for more on this, see e.g., [15], [16].

Let $\mathcal{E}$ denote the event that there was error in the recovery of the graph structure. Notice that this error occurs only when there is a particular conditional independence test that fails. If we let $\mathcal{E}_{ij|S}$ denote an error in the conditional independence test of X_i and X_j given X_S, then we have the following

$$\mathbb{P}\left[\mathcal{E}\right] = \mathbb{P}\left[\bigcup_{i,j\in[p],S\subset[p]\setminus\{i,j\}} \mathcal{E}_{ij|S}\right] \leq \sum_{i,j,S} \mathbb{P}\left[\mathcal{E}_{ij|S}\right]. \tag{3}$$

Let us now consider one of the terms in (3) Notice that

$$\begin{aligned}\mathbb{P}\left[\mathcal{E}_{ij|S}\right] &= \mathbb{P}\left[\left|\widehat{\rho}_{i,j|S}\right| \geq \eta\right] \mathbb{1}\left\{X_i \perp\!\!\!\perp X_j \mid X_S\right\} \\ &+ \mathbb{P}\left[\left|\widehat{\rho}_{i,j|S}\right| \leq \xi\right] \mathbb{1}\left\{X_i \not\perp\!\!\!\perp X_j \mid X_S\right\}\end{aligned} \tag{4}$$

Now, we will consider the first term in (4). This term is non-zero only when $X_i \perp\!\!\!\perp X_j \mid X_S$. Therefore, we will assume that this is the case and note that this implies $\rho_{i,j|S} = 0$. We then have

$$\mathbb{P}\left[\left|\widehat{\rho}_{i,j|S}\right| \geq \xi\right] \mathbb{1}\left\{X_i \perp\!\!\!\perp X_j \mid X_S\right\} \leq \mathbb{P}\left[\left|\widehat{\rho}_{i,j|S} - \rho_{i,j|S}\right| \geq \xi\right]$$

Similarly, the second term in (4) is nonzero only when $X_i \not\perp\!\!\!\perp X_j \mid X_S$. This implies, by (A2), that $\rho_{i,j|S} \geq m$. Now, observe that the conditions $\left|\widehat{\rho}_{i,j|S}\right| \leq \xi$ and $\left|\rho_{i,j|S}\right| \geq m$ together imply that $\left|\rho_{i,j|S}\right| - \left|\widehat{\rho}_{i,j|S}\right| \geq m - \xi \Rightarrow \left|\rho_{i,j|S} - \widehat{\rho}_{i,j|S}\right| \geq m - \xi$, since we will choose $m > \xi$. Therefore, we have

$$\begin{aligned}&\mathbb{P}\left[\left|\widehat{\rho}_{i,j|S}\right| \leq \xi\right] \mathbb{1}\left\{X_i \not\perp\!\!\!\perp X_j \mid X_S\right\} \\ &\qquad \leq \mathbb{P}\left[\left|\widehat{\rho}_{i,j|S} - \rho_{i,j|S}\right| \geq m - \xi\right]\end{aligned} \tag{5}$$

Putting these two expressions in (4) and setting $\xi = m/2$, we have the following upper bound

$$\mathbb{P}\left[\mathcal{E}_{ij|S}\right] \leq \mathbb{P}\left[\left|\widehat{\rho}_{ij|S} - \rho_{ij|S}\right| \geq m/2\right]. \tag{6}$$

Therefore, to conclude establishing an estimate on the above probability it suffices to establish a concentration result on the partial correlation coefficient. We note here that while standard results for the concentration of the partial correlation exist in the literature (e.g., [16]), these do not apply here since we do not measure the variables X_i, X_j, X_S simultaneously. We instead establish a (weaker) concentration result using the recursive formula for partial correlations.

Lemma 1. *Let us suppose that $\left|\widehat{\rho}_{ij|S} - \rho_{ij|S}\right| \leq \varepsilon_1$ for all pairs $i, j \in [p]$ and all subsets $S \subset [p]$ that have cardinality $|S| \leq k-1$. This implies that there is a constant $C > 0$ such that the following bound holds for all pairs of vertices $i, j \in [p]$ and subsets $S \subset [p]$ with $|S| = k$: $\left|\widehat{\rho}_{ij|S} - \rho_{ij|S}\right| \leq C\varepsilon_1$, where C is only a function of M from Assumption (A2).*

Proof: (Sketch) This result can be shown to be true by bounding the partial derivatives of the function in (1) with respect to $\rho_{ij|S_{-k}}, \rho_{ik|S_{-k}}, \rho_{jk|S_{-k}}$. We will omit the details of this proof in this short abstract.

