Optimal Estimation of Gaussian DAG Models

Ming Gao
University of Chicago

Wai Ming Tai
University of Chicago

Bryon Aragam
University of Chicago

Abstract

We study the optimal sample complexity of learning a Gaussian directed acyclic graph (DAG) from observational data. Our main results establish the minimax optimal sample complexity for learning the structure of a linear Gaussian DAG model in two settings of interest: 1) Under equal variances without knowledge of the true ordering, and 2) For general linear models given knowledge of the ordering. In both cases the sample complexity is $n \asymp q \log(d/q)$, where q is the maximum number of parents and d is the number of nodes. We further make comparisons with the classical problem of learning (undirected) Gaussian graphical models, showing that under the equal variance assumption, these two problems share the same optimal sample complexity. In other words, at least for Gaussian models with equal error variances, learning a directed graphical model is statistically no more difficult than learning an undirected graphical model. Our results also extend to more general identification assumptions as well as subgaussian errors.

1 INTRODUCTION

A significant open question in the literature on structure learning is the optimal sample complexity of learning a directed acyclic graphical model. The problem of deriving upper bounds on the sample complexity for this problem goes back decades (Zuk et al., 2006; Friedman and Yakhini, 1996), and in recent years there has been significant progress (Ghoshal and Honorio, 2017a, 2018; Chen et al., 2019; Park and Raskutti, 2017; Park, 2018; Park and Park, 2019; Park, 2020; Wang and Drton, 2020; Gao et al., 2020; Gao and Aragam, 2021). Nonetheless, despite these upper bounds, a tight characterization of the optimal sample complexity is missing. This is to be contrasted with the situation for learning undirected graphs (UGs), also known as Markov random fields (MRFs), for which optimal rates were established approximately ten years ago (Santhanam and Wainwright, 2012; Wang et al., 2010), alongside similar results for support recovery in linear models (Wainwright, 2009a,b). In fact, this is unsurprising given the connection between these two problems via neighbourhood regression. Unfortunately, learning a directed acyclic graph (DAG) does not reduce to neighbourhood regression as it involves a more difficult order recovery step.

In this paper, we resolve this question for the special case of linear Gaussian DAG models with equal error variances. The identifiability of these models was established in Peters and Bühlmann (2013), and eventually led to the development of several polynomial-time algorithms under the equal variance assumption (Ghoshal and Honorio, 2017a, 2018; Chen et al., 2019; Gao et al., 2020). Nonetheless, it was not known whether or not any of these algorithms were optimal for this precise statistical setting. We will show that a variant of the EQVAR algorithm from Chen et al. (2019) is indeed optimal. This involves the derivation of new lower bounds and a novel analysis of the EQVAR algorithm that sharpens the existing sample complexity upper bound from $O(q^2 \log d)$ to $O(q \log d/q)$, where q is the maximum number of parents in the DAG and d is the number of nodes. This upper bound is optimal up to constants, and allows for the high-dimensional regime with $d \gg n$, where as usual n denotes the sample size. Moreover, in Section 4, we extend this result to the case of general Gaussian models with known ordering. Our results also extend to more general identification assumptions (e.g. allowing for unequal error variances) as well as subgaussian error terms; see Remark 1.

As a problem of independent interest, we further compare the complexity of learning Gaussian graphical models (GGMs) and Gaussian DAG models under the equal variance assumption. Given the additional complexity of the order recovery problem in DAG learning,

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

the folklore has generally been that learning DAGs is harder than learning UGs. Despite this folklore, few results are available to rigorously characterize the hardness of these problems on an equal footing (besides known NP-hardness results for both problems, see Srebro, 2003; Chickering, 1996; Chickering et al., 2004). The equal variance assumption gives us the opportunity to make an apples-to-apples comparison under the same assumptions. As we will show, the optimal sample complexity for *both problems* scales as $O(q \log d/q)$. In other words, learning a DAG is *statistically* no harder than learning a GGM under the equal variance assumption. It is worth emphasizing that this comparison is purely statistical: The *computational* complexity of the algorithm we analyze is exponential in q whereas learning GGMs can be done efficiently; see also Remark 2.

To the best of our knowledge, these are the first results giving a tight characterization of the optimal sample complexity for learning DAG models from observational data.

The rest of this paper is organized as follows: In the remainder of Section 1, we discuss related work and the problem setting. In Sections 2 and 3 we present our main results for learning equal variance DAGs. Then in Section 4 we consider the special case of known ordering, and in Section 5 make further comparisons with learning undirected GGMs. An illustrative simulation study is presented in Section 6 before concluding with some open questions in Section 7.

Notation and preliminaries Given a directed graph $G = (V, E)$ with $|V| = d$ nodes, we make the following standard definitions:

- The parents $\mathrm{pa}(k) = \{\ell : (\ell, k) \in E\}$;
- The descendants $\mathrm{de}(k)$ to which k has at least one directed path;
- The nondescendents $\mathrm{nd}(k) = V \setminus \mathrm{de}(k)$;
- The ancestors $\mathrm{an}(k)$ any of which has at least one directed path to k.

When $(\ell, k) \in E$ we will often write $\ell \to k$ for short. A source node is any $k \in V$ with $\mathrm{an}(k) = \emptyset$. A subgraph $G[V \setminus S]$ is the original graph with nodes in S and edges related to S removed. Given a DAG G, the moralized graph $\mathcal{M}(G)$ is constructed by dropping the orientations of all directed edges and then connecting all nodes within $\mathrm{pa}(k)$ for all $k \in [d]$. A *topological sort* (also called an *ordering*) of a DAG G is an ordering of the nodes V such that $j \to k \implies j \prec k$.

Given a random vector $X = (X_1, \ldots, X_d)$, we say that G is a Bayesian network for X (or more precisely, its joint distribution P), if the following factorization holds:

$$P(X) = \prod_{k=1}^{d} P(X_k \mid \mathrm{pa}(k)). \tag{1}$$

In this case, we abuse notation by identifying the random vector X with the vertex set V, i.e. $V = X = [d] = \{1, 2, \ldots, d\}$. We denote the class of all DAGs with d nodes and at most q parents per node (i.e. in-degree $\leq q$) by $\mathcal{G}_{d,q}$.

1.1 Related Work

To provide context, we begin by reviewing the related problem of learning the structure of an undirected graph (e.g. MRF, GGM, etc.) from data. Early work establishing consistency and rates of convergence includes Meinshausen and Bühlmann (2006); Banerjee et al. (2008); Ravikumar et al. (2010), with information-theoretic lower bounds following in Santhanam and Wainwright (2012); Wang et al. (2010). More recently, sample optimal and computationally efficient algorithms have been proposed (Vuffray et al., 2016; Misra et al., 2020). Part of the reason for the early success of MRFs is owed to the identifiability and convexity of the underlying problems. By contrast, DAG learning is notably *nonidentifiable* and *nonconvex*. This has led to a line of work to better understand identifiability (e.g. Hoyer et al., 2009; Zhang and Hyvärinen, 2009; Peters et al., 2014; Peters and Bühlmann, 2013; Park and Raskutti, 2017) as well as efficient algorithms that circumvent the nonconvexity of the score-based problem (Ghoshal and Honorio, 2017a, 2018; Chen et al., 2019; Gao et al., 2020; Gao and Aragam, 2021). The latter class of algorithms begins by finding a topological sort of the DAG; once this is known the problem reduces to a variable selection problem. Our paper builds upon this line of work.

Other approaches include score-based learning, for which various consistency results are known (van de Geer and Bühlmann, 2013; Bühlmann et al., 2014; Loh and Bühlmann, 2014; Aragam et al., 2015; Nowzohour and Bühlmann, 2016; Nandy et al., 2018; Rothenhäusler et al., 2018; Aragam et al., 2019), but for which optimality results are missing. It is interesting to note that recent work has explicitly connected the equal variance assumption we use here to score-based learning via a greedy search algorithm (Rajendran et al., 2021). We also note here important early work on the constraint-based PC algorithm, which also establishes finite-sample rates under the strong faithfulness assumption (Kalisch and Bühlmann, 2007).

For completeness, we pause for a more detailed comparison with existing sample complexity upper bounds from

the literature. van de Geer and Bühlmann (2013) studied the ℓ_0-penalized MLE and showed that $n/\log n \gtrsim d$ samples suffice, which was later improved to $n \gtrsim s \log d$ (Aragam et al., 2019). Using a different approach, Ghoshal and Honorio (2017a) proved that $n \gtrsim s^4 \log d$ samples suffice, where s is the maximum Markov blanket size or equivalently the size of the largest neighbourhood in the conditional independence graph of X. The dependency on s arises from the way this algorithm uses the inverse covariance matrix $\Gamma = \Sigma^{-1}$. Moreover, their result additionally requires the restricted strong adjacency faithfulness assumption, which we do not impose. In a more recent work, Chen et al. (2019) show that $n \gtrsim q^2 \log d$ samples suffices to learn the ordering of the underlying DAG, but do not establish results for learning the full DAG. Similar to our work, Chen et al. (2019) do not make any faithfulness or restricted faithfulness-type assumptions. We note also the work of Park (2020) that establishes rates of convergence assuming $n > d$, but which precludes the high-dimensional scenario $d \gg n$. For comparison, we improve these existing bounds to $n \gtrsim q \log d/q$ for the full DAG and moreover prove a matching lower bound (up to constants). Ghoshal and Honorio (2017b) have also established lower bounds for a range of DAG learning problems up to Markov equivalence. For example, their lower bound for sparse Gaussian DAGs is $\sigma^2(q \log d/q + q^2/d)/(\sigma^2 + 2w_{\max}^2(1 + w_{\max}^2))$, where $w_{\max}$ depends on the ℓ_2 norms of the regression coefficients. By contrast, our lower bounds depend instead on $(\beta_{\min}, M)$ (cf. (4), (5) for definitions).