■

This means that if we set $\alpha = \frac{3\sqrt{1-M^2}}{M^5}$, and if $|\widehat{\rho}_{ij} - \rho_{ij}| \leq m/2\alpha^d$ for all pairs $i, j \in [p]$, then we have the following relationship for i, j, S, where $S \subset [p]$ and $|S| \leq d$: $\left|\widehat{\rho}_{ij|S} - \rho_{ij|S}\right| \leq \frac{m}{2}$.

The probability of making an error in a conditional independence test therefore is bounded from above by the probability that there is a pair of vertices $i, j \in [p]$ such that $|\widehat{\rho}_{ij} - \rho_{ij}| \geq \frac{m}{2\alpha^d} = \varepsilon_0$. We can now use the following concentration result for the standard correlation coefficient; see for instance [16]:

Lemma 2. *Provided (A2) holds, given n samples from the pair (X_i, X_j), the empirical correlation coefficient $\widehat{\rho}_{i,j}$ satisfies the following*

$$\mathbb{P}\left[|\widehat{\rho}_{i,j} - \rho_{i,j}| \geq \epsilon\right] \leq C_1 (n-2) e^{-(n-4)\log\left(\frac{4+\epsilon^2}{4-\epsilon^2}\right)}, \tag{7}$$

where $C_1 > 0$ is a constant that depends on M from (A2).

From Lemmas 1 and 2, we can bound the probability of error as follows

$$\mathbb{P}\left[\bigcup_{i,j} |\widehat{\rho}_{ij} - \rho_{ij}| \geq \varepsilon_0\right] \leq \binom{p}{2} C_1 (n-2) e^{-(n-4)\log\left(\frac{4+\varepsilon_0^2}{4-\varepsilon_0^2}\right)}. \tag{8}$$

It is not hard to verify that there is a constant C_2 such that if we choose $n \geq C_2 \varepsilon_0^2 \log\left(\frac{C_2 p^2}{\delta}\log\left(\frac{p^2}{\delta}\right)\right)$.

then, the above probability can be guaranteed to be less than δ. To conclude the proof, we simply observe that there are $\binom{p}{2}$ pairs, and the value of $\varepsilon_0 = \frac{m}{2}(\frac{1}{\alpha})^d$; this gives us the total $2-$sample complexity stated in the theorem.

■

IV. The $r \geq 3$ case: Connection to Combinatorial Design Theory

In this section, we will consider the same problem when we are allowed to measure 3 or more variables at once. It is of course easy to see that we can achieve an $r-$sample complexity of the order stated in Theorem 1 – to compute the correlation coefficient between a pair of variables X_i and X_j, one could simply ignore the fact that we're allowed to measure three variables at once and just measure the pair (or, equivalently measure these two along with a third arbitrary variable X_k). Therefore, to understand the statistical benefit of measuring several random variables at once, we will instead focus on whether we can *cover* all the elements of the covariance matrix using correlations measured from variables r at a time. Towards this end, we will define the following.

Problem 1 ($r-$covering)**.** *Given a family of subsets $S_k \subset [p], k = 1, \ldots, m$, we say that $\{S_k\}$ is an $r-$covering of $[p]$ if (a) $|S_k| \leq r$ for all $k \in [m]$ and (b) for each pair $i, j \in [p]$, there is a $k \in [m]$ such that $i, j \in S_k$.*

We will denote the smallest m such that there is an $r-$covering of $[p]$ as m_{rp}, and call this the $r-$covering number. Clearly $m_{r,p} \geq \frac{\binom{p}{2}}{\binom{r}{2}}$ since each set S_k can only contain a maximum of $\binom{r}{2}$ pairs and there are a total of $\binom{p}{2}$ pairs that need to be covered. The $r-$covering number is directly related to the data requirement for the graphical model selection problem in question.

Theorem 2. *Consider the graphical model selection problem for (G, Σ) where the learning algorithm is allowed to measure up to r variables at once. Suppose that $m_{r,p}$ is the $r-$covering number for this value of r and p. Then, for any $\delta \in (0, 1)$, there is an algorithm that succeeds in recovering the edge set of the graph G with probability greater than $1 - \delta$, provided the total sample complexity n satisfies*

$$n \geq C_0^d m_{r,p} \log\left(\frac{p^2}{\delta} \log\left(\frac{p^2}{\delta}\right)\right). \tag{9}$$

Proof: (Sketch) We will present a brief sketch of the proof in this short abstract. Suppose that $S_k, k = 1, 2, \ldots, m_{rp}$ are the family of subsets that correspond to the $r-$cover, then the reconstruction algorithm would proceed as follows. For each pair $i, j \in [p]$, the algorithm will choose the $k \in [m_{rp}]$ such that $S_k \ni i, j$. The algorithm will then estimate $\widehat{\rho}_{ij}$ using the samples from X_{S_k} and use these estimated correlations in an exhaustive search, exactly as in Algorithm 1. ■

Notice that Theorem 2 simply provides guarantees on the $r-$sample complexity without explicitly providing an algorithm. In particular, the algorithm that achieves the stated sample complexity needs an explicit $r-$covering of size m_{rp}. The astute reader will have observed $r-$coverings simply depend on the parameters p and r. For instance, for $p = 7$ and $r = 3$, the following family of subsets is a minimal $r-$covering: {1,2,3}, {1,4,5}, {1,6,7}, {2,4,6},{2,5,7}, {3,4,7}, {3,5,6}. In fact, such combinatorial constructions (called block designs) have been the subject of a long line of research broadly termed combinatorial design theory, and the interested reader may refer to [18], [19], for instance.

While several lines of work in this area consider the issues of the existence of exact block designs, where the target is to construct the $r-$covering set of minimal size (when the obvious divisibility conditions are satisfied), for the purposes of this paper, we will be satisfied $r-$coverings that are only close to minimal. Towards this end, there exist efficient randomized algorithms (cf., Rödl's Nibble [20], [21]) that allows one to obtain a covering of size $(1 + o(1))\binom{p}{2}/\binom{r}{2}$. This allows us to state the following corollary.

Corollary 1. *There is an (explicit) algorithm for reconstructing a graphical model (G, Σ) that succeeds in reconstructing the edges exactly with probability greater $1 - \delta$ provided the total number of scalar samples n satisfies $n_{\text{tot}} \geq C_0^d \frac{\binom{p}{2}}{\binom{r}{2}} \log\left(\frac{p^2}{\delta} \log\left(\frac{p^2}{\delta}\right)\right)$.*

Given that this at the heart of this construction is a nearly optimal $r-$covering of the covariance matrix, one might wonder if this can be improved. In the next section, we show that this total sample complexity can be vastly improved if the reconstruction algorithm is allowed to be interactive.

V. Active Learning for Graphical Model Selection from Size-Constrained Measurements

Inspired by [12], we ask whether one might improve the total sample complexity by allowing the reconstruction algorithm to be interactive. In particular, if the algorithm can sequentially and adaptively select which subsets to measure and decide how many samples it needs from these subsets, can one hope to improve the sample complexity? The answer turns out to be affirmative in most realistic situations where the graph has an inhomogeneous degree distribution. The algorithm we will present will assume we have access to a subroutine **`rCover`**(p, r) (say, like in [20], [21]) that produces a nearly optimal $r-$cover of $[p]$. We will now present a modification of the **`AdPaCT`** algorithm of [12] that applies to the size constrained graphical model selection problem.