1.2 Problem Setting

Although our results extend to more general settings, we focus on the special case of linear Gaussian Bayesian networks under equal variances. See Remark 1 for a discussion of generalizations. Specifically, let $X = (X_1, \dots, X_d)$,

$$X_k = \langle \beta_k, X \rangle + \epsilon_k, \qquad \operatorname{var}(\epsilon_k) \equiv \sigma^2, \qquad \mathbb{E}[\epsilon_k] = 0,$$

that is, each node is a linear combination of its parents with independent Gaussian noise. The variance σ^2 of each noise term is assumed to be the same; this is the key identifiability assumption that is imposed on the model.

More compactly, let $B = (\beta_{jk})$ denote the coefficient matrix such that $\beta_{jk} \neq 0$ is equivalent to the existence of the edge $j \to k$. Then letting $\epsilon = (\epsilon_1, \dots, \epsilon_d)$ we have

$$X = B^\top X + \epsilon. \tag{2}$$

The matrix B defines a graph $G = G(B)$ by its nonzero entries, i.e.

$$G(B) = (V, E(B)), \quad \begin{cases} V = X, \\ E(B) = \{(j,k) : \beta_{jk} \neq 0\}. \end{cases}$$

Whenever G is acyclic, it is easy to check that (1) holds, and hence G is a Bayesian network for X. In the sequel we assume that G is acyclic.

The following quantities are important in the sequel: The largest in-degree of any node is denoted by q, i.e.

$$q = q(B) := \sup_k |\operatorname{pa}(k)| = \sup_k |\operatorname{supp}(\beta_k)|. \tag{3}$$

The absolute values of the coefficients are lower bounded by $\beta_{\min}$, i.e.

$$\beta_{\min} = \beta_{\min}(B) := \min\{|\beta_{jk}| : \beta_{jk} \neq 0\}. \tag{4}$$

Furthermore, assume the covariance matrix $\Sigma = \mathbb{E}[XX^\top]$ satisfies

$$M^{-1} \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq M \tag{5}$$

for some $M > 1$.

Let the class of distributions satisfying the above conditions (3), (4), and (5) be denoted by $\mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)$. For any $F \in \mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)$ we have

$$\Sigma = \sigma^2 (I - B)^{-T} (I - B)^{-1}. \tag{6}$$

This follows directly from (2) and $\operatorname{cov}(\varepsilon) = \sigma^2 I$. Since the DAG is identifiable from the observational distribution, we denote $G(F)$ to be the DAG associated with the distribution $F \in \mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)$. Finally, we introduce the variance gap:

$$\Delta \equiv \min_k \min_{\substack{A \subseteq \operatorname{nd}(k) \\ \operatorname{pa}(k) \setminus A \neq \emptyset \\ A \subseteq \operatorname{nd}(\operatorname{pa}(k) \setminus A)}} \mathbb{E}_A \operatorname{var}(X_k \mid A) - \sigma^2$$

where the subscript indicates that the expectation is being taken over the random variables in A. This is the missing conditional variance on ancestors if not all the parents are conditioned on, which serves as the identifiability signal for the main algorithm. It turns out it can be explicitly expressed in terms of the edge coefficient and noise variance:

Lemma 1.1. $\Delta = \beta_{\min}^2 \sigma^2 > 0$.

The proof of this lemma is a straightforward calculation; see Appendix E for details.

Remark 1. Both our upper and lower bounds can be generalized as follows: Although we assume Gaussianity for simplicity, everything extends to subgaussian families without modification. This is because the upper bound analysis relies only on subgaussian concentration, and the lower bounds easily extend to subgaussian

models (i.e. since subgaussian also contains Gaussian as a subclass). Furthermore, the equal variance assumption can be relaxed to more general settings as long as B can be identified by Algorithm 1. Examples include (a) the "unequal variance" condition from Ghoshal and Honorio (2018) (see Assumption 1 therein) and (b) if noise variances are known up to some ratio as in Loh and Bühlmann (2014). Moreover, both of these identifiability conditions include the naive equal variance condition as a special case, hence the lower bounds still apply. This implies more general optimality results for a wider class of Bayesian networks.

2 ALGORITHM AND UPPER BOUND

We begin with stating the sufficient conditions on the sample size for DAG recovery under the equal variance assumption. Namely, we present an algorithm (Algorithm 1) that takes samples from a distribution $F \in \mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)$ as an input and returns the DAG $G(F)$ with high probability. We first state an upper bound for the number of samples required in Algorithm 1 in Theorem 2.1.

Theorem 2.1. *For any $F \in \mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)$, let $\widehat{G}$ be the DAG return by Algorithm 1 with $\gamma = \Delta/2$. If*

$$n \gtrsim \frac{M^5}{\Delta}\left(q \log \frac{d}{q} + \log \delta\right),$$

then $P(\widehat{G} = G(F)) \gtrsim 1 - \delta$.

The proof of this result can be found in Appendix A. The obtained sample complexity depends on the variance gap Δ, which serves as signal strength, and covariance matrix norm M, which shows up when estimating conditional variances. Treating these parameters as fixed, the sample complexity scales with $q \log d/q$. The order of this complexity arises mainly from counting all possible conditioning sets. The proof follows the correctness of Algorithm 1, which consists of two main steps: *Learning ordering* and *Learning parents*.

Algorithmically, the first step is the same as Chen et al. (2019), however, our analysis is sharper: We separately analyze the estimation of each conditional variance directly rather than indirectly via the inverse covariance matrix. This leads to the improved sample complexity in Theorem 2.1. This step is where we exploit the equal variance assumption: The conditional variance $\mathrm{var}(X_k \,|\, C)$ of each random variable X_k is a constant σ^2 if and only if $\mathrm{pa}(k) \subseteq C$ for any nondescendant set C. This implies that the variance of any non-source node in the corresponding subgraph would be larger than σ^2. Therefore, when all conditional variances v_{kC} are correctly estimated with error within some small factor of the signal Δ (see Lemma A.1), identifying the node with the smallest σ_k yields a source node in the underlying subgraph. Recall that σ_k is the minimum variance estimation that node k can achieve conditioned on at most q nondescendants. Finally, recursively applying the above step leads to a valid topological sort.

Algorithm 1 LearnDAG algorithm

Input: Sample covariance matrix $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$, backward phase threshold γ
Output: $\widehat{G}$.

1. *Learning Ordering*:
   (a) Initialize empty ordering $\widehat{\tau} = []$
   (b) Denote $v_{kC} := \widehat{\Sigma}_{kk} - \widehat{\Sigma}_{kC}\widehat{\Sigma}_{CC}^{-1}\widehat{\Sigma}_{Ck}$
   (c) For $j = 1, 2, \ldots, d$
      i. Calculate $\sigma_k := \min_{C \subseteq \widehat{\tau}, |C| \le q} v_{kC}$
      ii. Update $\widehat{\tau} = [\widehat{\tau}, \arg\min_k \sigma_k]$
2. *Learning Parents*:
   (a) Initialize empty graph $\widehat{G} = \emptyset$
   (b) For $j = 1, 2, \ldots, d$
      i. Let $C_j = \arg\min_{C \subseteq \widehat{\tau}_{[1:j-1]}, |C| \le q} v_{\widehat{\tau}_j C}$
      ii. Set

$$\mathrm{pa}_{\widehat{G}}(\widehat{\tau}_j) = C_j \setminus \left\{ i \in C_j \,\middle|\, |v_{\widehat{\tau}_j C_j} - v_{\widehat{\tau}_j C_j \setminus i}| \le \gamma \right\}$$

3. *Return* $\widehat{G}$

In the second step, given the correct ordering, we use Best Subset Selection (BSS) along with a backward phase to learn the parents for each node. Note that BSS is already applied in the step **1.(c).i.** of Algorithm 1 and the candidate set C_j can be stored for each $\widehat{\tau}_j$, thus there is no additional computational cost. Again, when all conditional variances are well approximated by their sample counterpart v_{kC}, C_j would be a superset of the true parents of current node $\widehat{\tau}_j$, otherwise the minimum would not be achieved. Meanwhile, removal of any true parent $i \in \mathrm{pa}(\widehat{\tau}_j)$ from C_j would induce a significant change in conditional variances, which is quantified by Δ as well. This is used to design a tuning parameter γ in the backward phase for pruning C_j. Finally, we show the tail probability of conditional variance estimation error is well bounded to get the desired sample complexity in Lemma A.2.

When the true variance gap Δ is unknown, we can select the tuning parameter γ according to the following theorem:

Theorem 2.2. *For any $F \in \mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)$, let $\widehat{G}$*

be the DAG return by Algorithm 1 with tuning parameter

$$\gamma \asymp \frac{2M^5 q \log d/q}{n}.$$

If

$$n \gtrsim \frac{M^5}{\Delta} q \log \frac{d}{q},$$

then $P(\widehat{G} = G(F)) \gtrsim 1 - \exp(-q \log d/q)$.

The proof of this result can be found in Appendix A.3.

Remark 2. A computationally attractive alternative to BSS is the Lasso, or ℓ_1-regularized least squares regression. Unlike BSS, the Lasso requires restrictive incoherence-type conditions. If these conditions (or related conditions such as irrepresentability) are imposed on each parent set, then the Lasso can be used to recover the full DAG under a similar sample complexity scaling (see e.g. Wainwright, 2009a). Furthermore, these incoherence-type conditions can be further relaxed through the use of nonconvex regularizers such as the MCP (Zhang, 2010) or SCAD (Fan and Li, 2001); see also Loh and Wainwright (2014).

3 LOWER BOUND

We will now present the necessary conditions on the sample size for DAG recovery under the equal variance assumption. Namely, we present a subclass of $\mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)$ such that any estimator that successfully recovers the underlying DAG in this subclass with high probability requires a prescribed minimum sample size. For this, we rely on Fano's inequality, which is a standard technique for establishing necessary conditions for graph recovery. See Corollary B.2 for the exact variant we use.