Algorithm 2 **`SC-AdPaCT`**: Size-Constrained Adaptive Partial Correlation Testing

Require: : Threshold $\xi > 0$

Initialize: $\ell = 1$, $\widehat{N}(i), \forall i \in [p]$, NBDFOUND, SETTLED to $\emptyset$(the empty set)
repeat
  Obtain an $r-$covering $\{S_k\}$ of $[p] \backslash$SETTLED from **`rCover`**$(p - |\text{SETTLED}|, r)$
  Obtain $n_\ell = \alpha^\ell \log\left(\frac{p^2}{\delta} \log\left(\frac{p^2}{\delta}\right)\right)$ independent samples from each X_{S_k}.
  for each $i \in$ NBDFOUND^c **do**
    $\mathcal{S} = \{S \subset \text{SETTLED}^c : |S| = \ell, \max_{j \notin S} |\widehat{\rho}_{ij|S}| \leq \xi\}$
    if $\mathcal{S} = \emptyset$ **then**
      continue
    else
      Set $\widehat{N}(i) = \arg\min_{S \in \mathcal{S}} |\widehat{\rho}_{ij|S}|$
    end if
  end for
  $S = \{i \in \text{NBDFOUND} : \widehat{N}(i) \subset \text{NBDFOUND}\}$
  SETTLED = SETTLED $\bigcup$ S
  $\ell = \ell \times 2$
until NBDFOUND$= [p]$

We will now briefly describe Algorithm 2, which is a modification of the AdPaCT algorithm of [12] and allows for size constrained graphical model selection. We start with an empty graph on $[p]$ and initialize counter ℓ to 1 and sets NBDFOUND, SETTLED to $\emptyset$. NBDFOUND will be used to keep track of the vertices whose neighborhood estimates

the algorithm is confident about and SETTLED keeps track of the vertices that no longer need to be sampled from. Notice that the faster SETTLED is populated, the better the performance of Algorithm 2, since in successive stages only the vertices in SETTLED^c are sampled. The algorithm then loops over ℓ (by doubling) until NBDFOUND $= [p]$. At each iteration, the algorithm obtains first a covering of the "unsettled" vertices using the **rCover** subroutine and obtains $n_\ell = \alpha^\ell \log\left(\frac{p^2}{\delta}\log\left(\frac{p^2}{\delta}\right)\right)$ independent samples from X_{S_k} for each S_k in the covering. Next, for each vertex $i \notin$ NBDFOUND the algorithm uses partial correlation testing (as in Algorithm 1) to obtain an estimate of a neighborhood of i of size at most ℓ. If this is possible to find, the algorithm adds i to the set **nbdFound**. Finally, the set SETTLED gets updated. Any i in NBDFOUND whose entire estimated neighborhood is in NBDFOUND gets enrolled in SETTLED and does not get sampled henceforth. That is, the algorithm "settles" a vertex $i \in [p]$ if it is both confident about the vertex's neighborhood and about the neighborhood of i's neighbors. It is this step that gives our algorithm its improved total sample complexity.

We will now state a performance guarantee for this algorithm. Towards this end, we will define the following notion of complexity of the graph G, for each $\ell = 1, 2, \ldots, d$, we will p^ℓ be defined as follows

$$p^{(\ell)} = p - |\{i \in [p] : d_i < \ell, \text{ and } \forall j \in N(i), d_j < \ell\}|\,.$$

Notice that $p^{(1)} = p$ and $p^{(\ell)} = 0$ for $\ell > d$. Also, notice that for graphs whose degree distributions are homogenous, the sequence $p^{(1)} \approx p^{(2)} \approx \cdots \approx p^{(d)} \approx p$. On the other hand, for graphs whose degree sequence is rather inhomogeneous (as is the case in most real world graphs, see e.g. [22]), this sequence of numbers would rapidly fall. The following theorem will show that in this latter case, the total sample complexity of the interactive SC-AdPaCT algorithm is significantly better than the passive algorithm.

Theorem 3. *Fix $\delta \in (0,1)$. Algorithm 2 succeeds in reconstructing the graph G with probability at least $1-\delta$ and has a total sample complexity given by the following expression*

$$\sum_{\ell=1,2,4,\ldots,d} \frac{\binom{p^{(\ell)}}{2}}{\binom{r}{2}} C_0^\ell \log\left(\frac{p^2}{\delta}\log\left(\frac{p^2}{\delta}\right)\right). \tag{10}$$

Notice that this implies that Algorithm 2 is guaranteed to never have a worse sample complexity than its passive counterpart Algorithm 1. Moreover, the faster the sequence $\{p^{(1)}, \ldots, p^{(d)}\}$ drops to zero, the more drastic the improvements provided by adaptivity. We will now briefly sketch a proof of this result.