Theorem 3.1. *Assume* $q \leq d/2$. *If*

$$n \leq (1 - 2\delta) \max\left(\frac{\log d}{\beta_{\min}^2}, \frac{q \log d/q}{M^2 - 1}\right)$$

then for any estimator $\widehat{G}$,

$$\sup_{F \in \mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)} P(\widehat{G} \neq G(F)) \geq \delta - \frac{\log 2}{\log d}.$$

In Theorem 3.1, we state two sample complexity lower bounds: $q \log d/q/(M^2 - 1)$ and $\log d/\beta_{\min}^2$. Though the first one dominates when fixing other parameters as constants, the second one reveals the dependency on the signal strength $\beta_{\min}$. This can also be seen from the upper bound in Theorem 2.1 by replacing $\Delta = \beta_{\min}^2 \sigma^2$ (cf. Lemma 1.1). We will present two ensembles for each bound. The first one is the whole set of sparse DAGs $\mathcal{G}_{d,q}$, and the second is the set of DAGs with only one edge, which is constructed to study the dependency on the coefficient $\beta_{\min}$.

For the first ensemble, we borrow the ideas from Santhanam and Wainwright (2012) to count the number of DAGs in $\mathcal{G}_{d,q}$. The only difference is we consider DAGs instead of undirected graphs. Also, it is easy to bound the KL divergence between any two distributions in this ensemble due to Gaussianity, which would lead to the bound $q \log d/q/(M^2 - 1)$. For the second ensemble, it is easy to count the size of this ensemble since we consider the DAGs with only one edge. Then all possibilities of any different pair of edges are analyzed to bound the KL divergence. This ensemble gives us the bound $\log d/\beta_{\min}^2$. The detailed proof can be found in Appendix B.

For comparison, Ghoshal and Honorio (2017b) previously established a lower bound for general Gaussian DAGs (i.e. without equal variances) of

$$\Omega\Big(\sigma^2 \frac{q \log d/q + q^2/d}{\sigma^2 + 2w_{\max}^2(1 + w_{\max}^2)}\Big),$$

where $w_{\max}$ depends on the ℓ_2 norms of the regression coefficients. Holding σ^2 constant, $w_{\max}^2$ is similar to the maximum marginal variance of the variables, which is comparable with our definition of M as an upper bound on $\|\Sigma\|$. By contrast, under the stronger assumption of equal variances, our lower bound is

$$\Omega\Big(\frac{\log d}{\beta_{\min}^2} \vee \frac{q \log d/q}{M^2 - 1}\Big),$$

which is a comparable lower bound. This is interesting since by restricting to simpler equal variance models (i.e. a smaller family), the problem should become easier, however, our analysis shows this is not the case. In particular, our lower bounds do not follow from previous work, and require a slightly different analysis as outlined in Appendix B.

4 RECONSTRUCTING A DAG FROM ITS ORDERING

The second step of Algorithm 1 may be of interest in its own right: Abstracted away, this step seeks to reconstruct a DAG from knowledge of its topological sort. We claim that the second step of Algorithm 1 is in fact sample optimal for learning the parents of each node (and hence all of G) given the true ordering of G under more general assumptions.

Dropping the equal variance condition from $\mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)$, define $\sigma_k^2 := \mathrm{var}(\epsilon_k)$ and let $\overline{\mathcal{F}}_{d,q}(\beta_{\min}, \sigma_{\max}^2, M)$ denote the class of Gaussian distributions such that (3), (4), and (5) hold and

$$\sup_k \sigma_k^2 \leq \sigma_{\max}^2,$$

i.e. σ_k^2 is allowed to depend on k. Note that $\mathcal{F}_{d,q}(\beta_{\min}, \sigma^2_{\max}, M) \subset \overline{\mathcal{F}}_{d,q}(\beta_{\min}, \sigma^2_{\max}, M)$. Furthermore, we modify the definition of the variance gap for $\overline{\mathcal{F}}_{d,q}(\beta_{\min}, \sigma^2_{\max}, M)$ as follows:

$$\overline{\Delta} \equiv \min_k \min_{\substack{A \subseteq \mathrm{nd}(k) \\ \mathrm{pa}(k) \setminus A \neq \emptyset}} \mathbb{E}_A \operatorname{var}(X_k \,|\, A) - \sigma_k^2 \,.$$

Finally, given a known topological sort τ of G, let $\widehat{G}(\tau)$ be the DAG returned by the second step of Algorithm 1 with $\gamma = \overline{\Delta}/2$.

Remark 3. As with the rest of our results, these result extend to subgaussian models without issue. See Remark 1.

Using the second part of Lemma A.1, Lemma A.2 and following the proof in Appendix A.2, we have an upper bound on the sample complexity for recovering G from its ordering:

Proposition 4.1. *For any $F \in \overline{\mathcal{F}}_{d,q}(\beta_{\min}, \sigma^2_{\max}, M)$, given a valid topological sort τ of G, let $\widehat{G}(\tau)$ be the DAG returned by the second step of Algorithm 1 with $\gamma = \overline{\Delta}/2$. If*

$$n \gtrsim \frac{M^5}{\overline{\Delta}} \left(q \log \frac{d}{q} + \log \delta \right),$$

then $P(\widehat{G}(\tau) = G(F) \,|\, \tau) \gtrsim 1 - \delta$.

The "given τ" in the probability is to emphasize that the estimator has the access to the true ordering τ.

Unsurprisingly, this approach of using best subset selection with a backwards phase is indeed optimal: We have a matching lower bound (up to constants).

Proposition 4.2. *If*

$$n \leq \frac{\sigma^2_{\max}}{8 M \beta^2_{\min}} q \log \frac{d}{q} \,,$$

then given the knowledge of true ordering τ of DAG G, for any estimator $\widehat{G}$,

$$\sup_{F \in \overline{\mathcal{F}}_{d,q}(\beta_{\min}, \sigma^2_{\max}, M)} P(\widehat{G} \neq G(F) \,|\, \tau) \geq \frac{1}{2} \,.$$

The proof uses known lower bounds from the sparse support recovery literature (Wainwright, 2009b); see Appendix C for details.

This more general optimality result for the second step shows that it is only in the first step (learning parents) that the equal variance assumption is operational. Moreover, although it may be possible to improve the sample complexity of the second step for the smaller class $\mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)$, since the sample complexity upper bound of the first step of Algorithm 1 matches the lower bound for recovering the whole graph, such improvements would not change the optimal sample complexity for learning G.

5 COMPARISON WITH UNDIRECTED GRAPHS

Our results on DAG learning under equal variances raise an interesting question: *Is learning an equal variance DAG statistically more difficult than learning its corresponding Gaussian graphical model (i.e. inverse covariance matrix)?* This is especially intriguing given the folklore intuition that learning a DAG is more difficult than learning an undirected graph (UG). In fact, it is common to learn an undirected graph *first* as a pre-processing step in order to reduce the search space and sample complexity for DAG learning (Perrier et al., 2008; Loh and Bühlmann, 2014; Bühlmann et al., 2014; Aragam et al., 2019). In this section we explore this question and show that in fact, at least in the special case of equal variance Gaussian models, the sample complexity of both problems is the same.

5.1 Gaussian Graphical Models

First, let us recall some basics about undirected graphical models, also known as Markov random fields (MRFs). When $X \sim \mathcal{N}(0, \Sigma)$ as in this paper, an MRF can be read off from the inverse covariance matrix $\Gamma = (\gamma_{jk}) := \Sigma^{-1}$. More precisely, the zero pattern of Γ defines an undirected graph $U = U(\Gamma)$ that is automatically an MRF for X:

$$U(\Gamma) = (V, E(\Gamma)), \quad \begin{cases} V = X, \\ E(\Gamma) = \{(j,k) : \gamma_{jk} \neq 0\}. \end{cases}$$

Let $\mathrm{ne}(k) = \{\ell \in V : (\ell, k) \in E\}$ be the neighbours of node k, i.e. they are connected by some edge. Note the distinction between the parents of k in a directed graph vs. the neighbours of k in an undirected graph. This model is often referred as the Gaussian graphical model (GGM).

Wang et al. (2010) showed that the optimal sample complexity for learning a GGM is $n \asymp s \log d$, where

$$s := \max_k |\mathrm{ne}(k)| \tag{7}$$

is the degree of U or maximum neighborhood size, and Misra et al. (2020) developed an efficient algorithm that matches this information-theoretic lower bound. Given $F \in \mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)$, let $U(F)$ be the undirected graph induced by the covariance matrix of F (cf. 6). It follows that for any $F \in \mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)$ we can learn the structure $U(F)$ with $\Theta(s \log d)$ samples. Note that this sample complexity scales with s instead of q.

Example 1. Consider the DAGs G_1 and G_2 in Figure 1. In G_1, we have $q = O(d)$ since T has d parents, whereas in G_2 we have $q = O(1)$ since each S_k has only one parent. Thus, we expect that learning G_1

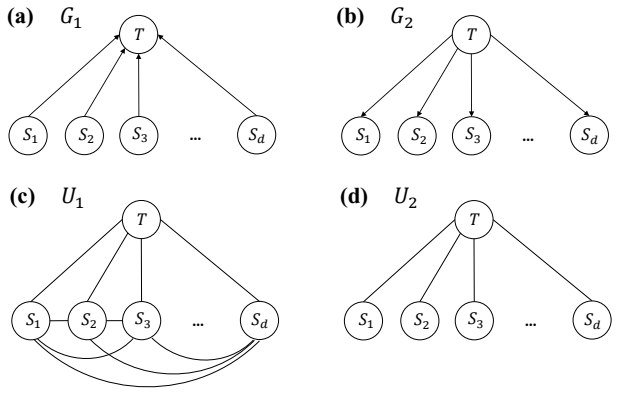

Figure 1: Illustrating examples to compare sample complexities for learning DAG and UG. **(a) & (b):** The original DAGs; **(c) & (d):** The UGs generated under equal variance condition by DAGs in (a) & (b) respectively.

will require $\Theta(d)$ samples and learning G_2 will require $\Theta(\log d)$ samples. By comparison, the UGs associated with each model, given by U_1 and U_2 have $s = d$, and hence if we use the previous approaches to learn each U_k we will need $\Theta(d \log d)$ samples each. Of course, this is to be expected: One should expect that a specialized estimator that exploits the structure of the family $\mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)$ to perform better.