Proof: (Sketch) This result can be proved by adopting a strategy similar to that of [12]. The main idea is that one can bound the probability that the algorithm fails at step ℓ, when it has succeeded in all the preceding steps. Of course, by definition, conditioning on the event that all the preceding steps have succeeded implies that at step ℓ we will be left with $p^{(\ell)}$ vertices in the set SETTLED^c. Furthermore, at step ℓ, we only need to guarantee partial correlation tests when conditioned by sets of size ℓ. Now, using the **rCover** on these $p^{(\ell)}$ vertices, one can replicate the argument in the proof of Theorem 2 to obtain a sample requirement given by

$$\frac{\binom{p^{(\ell)}}{2}}{\binom{r}{2}} C_0^\ell \log\left(\frac{p^2}{\delta}\log\left(\frac{p^2}{\delta}\right)\right). \tag{11}$$

Summing up these terms yields the desired guarantee on the sample complexity. ■

REFERENCES

[1] V. A. Marčenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Sbornik: Mathematics*, vol. 1, no. 4, pp. 457–483, 1967.

[2] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Annals of statistics*, pp. 295–327, 2001.

[3] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, "High-dimensional ising model selection using ℓ1-regularized logistic regression," *The Annals of Statistics*, vol. 38, no. 3, pp. 1287–1319, 2010.

[4] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ1-penalized log-determinant divergence," *Electronic Journal of Statistics*, vol. 5, pp. 935–980, 2011.

[5] A. Anandkumar, V. Y. Tan, F. Huang, A. S. Willsky, *et al.*, "High-dimensional structure estimation in ising models: Local separation criterion," *The Annals of Statistics*, vol. 40, no. 3, pp. 1346–1375, 2012.

[6] A. Anandkumar, V. Y. Tan, F. Huang, and A. S. Willsky, "High-dimensional gaussian graphical model selection: Walk summability and local separation criterion," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 2293–2337, 2012.

[7] C. Stosiek, O. Garaschuk, K. Holthoff, and A. Konnerth, "In vivo two-photon calcium imaging of neuronal networks," *Proceedings of the National Academy of Sciences*, vol. 100, no. 12, pp. 7319–7324, 2003.

[8] J. Elson and D. Estrin, *Time synchronization for wireless sensor networks*. IEEE, 2001.

[9] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5721, pp. 523–529, 2005.

[10] S. Keshri, E. Pnevmatikakis, A. Pakman, B. Shababo, and L. Paninski, "A shotgun sampling solution for the common input problem in neural connectivity inference," *arXiv preprint arXiv:1309.3724*, 2013.

[11] S. Turaga, L. Buesing, A. M. Packer, H. Dalgleish, N. Pettit, M. Hausser, and J. Macke, "Inferring neural population dynamics from multiple partial recordings of the same neural circuit," in *Advances in Neural Information Processing Systems*, pp. 539–547, 2013.

[12] G. Dasarathy, A. Singh, M.-F. Balcan, and J. H. Park, "Active learning algorithms for graphical model selection," in *Artificial Intelligence and Statistics*, pp. 1356–1364, 2016.

[13] J. Scarlett and V. Cevher, "Lower bounds on active learning for graphical model selection," *arXiv preprint arXiv:1607.02413*, 2016.

[14] S. L. Lauritzen, *Graphical models*. Oxford University Press, 1996.

[15] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*, vol. 81. MIT press, 2000.

[16] M. Kalisch and P. Bühlmann, "Estimating high-dimensional directed acyclic graphs with the pc-algorithm," *The Journal of Machine Learning Research*, vol. 8, pp. 613–636, 2007.

[17] J. Pearl, *Causality: models, reasoning and inference*, vol. 29. Cambridge Univ Press, 2000.

[18] D. R. Stinson, *Combinatorial designs: constructions and analysis*. Springer Science & Business Media, 2007.

[19] C. J. Colbourn, *CRC handbook of combinatorial designs*. CRC press, 2010.

[20] D. M. Gordon, O. Patashnik, G. Kuperberg, and J. H. Spencer, "Asymptotically optimal covering designs," *journal of combinatorial theory, Series A*, vol. 75, no. 2, pp. 270–280, 1996.

[21] V. Rodl, "On a packing and covering problem," *European Journal of Combinatorics*, vol. 6, no. 1, pp. 69–78, 1985.

[22] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.