5.2 Optimal Estimation of Equal Variance GGMs

Example 1 shows that there is a gap between existing "universal" algorithms for learning GGMs (i.e. algorithms that do not exploit the equal variance assumption) and the sample complexity for learning equal variance DAGs. A natural question then is: *What is the optimal sample complexity for learning the structure of $U(F)$ for any $F \in \mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)$?*

We begin by establishing a lower bound that matches the lower bound in Theorem 3.1 (up to constants):

Theorem 5.1. *Assume $q \le d/2$. If*

$$n \le \max\left(\frac{2(1-\delta)\log d}{\beta_{\min}^2}, \frac{(1-2\delta)q\log d/q}{M^2-1}\right),$$

then for any estimator $\widehat{U}$,

$$\sup_{F \in \mathcal{F}_{d,q}(\beta_{\min},\sigma^2,M)} P(\widehat{U} \neq U(F)) \ge \delta - \frac{1}{\log d}.$$

The proof is deferred to Appendix D.

To derive an upper bound for this problem, we use the well-known trick of moralization; see Lauritzen (1996) for details. Since $F \in \mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)$, we can first learn the DAG $G = G(F)$ via Algorithm 1. Given the output $\widehat{G}$, we then form the moralized graph and define $\widehat{U} := \mathcal{M}(\widehat{G})$. See Algorithm 2.

Algorithm 2 LearnUG algorithm

Input: Sample covariance matrix $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$, backward phase threshold γ
Output: $\widehat{U}$.

1. *Learn DAG*: Let $\widehat{G} = \textsc{LearnDAG}(\widehat{\Sigma}, \gamma)$
2. *Moralization*: Set $\widehat{U} = \mathcal{M}(\widehat{G})$
3. *Return* $\widehat{U}$

This approach is justified by a result due to Loh and Bühlmann (2014). First, we need the following condition:

Condition 1. Let precision matrix $\Gamma = \Sigma^{-1} = [\mathrm{cov}(F)]^{-1}$, $\Gamma_{ij} = 0 \Leftrightarrow \beta_{ij} = 0$ and $\beta_{ik}\beta_{jk} = 0$ for all $k \neq i, j$.

When we sample nonzero entries of B from some continuous distribution independently, Condition 1 is satisfied except on a set of Lebesgue measure zero. For example, it is easy to check that the examples in Figure 1 satisfy this condition. Under this condition, moralization is guaranteed to return U:

Lemma 5.2 (Theorem 2, Loh and Bühlmann, 2014)**.** *If Condition 1 holds, then $U = \mathcal{M}(G(F))$.*

Under Condition 1, we have the following upper bound, which matches the lower bound in Theorem 5.1:

Theorem 5.3. *Assuming Condition 1, for any $F \in \mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)$ let $\widehat{U}$ be the UG returned by Algorithm 2 with $\gamma = \Delta/2$. If*

$$n \gtrsim \frac{M^5}{\Delta}\left(q\log\frac{d}{q} + \log\delta\right),$$

then $P(\widehat{U} = U(F)) \gtrsim 1 - \delta$.

The proof is straightforward by Theorem 2.1 and Lemma 5.2. This answers the question proposed at the beginning of this section: *Under the equal variance assumption, learning a DAG is no harder than learning its corresponding UG.*

Remark 4. It is an interesting question whether or not similar results hold without Condition 1; i.e. is there a direct estimator of U—not based on moralizing a DAG—that matches the sample complexity of learning an equal variance DAG?

Remark 5. To compare Theorem 5.1 with previous work, Wang et al. (2010) showed the optimal sample complexity is $\Theta(s \log d/\lambda^2)$ where $\lambda^2 =$

$\min_{(s,t)\in E} \Gamma_{st}^2/(\Gamma_{ss}\Gamma_{tt})$, which is dominated by $\beta_{\min}^2$ when $\beta_{\min}$ is small. For equal variance GGMs, our result gives $\Omega(q \log d/q/M^2 + \log d/\beta_{\min}^2)$. Under Condition 1, s is always greater than q, so our new lower bound is strictly smaller, along with a matching upper bound that shows the s dependence for general GGMs is suboptimal for equal variance GGMs. Another related work (Cai et al., 2016) derives lower bounds on precision matrix estimation under certain matrix norms, which is distinct from the graph recovery problem we consider in this work. For comparison, put in our setting, their lower bound becomes $\Omega(s^2 \log d)$, which again depends on s instead of q.

6 EXPERIMENTS

To illustrate the effectiveness of Algorithm 1, we report the results of a simulation study. We note that existing variants of Algorithm 1 have been compared against other approaches such as greedy DAG search (GDS, Peters and Bühlmann, 2013), see Chen et al. (2019) for details. In our experiments, we controlled the number of parents when generating the DAG, fix all noise variances to be the same, and sample nonzero entries of β_k's uniformly from given intervals.

6.1 Experiment Settings

To generate random DAGs, we first randomly permute $[d]$ to obtain an ordering τ. Then for each $j \in [d]$, we randomly draw a set S of q nodes from $\tau_{[1:j-1]}$ and set

$$\mathrm{pa}_G(\tau_j) = \begin{cases} \tau_{[1:j-1]}, & j \le q+1 \\ S, & \text{otherwise.} \end{cases}$$

We then generate random nonzero coefficients according to $\beta_j \sim \mathrm{Rad} \times \mathrm{Unif}(0.5, 1)$, where Rad is a Rademacher random variable. Finally, we generate data by the resulting Gaussian linear model:

$$X_k = \beta_k^\top X + \epsilon_k, \qquad \epsilon_k \sim \mathcal{N}(0, \sigma^2) \quad \forall k$$

with $\sigma = 0.3$.

We consider graphs with $d \in \{20, 30, \ldots, 90\}$ nodes and in-degree $q \in \{2, 3, 4\}$. For each setting, the total number of replications is $N = 100$. For each replication, we generate a random graph and a dataset with sample size $n \in \{80, 160, \ldots, 560\}$. Finally, we report $\#\{\widehat{G} = G\}/N$ to approximate $P(\widehat{G} = G)$.

6.2 Implementation

We implement the *learning ordering* phase of Algorithm 1 using code from Chen et al. (2019)[1], which inputs the oracle in-degree q and outputs a topological sort. For the *learning parents* phase of Algorihm 1, we use the R package `leaps` (Lumley and Lumley, 2013) for Best Subset Selection with BIC.

The experiments were conducted on an internal cluster using an Intel E5-2680v4 2.4GHz CPU with 64 GB memory.

6.3 Results

The results are shown in Figure 2. As expected, the probability of successfully recovering true DAG goes to one quickly across different settings of the number of nodes d and maximum in-degree q. Since Best Subset Selection is computationally expensive, we are not able to examine higher dimensions systematically. Nonetheless, to test higher dimensional cases, we checked several (random) cases for $d = 200, 300, 500$, and the results shows with 90% chance the DAG is successfully recovered for moderate sample sizes $n = 480, 560, 900$.

7 CONCLUSION

In this paper, we derived the optimal sample complexity for learning Gaussian linear DAG models under an equal variance condition that has been extensively studied in the literature. These results extend to subgaussian errors under similar assumptions as well as more general models with unequal variances as long as the DAG remains identifiable by the proposed algorithm, which is easy to implement and simulations corroborate our theoretical findings. We also investigated the sub-problem of learning a linear DAG from its ordering and made comparisons with the classical problem of learning GGMs, showing the sample complexity of both problems is the same.

We conclude with some open questions. Although our algorithm is sample optimal, it is not computationally efficient. As noted in Remark 2, an easy fix is to use ℓ_1-regularization, however, this would require imposing restrictive incoherence assumptions. It would be interesting to find efficient algorithms without such conditions, along the lines of Misra et al. (2020) for GGMs. We conjecture that there is such a polynomial-time algorithm that achieves the optimal sample complexity bound without such restrictive conditions. It also is not known whether or not score-based approaches (e.g. van de Geer and Bühlmann, 2013; Loh and Bühlmann, 2014; Nandy et al., 2018; Aragam et al., 2019; Rajendran et al., 2021) are sample optimal. Another interesting question posed in Remark 4 is whether or not there is a moralization-free algorithm that achieves the optimal sample complexity for learning $U(F)$ established in Theorem 5.1. This would allow Condition 1

[1]The code can be found at `https://github.com/WY-Chen/EqVarDAG/blob/master/R/EqVarDAG_HD_TD.R`.

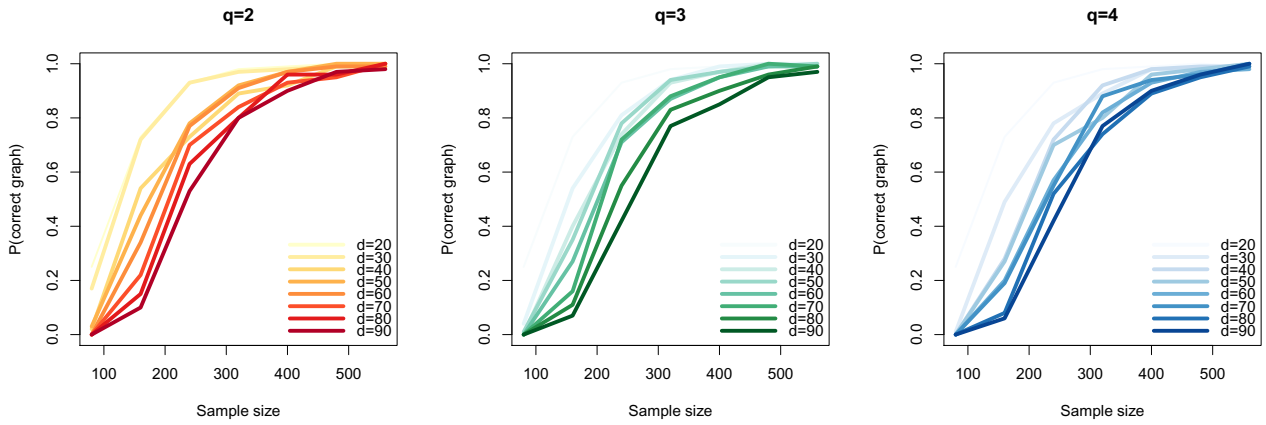

Figure 2: $P(\widehat{G} = G)$ v.s. sample size n. Different in-degree q are presented by different colors. The darker shade indicates large number of nodes d.

to be relaxed or removed entirely.

Finally, it would be of interest to generalize the results in Section 5 to more general families, i.e. beyond equal variances and its generalizations (Ghoshal and Honorio, 2018). This would require the derivation of new identifiability conditions, as in Gao and Aragam (2021) and Rajendran et al. (2021).

Acknowledgements

B.A. gratefully acknowledges support via NSF IIS-1956330, NIH R01GM140467, and the Robert H. Topel Faculty Research Fund at the University of Chicago Booth School of Business.

References

B. Aragam, A. A. Amini, and Q. Zhou. Learning directed acyclic graphs with penalized neighbourhood regression. *arXiv:1511.08963*, 2015.

B. Aragam, A. Amini, and Q. Zhou. Globally optimal score-based learning of directed acyclic graphs in high-dimensions. In *Advances in Neural Information Processing Systems 32*, pages 4450–4462. 2019.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

P. Bühlmann, J. Peters, and J. Ernest. CAM: Causal additive models, high-dimensional order search and penalized regression. *Annals of Statistics*, 42(6):2526–2556, 2014.

T. T. Cai, W. Liu, and H. H. Zhou. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *The Annals of Statistics*, 44(2): 455–488, 2016.

W. Chen, M. Drton, and Y. S. Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 09 2019. ISSN 0006-3444. doi: 10.1093/biomet/asz049.

D. M. Chickering. Learning Bayesian networks is NP-complete. In *Learning from data*, pages 121–130. Springer, 1996.

D. M. Chickering, D. Heckerman, and C. Meek. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

N. Friedman and Z. Yakhini. On the sample complexity of learning bayesian networks. In *Uncertainty in Artifical Intelligence (UAI)*, 02 1996.

M. Gao and B. Aragam. Efficient bayesian network structure learning via local markov boundary search. *Advances in Neural Information Processing Systems*, 34, 2021.

M. Gao, Y. Ding, and B. Aragam. A polynomial-time algorithm for learning nonparametric causal graphs. *Advances in Neural Information Processing Systems*, 33, 2020.

A. Ghoshal and J. Honorio. Learning identifiable gaussian bayesian networks in polynomial time and sample complexity. In *Advances in Neural Information Processing Systems 30*, pages 6457–6466. 2017a.

A. Ghoshal and J. Honorio. Information-theoretic limits of Bayesian network structure learning. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 767–775, Fort Lauderdale, FL, USA, 20–22 Apr 2017b. PMLR.

A. Ghoshal and J. Honorio. Learning linear structural equation models in polynomial time and sample complexity. In A. Storkey and F. Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1466–1475, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

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

P.-L. Loh and P. Bühlmann. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15:3065–3105, 2014.

P.-L. Loh and M. J. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. *arXiv preprint arXiv:1412.5632*, 2014.

T. Lumley and M. T. Lumley. Package 'leaps'. *Regression subset selection. Thomas Lumley Based on Fortran Code by Alan Miller. Available online: http://CRAN. R-project. org/package= leaps (Accessed on 18 March 2018)*, 2013.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

S. Misra, M. Vuffray, and A. Y. Lokhov. Information theoretic optimal learning of gaussian graphical models. In *Conference on Learning Theory*, pages 2888–2909. PMLR, 2020.

P. Nandy, A. Hauser, and M. H. Maathuis. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A): 3151–3183, 2018.

C. Nowzohour and P. Bühlmann. Score-based causal learning in additive noise models. *Statistics*, 50(3): 471–485, 2016.

G. Park. Learning generalized hypergeometric distribution (ghd) dag models. *arXiv preprint arXiv:1805.02848*, 2018.

G. Park. Identifiability of additive noise models using conditional variances. *Journal of Machine Learning Research*, 21(75):1–34, 2020.

G. Park and S. Park. High-dimensional poisson structural equation model learning via ℓ_1-regularized regression. *Journal of Machine Learning Research*, 20 (95):1–41, 2019.

G. Park and G. Raskutti. Learning quadratic variance function (QVF) dag models via overdispersion scoring (ODS). *The Journal of Machine Learning Research*, 18(1):8300–8342, 2017.

E. Perrier, S. Imoto, and S. Miyano. Finding optimal bayesian network given a super-structure. *Journal of Machine Learning Research*, 9(Oct):2251–2286, 2008.

J. Peters and P. Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2013.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1): 2009–2053, 2014.

G. Rajendran, B. Kivva, M. Gao, and B. Aragam. Structure learning in polynomial time: Greedy algorithms, bregman information, and exponential families. *Advances in Neural Information Processing Systems*, 34, 2021.

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

D. Rothenhäusler, J. Ernest, P. Bühlmann, et al. Causal inference in partially linear structural equation models. *The Annals of Statistics*, 46(6A):2904–2938, 2018.

N. P. Santhanam and M. J. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012.

N. Srebro. Maximum likelihood bounded tree-width markov networks. *Artificial intelligence*, 143(1):123–138, 2003.

S. van de Geer and P. Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *Annals of Statistics*, 41(2):536–567, 2013.

M. Vuffray, S. Misra, A. Y. Lokhov, and M. Chertkov. Interaction screening: Efficient and sample-optimal learning of ising models. *arXiv preprint arXiv:1605.07252*, 2016.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (Lasso). *Information Theory, IEEE Transactions on*, 55(5):2183–2202, 2009a.

M. J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE transactions on information theory*, 55(12):5728–5741, 2009b.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

W. Wang, M. J. Wainwright, and K. Ramchandran. Information-theoretic bounds on model selection for gaussian markov random fields. In *2010 IEEE International Symposium on Information Theory*, pages 1373–1377. IEEE, 2010.

Y. S. Wang and M. Drton. High-dimensional causal discovery under non-gaussianity. *Biometrika*, 107(1): 41–59, 2020.

B. Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38 (2):894–942, 2010.

K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pages 647–655. AUAI Press, 2009.

O. Zuk, S. Margel, and E. Domany. On the number of samples needed to learn the correct structure of a bayesian network. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 560–567, 2006.

Supplementary Material: Optimal Estimation of Gaussian DAG Models

A PROOF OF UPPER BOUND

A.1 Preliminaries

We first show that if all conditional variances are estimated sufficiently well, then Algorithm 1 is able to identify the true DAG.

Lemma A.1. *If for all $k \in V$ and $C \subset V \setminus \{k\}$, $|C| \le q$,*

$$|\operatorname{var}(X_k \mid X_C) - v_{kC}| \le \Delta/4$$

then $\widehat{G} = G$.

Proof. We start by showing that $\widehat{\tau}$ is a valid ordering for G, which is equivalent to saying $\widehat{\tau}_j$ is a source node of the subgraph $G[V \setminus \widehat{\tau}_{[1:j-1]}]$ for all j. We proceed by induction. For $j = 1$, it reduces to compare marginal variances.

$$\begin{cases} \operatorname{var}(X_k) = \sigma^2 & k \text{ is a source node} \\ \operatorname{var}(X_\ell) > \sigma^2 + \Delta & \ell \text{ is not a source node} \end{cases}$$

For any non-source node ℓ and any source node k,

$$\begin{aligned} \sigma_\ell &= v_{\ell\emptyset} \\ &\ge \operatorname{var}(X_\ell) - \Delta/4 \\ &> \Delta \times \frac{3}{4} + \sigma^2 \\ &> \Delta/4 + \sigma^2 \\ &= \operatorname{var}(X_k) + \Delta/4 \\ &\ge v_{k\emptyset} = \sigma_k \end{aligned}$$

Thus k is preferred over ℓ. Given that $\widehat{\tau}_{[1:j-1]}$ correctly identified, by the equal variance assumption,

$$\begin{aligned} &\min_{C \subseteq \widehat{\tau}_{[1:j-1]}, |C| \le q} \operatorname{var}(X_k \mid C) = \operatorname{var}(X_k \mid \operatorname{pa}(k)) = \sigma^2 \\ &\quad \text{if } k \text{ is a source node of } G[V \setminus \widehat{\tau}_{[1:j-1]}] \\ &\min_{C \subseteq \widehat{\tau}_{[1:j-1]}, |C| \le q} \operatorname{var}(X_\ell \mid C) > \Delta + \sigma^2 \\ &\quad \text{if } \ell \text{ is not a source node of } G[V \setminus \widehat{\tau}_{[1:j-1]}] \end{aligned}$$

Therefore, for any ℓ that is not a source node and for any k that is a source node,

$$\begin{aligned} \sigma_\ell &= \min_{C \subseteq \widehat{\tau}_{[1:j-1]}, |C| \le q} v_{\ell C} \\ &\ge \min_{C \subseteq \widehat{\tau}_{[1:j-1]}, |C| \le q} \operatorname{var}(X_\ell \mid C) - \Delta/4 \\ &> \frac{3}{4}\Delta + \sigma^2 \\ &> \Delta/4 + \sigma^2 \\ &= \min_{C \subseteq \widehat{\tau}_{[1:j-1]}, |C| \le q} \operatorname{var}(X_k \mid X_C) + \Delta/4 \\ &\ge \min_{C \subseteq \widehat{\tau}_{[1:j-1]}, |C| \le q} v_{kC} = \sigma_k \end{aligned}$$

Thus the first step of Algorithm 1 will always include k instead of ℓ into $\widehat{\tau}$. This implies that $\widehat{\tau}$ is a valid topological ordering.

Now we look at the second step of Algorithm 1, this step is to remove false parents from candidate set returned by Best Subset Selection. For any j, let $\widehat{\tau}_j = j$ for ease of notation. Given that $\widehat{\tau}$ is a valid ordering, $\mathrm{pa}(j) \subseteq \widehat{\tau}_{[1:j-1]}$. We first conclude $\mathrm{pa}(j) \subseteq C_j$, otherwise there exists $C'_j \subseteq \widehat{\tau}_{[1:j-1]}$ with $\mathrm{pa}(j) \subseteq C'_j$ such that

$$\begin{aligned} v_{jC_j} &\geq \mathrm{var}(X_j \,|\, C_j) - \Delta/4 \\ &> \sigma^2 + \Delta - \Delta/4 \\ &> \sigma^2 + \Delta/4 \\ &= \mathrm{var}(X_j \,|\, C'_j) + \Delta/4 \\ &\geq v_{jC'_j} \end{aligned}$$

Then C_j should not lead to minimum. Then for any $k \in \mathrm{pa}(j)$ and any $\ell \in C_j \setminus \mathrm{pa}(j)$,

$$\begin{aligned} v_{jC_j} - v_{jC_j \setminus k} &\geq \mathrm{var}(X_j \,|\, C_j) - \mathrm{var}(X_j \,|\, C_j \setminus k) - \Delta/2 \\ &> \Delta/2 = \gamma \\ v_{jC_j} - v_{jC_j \setminus \ell} &\leq \mathrm{var}(X_j \,|\, C_j) - \mathrm{var}(X_j \,|\, C_j \setminus \ell) + \Delta/2 \\ &= \Delta/2 = \gamma \end{aligned}$$

Thus ℓ will be removed while k will stay. Then $\mathrm{pa}_{\widehat{G}}(j) = \mathrm{pa}_G(j)$ for all j. □

Next we bound the estimation error tail probability:

Lemma A.2. *For all $k \in V$ and $C \subset V \setminus \{k\}$, $|C| \leq q$,*

$$P(|\,\mathrm{var}(X_k \,|\, X_C) - v_{kC}| \geq \epsilon) \leq A_1 \exp(-A_2 n\epsilon/M^5 + q)$$

for some constants A_1, A_2.

Proof. Denote the covariance between k and set of nodes C at

$$\Sigma_t = \begin{pmatrix} \Sigma_{kk} & \Sigma_{kC} \\ \Sigma_{Ck} & \Sigma_{CC} \end{pmatrix}$$

Note that $\|\Sigma_{kC}\|, \|\Sigma_{CC}\|, \|\Sigma_{CC}^{-1}\| \leq \|\Sigma_t\| \leq \|\Sigma\| \leq M$.

Then the estimation error for conditional variance

$$\begin{aligned} &\left|\,\mathrm{var}(X_k \,|\, X_C) - \widehat{\mathrm{var}}(X_k \,|\, X_C)\right| \\ &= \left|(\Sigma_{kk} - \Sigma_{kC}\Sigma_{CC}^{-1}\Sigma_{Ck}) - (\widehat{\Sigma}_{kk} - \widehat{\Sigma}_{kC}\widehat{\Sigma}_{CC}^{-1}\widehat{\Sigma}_{Ck})\right| \\ &\leq \left|\widehat{\Sigma}_{kk} - \Sigma_{kk}\right| + \left|\widehat{\Sigma}_{kC}(\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1})\widehat{\Sigma}_{Ck}\right| \\ &\quad + \left|(\widehat{\Sigma}_{kC} - \Sigma_{kC})\Sigma_{CC}^{-1}\widehat{\Sigma}_{Ck}\right| + \left|\Sigma_{kC}\Sigma_{CC}^{-1}(\widehat{\Sigma}_{Ck} - \Sigma_{Ck})\right| \\ &\leq \left|\widehat{\Sigma}_{kk} - \Sigma_{kk}\right| + \|\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1}\|\|\widehat{\Sigma}_{kC}\|^2 \\ &\quad + M\|\widehat{\Sigma}_{kC} - \Sigma_{kC}\|\|\widehat{\Sigma}_{Ck}\| + M^2\|\widehat{\Sigma}_{Ck} - \Sigma_{Ck}\| \\ &\leq \left|\widehat{\Sigma}_{kk} - \Sigma_{kk}\right| \\ &\quad + 2\|\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1}\|\|\widehat{\Sigma}_{kC} - \Sigma_{kC}\|^2 + 2M^2\|\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1}\| \\ &\quad + M\|\widehat{\Sigma}_{kC} - \Sigma_{kC}\|^2 + M\|\widehat{\Sigma}_{kC} - \Sigma_{kC}\| \\ &+ M^2\|\widehat{\Sigma}_{Ck} - \Sigma_{Ck}\| \\ &\leq \left|\widehat{\Sigma}_{kk} - \Sigma_{kk}\right| \\ &\quad + 2\|\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1}\|\|\widehat{\Sigma}_t - \Sigma_t\|^2 + 2M^2\|\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1}\| \\ &\quad + M\|\widehat{\Sigma}_t - \Sigma_t\|^2 + M\|\widehat{\Sigma}_t - \Sigma_t\| + M^2\|\widehat{\Sigma}_t - \Sigma_t\| \end{aligned}$$

The first inequality is by the triangular inequality, and the second simply bounds $\|\Sigma_{kC}\|$ by M. The third inequality introduces the estimation error of $\widehat{\Sigma}_{kC}$ and the final inequality replaces this with the estimation error of the full covariance matrix $\widehat{\Sigma}_t$. To set the RHS to be smaller than $\epsilon > 0$, we consider three estimation errors. The first two can be controlled via standard sub-exponential concentration, whereas the third can be controlled via Theorem 6.5 from Wainwright (2019):

$$P(|\widehat{\Sigma}_{kk} - \Sigma_{kk}| \geq \zeta) \leq \exp(-A_3 n\zeta/M)$$
$$P(\|\widehat{\Sigma}_t - \Sigma_t\| \geq \zeta) \leq \exp(-A_4 n\zeta/M + q)$$

for some constants A_3, A_4. The largest error is from

$$\begin{aligned}
\|\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1}\| &\leq \|\Sigma_{CC}^{-1}\|\|\widehat{\Sigma}_{CC}^{-1}\|\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\| \\
&\leq M\|\widehat{\Sigma}_{CC}^{-1}\|\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\| \\
&= M(\|\widehat{\Sigma}_{CC}^{-1}\| - \|\Sigma_{CC}^{-1}\|)\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\| \\
&\quad + M\|\Sigma_{CC}^{-1}\|\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\| \\
&\leq M\|\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1}\|\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\| \\
&\quad + M^2\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\|
\end{aligned}$$

After some arrangement, we have

$$\|\widehat{\Sigma}_{CC}^{-1} - \Sigma_{CC}^{-1}\| \leq \frac{M^2\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\|}{1 - M\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\|} \leq \zeta$$

as long as

$$\|\Sigma_{CC} - \widehat{\Sigma}_{CC}\| \leq \frac{\zeta}{M^2 + M\zeta} \leq \frac{\zeta}{M^2}\,.$$

This is just another Gaussian covariance matrix estimation error, i.e.

$$P(\|\widehat{\Sigma}_{CC} - \Sigma_{CC}\| > \zeta/M^2) \leq \exp(-A_5 n\zeta/M^3 + q)\,,$$

for some constant A_5. Now we require all the errors to be bounded by $\zeta = \epsilon/M^2$ such that the conditional variance estimation error is within ϵ. Thus

$$\begin{aligned}
&P(|\operatorname{var}(X_k|X_C) - \widehat{\operatorname{var}}(X_k|X_C)| > \epsilon) \\
&\leq \exp(-A_3 n\epsilon/M^3) + \exp(-A_4 n\epsilon/M^3 + q) \\
&\quad + \exp(-A_5 n\epsilon/M^5 + q) \\
&\leq A_1 \exp(-A_2 n\epsilon/M^5 + q)\,.
\end{aligned}$$

□

A.2 Proof of Theorem 2.1

Proof. Combine Lemma A.1 and A.2, we can have success probability:

$$\begin{aligned}
P(\widehat{G} \neq G) &\leq P\Bigg(\bigcup_{\substack{k\in V \\ C\subset V\setminus\{k\} \\ |C|\leq q}} \{|\operatorname{var}(X_k \,|\, X_C) - v_{kC}| \geq \epsilon\}\Bigg) \\
&\leq \sum_{\substack{k\in V \\ C\subset V\setminus\{k\} \\ |C|\leq q}} P\Big(|\operatorname{var}(X_k \,|\, X_C) - v_{kC}| \geq \epsilon\Big) \\
&\lesssim d \times \left(\binom{d-1}{1} + \ldots + \binom{d-1}{q}\right) \times \exp(q - n\epsilon/M^5) \\
&\lesssim d \times q \times e^{q\log(d/q)} \times \exp(q - n\epsilon/M^5) \\
&= \exp(q\log(d/q) + \log d + \log q + q - n\epsilon/M^5) \\
&\asymp \exp(q\log(d/q) - n\epsilon/M^5)\,.
\end{aligned}$$

The last equality is by $(\frac{d}{q})^q = (\frac{d}{q})^{q-1} \times (\frac{d}{q}) \geq 2^{q-1} \times (\frac{d}{q}) \geq q \times \frac{d}{q} = d$, thus $q \log(d/q) \gtrsim \log d$. In the end, replace ϵ by $\Delta/4$, solve for the sample size n such that

$$\exp(q \log(d/q) - n\Delta/M^5) \lesssim \delta \,,$$

we can have the desired sample complexity. □

A.3 Proof of Theorem 2.2

Proof. In the proof of Lemma A.1, denote the estimation error to be upper bounded by ϵ, i.e. for all $k \in V$ and $C \subset V \setminus \{k\}$, $|C| \leq q$,

$$|\operatorname{var}(X_k \,|\, X_C) - v_{kC}| \leq \epsilon \,,$$

then it suffices to have

$$2\epsilon < \gamma < \Delta/2 - 2\epsilon$$

for the correctness of second phase to proceed. Therefore, let $\gamma = 3\epsilon$ and require $\epsilon < \Delta/10$. Finally, set

$$\epsilon \asymp \frac{2M^5 \times q \log(d/q)}{n}$$

Then we have failure probability bounded:

$$\begin{aligned} P(\widehat{G} \neq G) &\lesssim \exp(q \log(d/q) - n\epsilon/M^5) \\ &= \exp(-q \log(d/q)) \,. \end{aligned}$$

And to satisfy the requirement $\epsilon < \Delta/10$, we need sample size

$$n \gtrsim \frac{M^5 q \log(d/q)}{\Delta} \,.$$ □

B PROOF OF LOWER BOUND

B.1 Preliminaries

Let's start with recalling Fano's inequality and its corollary under the structure learning setting. Let $\theta(F)$ be a parameter associated to some observational distribution F.

Lemma B.1 (Yu, 1997, Lemma 3)**.** *For a class of distributions $\mathcal{F}$ and its subclass $\mathcal{F}' = \{F_1, \dots, F_N\} \subseteq \mathcal{F}$,*

$$\inf_{\widehat{G}} \sup_{F \in \mathcal{F}} \mathbb{E}_F \mathbf{dist}(\theta(F), \widehat{\theta}) \geq \frac{s}{2}\left(1 - \frac{n\alpha + \log 2}{\log N}\right)$$

where

$$\begin{aligned} \alpha &= \max_{F_j \neq F_k \in \mathcal{F}'} \mathbf{KL}(F_j || F_k) \\ s &= \max_{F_j \neq F_k \in \mathcal{F}'} \mathbf{dist}(\theta(F_j), \theta(F_k)) \end{aligned}$$

Set $\theta(F) = G(F)$, $\mathbf{dist}(\cdot, \cdot) = \mathbf{1}\{\cdot \neq \cdot\}$. One consequence of Lemma B.1 is as follows:

Corollary B.2. *Consider some subclass $\mathcal{G}' = (G_1, \dots, G_N) \subseteq \mathcal{G}_{d,q}$, and let $\mathcal{F}' = \{F_1, \dots, F_N\}$, each of whose elements is generated by one distinct $G \in \mathcal{G}'$. If the sample size is bounded as*

$$n \leq \frac{(1 - 2\delta) \log N}{\alpha} \,,$$

then the any estimator for G is δ-unreliable:

$$\inf_{\widehat{G}} \sup_{F \in \mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)} P(\widehat{G} \neq G(F)) \geq \delta - \frac{\log 2}{\log N}$$

Thus the **strategy** for building lower bound is to find a subclass of original problem such that

- Has large cardinality N;
- Pairwise KL divergence between any two distributions is small.

Now we do some counting for the number of DAGs with d nodes and in-degree bounded by q.

Lemma B.3. *For $q \le d/2$, the number of DAGs with d nodes and in-degree bounded by q scales as $\Theta(dq \log d/q)$.*

Proof. The proof construction is similar to Santhanam and Wainwright (2012). We can upper bound by number of directed graphs (DG), and lower bound by one particular subclass of DAGs.

For upper bound, note that a DG has at most d^2 many, and a sparse DAG has at most dq any edges. Since $q \le d/2$, then we have $dq \le d^2/2$. Since for $\ell < k \le d^2/2$, we have $\binom{d^2}{\ell} < \binom{d^2}{k}$, and there are $\binom{d^2}{k}$ DGs with exactly k directed edges, then the numbder of DGs is upper bounded by

$$\log |DGs| = \log \sum_{k=1}^{dq} \binom{d^2}{k} \le \log dq \binom{d^2}{dq} \asymp dq \log \frac{d}{q}$$

For lower bound, we look at one subclass of DAGs. Suppose $d/(q+1)$ is an integer, otherwise discard remaining nodes. First partition d nodes into $q+1$ groups with equal size $d/(q+1)$. Then for the first group, build directed edges from nodes in group $2, 3, \ldots, q+1$ to group 1, which requires q permutations on $d/(q+1)$ nodes within one particular group. Then the nodes in group 1 has exactly degree q. Similarly, for group 2, build directed edges from group $3, 4, \ldots, q+1$, which requires $q-1$ permutations on $d/(q+1)$ nodes. Therefore, for the subclass of DAGs generated in this way of partition, we have

$$(\frac{d}{q+1}!)^{(q)+(q-1)+\ldots+1} = (\frac{d}{q+1}!)^{q(q+1)/2}$$

many DAGs, any of which is valid DAG and has degree bounded by q. Then the cardinality

$$\log |DAGs| \ge \log (\frac{d}{q+1}!)^{q(q+1)/2} \asymp dq \log \frac{d}{q}$$

Thus the total number of DAGs scales as $\Theta(dq \log \frac{d}{q})$. □

B.2 Proof of Theorem 3.1

Proof. We consider two ensembles:

Ensemble A In this ensemble, we consider all possible DAGs with bounded in-degree. Note that by Lemma B.3, we know $N \asymp dq \log d/q$, it remains to provide an upper bound for the KL divergence between any two distributions within the class. For any two $F_j, F_k \in \mathcal{F}_{d,q}(\beta_{\min}, \sigma^2, M)$, denote their covariance matrices to be Σ_j, Σ_k. Due to Gaussianity, It is easy to see that

$$\begin{aligned}
\mathbf{KL}(F_i || F_j) &= \frac{1}{2}\Big(\mathbb{E}_{F_j}[X^\top \Sigma_k^{-1} X] - \mathbb{E}_{F_j}[X^\top \Sigma_j^{-1} X]\Big) \\
&= \frac{1}{2}\Big(\mathbb{E}_{F_j}[X^\top \Sigma_k^{-1} X] - d\Big) \\
&= (\mathrm{tr}(\Sigma_k^{-1}\Sigma_j) - d)/2 \\
&\le (\sqrt{\mathrm{tr}(\Sigma_j^{-2})\,\mathrm{tr}(\Sigma_k^2)} - d)/2 \\
&\le (M^2 - 1)d\,.
\end{aligned}$$

Therefore, we can establish the first lower bound that

$$n \gtrsim \frac{q \log d/q}{M^2 - 1}$$

Ensemble B For this ensemble, we consider the DAGs with exactly one edge $u \to v$ and coefficient $\beta_{\min}$, denoted as G^{uv}. There are 2 directions and $d(d-1)/2$ many edges, so the cardinality of this ensemble would be $N = d(d-1) \asymp d^2$. Then denote the distribution defined according to G^{uv} as F^{uv}, the log likelihood

$$-\log F^{uv} \propto \frac{1}{2\sigma^2}\left[\sum_{i\notin\{u,v\}} X_i^2 + X_u^2 + (X_v - \beta_{\min}X_u)^2\right] + d\log\sigma$$

and the difference between any two cases is

$$\log F^{uv} - \log F^{jk} = \frac{1}{2\sigma^2}\Big[X_v^2 + (X_k - \beta_{\min}X_j)^2 - X_k^2 - (X_v - \beta_{\min}X_u)^2\Big]$$

Then take expectation over F^{uv} we get the KL divergence:

$$\mathbf{KL}(F^{uv}||F^{jk}) = \frac{1}{2\sigma^2}\left[\beta_{\min}^2\sigma^2 + \mathbb{E}_{F^{uv}}(\beta_{\min}^2 X_j^2 - 2X_kX_j)\right]$$

For fixed edge (u, v), any other edges (j, k) has relationship and corresponding KL divergence below:

- $j \neq u, k \neq v$, $\mathbf{KL}(F^{uv}||F^{jk}) = \beta_{\min}^2$
- $j = u, k \neq v$, $\mathbf{KL}(F^{uv}||F^{jk}) = \beta_{\min}^2$
- $j \neq u, k = v$, $\mathbf{KL}(F^{uv}||F^{jk}) = \beta_{\min}^2$
- $j = v, k = u$, $\mathbf{KL}(F^{uv}||F^{jk}) = \beta_{\min}^2 + \beta_{\min}^4/2 - \beta_{\min}$
- $j = v, k \neq u$, $\mathbf{KL}(F^{uv}||F^{jk}) = \beta_{\min}^2 + \beta_{\min}^4/2$
- $j \neq v, k = u$, $\mathbf{KL}(F^{uv}||F^{jk}) = \beta_{\min}^2$

Among them, the largest KL between F^{uv}, F^{jk} is $\beta_{\min}^2 + \beta_{\min}^4/2$. Therefore, we can conclude a lower bound

$$n \gtrsim \frac{1}{\beta_{\min}^4 + 2\beta_{\min}^2}\log d \asymp \frac{\log d}{\beta_{\min}^2}.$$ □

C PROOF OF PROPOSITION 4.2

Proof. For simplicity we consider DAGs with $d+1$ nodes. We first recall a known lower bound for sparsity recovery: Consider the linear model $Y = \beta^\top X + \epsilon$ with $X \sim \mathcal{N}(0, \Sigma)$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The support of β is $S \subset [d]$ and $|S| = q$. Let $\beta_{\min} := \min_{j:\beta_j\neq 0}|\beta_j|$. Then informally,

Lemma C.1 (Wainwright (2009b), Theorem 2)**.** *If*

$$n \leq \frac{\log\binom{d}{q}}{8\omega_{bu}(\Sigma)\frac{\beta_{\min}^2}{\sigma^2}}$$

where

$$\omega_{bu}(\Sigma) := \mathbb{E}_S\left[\min_{|z_j|\geq 1\,\forall j} z_S^\top \Sigma_{SS} z_S\right],$$

then with q known, for any instance from the linear model and any estimator $\widehat{S}$ for S,

$$P(\widehat{S} \neq S) \geq \frac{1}{2}.$$

Now we adapt this result to our setting, $\beta_{\min}$ remains the same, and clearly $\omega_{bu}(\Sigma) \le M$ according the definition of M. Additionally, since every Gaussian with positive definite Σ has a minimal I-map and given ordering, the parents can be read off through regression, i.e. the model class in Lemma C.1 is equivalent to the one generated in $\overline{\mathcal{F}}_{d,q}(\beta_{\min}, \sigma^2_{\max}, M)$. For any estimator $\widehat{G} = \widehat{G}(\tau)$, denote $\widehat{\mathrm{pa}}(k) := \mathrm{pa}_{\widehat{G}}(k)$ for any node k. Then if

$$n \le \frac{\sigma^2_{\max}}{8M\beta^2_{\min}} d \log \frac{d}{q},$$

we have

$$\begin{aligned}
\inf_{F \in \overline{\mathcal{F}}_{d,q}(\beta_{\min}, \sigma^2_{\max}, M)} P(\widehat{G} = G(F) \,|\, \tau) &= \inf_{F \in \overline{\mathcal{F}}_{d,q}(\beta_{\min}, \sigma^2_{\max}, M)} P(\widehat{\mathrm{pa}}(k) = \mathrm{pa}(k)\, \forall k \,|\, \tau) \\
&\le \inf_{F \in \overline{\mathcal{F}}_{d,q}(\beta_{\min}, \sigma^2_{\max}, M)} P(\widehat{\mathrm{pa}}(\tau_d) = \mathrm{pa}(\tau_d) \,|\, \tau) \\
&\le \inf_{\substack{F \in \overline{\mathcal{F}}_{d,q}(\beta_{\min}, \sigma^2_{\max}, M) \\ |\mathrm{pa}(\tau_d)| = q, \mathrm{var}(\epsilon_{\tau_d}) = \sigma^2_{\max}}} P(\widehat{\mathrm{pa}}(\tau_d) = \mathrm{pa}(\tau_d) \,|\, \tau) \\
&\le \inf_{\substack{F \in \overline{\mathcal{F}}_{d,q}(\beta_{\min}, \sigma^2_{\max}, M) \\ |\mathrm{pa}(\tau_d)| = q, \mathrm{var}(\epsilon_{\tau_d}) = \sigma^2_{\max}}} P(\widehat{\mathrm{pa}}(\tau_d) = \mathrm{pa}(\tau_d) \,|\, |\mathrm{pa}(\tau_d)| = q, \tau) \\
&< \frac{1}{2}
\end{aligned}$$

The first inequality is by relaxing the problem to simply finding the parents of the last node from all preceding nodes. The second inequality is because we can restrict at a sub-ensemble of $\overline{\mathcal{F}}_{d,q}(\beta_{\min}, \sigma^2_{\max}, M)$ whose last node of ordering has q parents and maximum noise $\mathrm{var}(\epsilon_{\tau_d}) = \sigma^2_{\max}$. The third inequality is because knowing the number of parents only makes the problem easier. The final inequality is by noticing the equivalence to sparsity recovery problem and applying Lemma C.1. □

D PROOF OF LOWER BOUND OF GGM (THEOREM 5.1)

Proof. We introduce two useful lemmas from Wang et al. (2010):

Lemma D.1 (Wang et al. (2010), Section IV.A)**.** *Consider a restricted ensemble $\widetilde{\mathcal{U}} \subseteq \mathcal{U}$ consisting of $N = |\widetilde{\mathcal{U}}|$ models, and let model index θ be chosen uniformly at random from $\{1, ..., N\}$. Given the observations X, the error probability for any estimator $\widehat{U}$*

$$\max_{U \in \mathcal{U}} P(\widehat{U} \ne U) \ge \max_{j=1,\dots,N} P(\widehat{U} \ne \widetilde{U}_j) \ge 1 - \frac{I(\theta; X) + 1}{\log N}$$

Lemma D.2 (Wang et al. (2010), Section IV.A)**.** *Define the averaged covariance matrix*

$$\bar{\Sigma} := \frac{1}{N} \sum_{j=1}^{N} \Sigma(\widetilde{U}_j)$$

The mutual information is upper bounded by $I(\theta; X) \le \frac{n}{2} R(\widetilde{\mathcal{U}})$, where

$$R(\widetilde{\mathcal{U}}) = \log\det \bar{\Sigma} - \frac{1}{N} \sum_{j=1}^{N} \log\det \Sigma(\widetilde{U}_j)$$

Another lemma for ease of presentation:

Lemma D.3. *For a matrix of dimension p*

$$A = \begin{pmatrix} 1+a & b & \cdots & b \\ b & 1+a & \cdots & b \\ & & \cdots & \\ b & b & \cdots & 1+a \end{pmatrix}$$

with $a, b \to 0$ and $pb \to 0$, the determinant $\log\det A \approx pa$.

Proof.

$$\begin{aligned}
A &= (1+a-b)I_p + b\mathbf{1}_p\mathbf{1}_p^\top \\
&= (1+a-b)\left(I_p + \frac{b}{1+a-b}\mathbf{1}_p\mathbf{1}_p^\top\right) \\
\det A &= (1+a-b)^p \det\left(I_p + \frac{b}{1+a-b}\mathbf{1}_p\mathbf{1}_p^\top\right) \\
&= (1+a-b)^p\left(1+\frac{bp}{1+a-b}\right) \\
&= (1+a-b)^{p-1}(1+a+(p-1)b) \\
\log\det A &= (p-1)\log(1+a-b) + \log(1+a+(p-1)b) \\
&\approx (p-1)(a-b) + a + (p-1)b = pa.
\end{aligned}$$

□

Finally, let's consider three ensembles of UGs generated by DAGs. We describe the ensembles by showing how the DAGs generate the UGs.

Ensemble A In this first Ensemble, we consider an empty DAG, then add one edge from node S to T with linear coefficient $\beta_{\min}$. Specifically,

$$\begin{cases} X_\ell = \beta_{\min} X_S + \epsilon_\ell & \ell = T \\ X_\ell = \epsilon_\ell & \ell \neq T \end{cases}$$

Without loss of generality, let $\mathrm{var}(\epsilon_\ell) = 1$, general variance σ^2 would not affect the final results. There are $N = d(d-1)$ possibilities, thus $\log N \asymp \log d$.

It remains to figure out the structure of covariance matrix and find out the corresponding determinants. Without loss of generality, let the first two nodes to be S, T, then the covariance matrix of any model (denoted as jth) is

$$\Sigma_j = \begin{pmatrix} 1 & \beta_{\min} & \mathbf{0} \\ \beta_{\min} & 1+\beta_{\min}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & I_{d-2} \end{pmatrix}$$

It is easy to see that $\log\det\Sigma_j = \log(1+\beta_{\min}^2 - \beta_{\min}\times\beta_{\min}) = 0$ for all models in this subclass. To compute the average $\bar{\Sigma}$, by symmetry, all diagonal and off-diagonal entries are the same respectively. For entries on diagonal, there are two situations: whether it corresponds to node T or not. For off-diagonal entries, there are two situations: corresponds to edge $S-T$ or not. Different situations behave differently Table 1 with total counts N:

Table 1: Summary of situations of sub-covariance matrix entry in Ensemble A.

	diagonal	
situations	T	Otherwise
values	$\beta_{\min}^2+1$	1
counts	$d-1$	$(d-1)^2$
	off-diagonal	
situations	$T-S$	Otherwise
values	$\beta_{\min}$	0
counts	2	$d(d-1)-2$

Thus we conclude the entries in $\bar{\Sigma}$:

$$\bar{\Sigma}_{ik} = \begin{cases} 1+\frac{\beta_{\min}^2}{d} := 1+a & i=k \\ \frac{2\beta}{d(d-1)} := b & i\neq k \end{cases}$$

Using Lemma D.3, we conclude $\log\det\bar{\Sigma} \asymp \beta_{\min}^2$, and invoking Lemma D.1 and D.2, we obtain a lower bound as

$$n \gtrsim \frac{\log d}{\beta_{\min}^2}.$$

Ensemble B Here we can adopt the same construction as the first ensemble for DAG in Appendix A, which applies analogously through Lemma B.1. Since the joint distribution remains to be the same, we have KL divergence upper bounded by $(M^2-1)d$.

For number of models inside this class, firstly we know that for a UG with degree bounded by s, there are $\Theta(ds\log d/s)$ many UGs (Lemma 1(b) of Santhanam and Wainwright (2012)). By Lemma 5.2, $U=\mathcal{M}(G)$, so $q\le s$, thus the number of UGs would be greater than $\Theta(dq\log d/q)$, which leads to the same lower bound:

$$n \gtrsim \frac{q\log d/q}{M^2-1}\,. \qquad \square$$

E PROOF OF LEMMA 1.1

Proof. Immediate from the law of total variance:

$$\begin{aligned}
\Delta &\equiv \min_k \min_{\substack{A\subseteq \mathrm{nd}(k)\\ \mathrm{pa}(k)\setminus A\neq\emptyset\\ A\subseteq \mathrm{nd}(\mathrm{pa}(k)\setminus A)}} \mathbb{E}_A \operatorname{var}(X_k \mid A) - \sigma^2 \\
&= \min_k \min_{\substack{A\subseteq \mathrm{nd}(k)\\ \mathrm{pa}(k)\setminus A\neq\emptyset\\ A\subseteq \mathrm{nd}(\mathrm{pa}(k)\setminus A)}} \mathbb{E}_A \mathbb{E}_{\mathrm{pa}(k)\setminus A} \operatorname{var}(X_k \mid \mathrm{pa}(k)) \\
&\quad + \mathbb{E}_A \operatorname{var}_{\mathrm{pa}(k)\setminus A} \mathbb{E}(X_k \mid \mathrm{pa}(k)) - \sigma^2 \\
&= \min_k \min_{\substack{A\subseteq \mathrm{nd}(k)\\ \mathrm{pa}(k)\setminus A\neq\emptyset\\ A\subseteq \mathrm{nd}(\mathrm{pa}(k)\setminus A)}} \mathbb{E}_A \operatorname{var}_{\mathrm{pa}(k)\setminus A} \left[\beta_{\mathrm{pa}(k)\setminus A}^\top X_{\mathrm{pa}(k)\setminus A} \mid A\right] \\
&= \beta_{\min}^2 \sigma^2.
\end{aligned}$$

$\square$