Information-theoretic limits of Bayesian network structure learning

Asish Ghoshal
aghoshal@purdue.edu

Jean Honorio
jhonorio@purdue.edu

Department of Computer Science, Purdue University, West Lafayette, IN, U.S.A. - 47907.

Abstract

In this paper, we study the information-theoretic limits of learning the structure of Bayesian networks (BNs), on discrete as well as continuous random variables, from a finite number of samples. We show that the minimum number of samples required by *any procedure* to recover the correct structure grows as $\Omega(m)$ and $\Omega\left(k \log m + {}^{k^2}/_{m}\right)$ for non-sparse and sparse BNs respectively, where m is the number of variables and k is the maximum number of parents per node. We provide a simple recipe, based on an extension of the Fano's inequality, to obtain information-theoretic limits of structure recovery for any exponential family BN. We instantiate our result for specific conditional distributions in the exponential family to characterize the fundamental limits of learning various commonly used BNs, such as conditional probability table based networks, Gaussian BNs, noisy-OR networks, and logistic regression networks. En route to obtaining our main results, we obtain tight bounds on the number of sparse and non-sparse essential-DAGs. Finally, as a byproduct, we recover the information-theoretic limits of sparse variable selection for logistic regression.

1 Introduction

Motivation. Bayesian Networks (BNs) are a class of probabilistic graphical models that describe the conditional dependencies between a set of random variables as a directed acyclic graph (DAG). However, in many problems of practical interest, the structure of the network is not known a priori and must be inferred from data.

Although, many algorithms have been developed over the years for learning BNs (cf. [1] and [2] for a detailed survey of algorithms), an important question that has hitherto remained unanswered is the fundamental limits of learning BNs, i.e., "What is the minimum number of samples required by any procedure to recover the true DAG structure of a BN?". The answer to this question would help shed light on the fundamental limits of learning the DAG structure of BNs, and also help determine if existing algorithms are optimal in terms of their sample complexity or if there exists a gap between the state-of-the-art estimation procedures and the information-theoretic limits. In this paper we obtain lower bounds on the minimum number of samples required to learn BNs over m variables, and sparse BNs over m variables with maximum in-degree of k.

Contribution. In this paper, we make the following contributions. We derive necessary conditions on the sample complexity of recovering the DAG structure of non-sparse and sparse BNs. We show that $\Omega(m)$ samples are necessary for consistent recovery of the DAG structure of BNs, while for sparse networks $\Omega\left(k \log m + {}^{k^2}/_{m}\right)$ samples are necessary. We provide a simple recipe for obtaining the information-theoretic limits of learning any exponential family BN, and we instantiate our result for specific conditional distributions to determine the fundamental limits of learning the structure of various widely used BNs, namely, conditional probability table (CPT) based networks, Gaussian networks, noisy-OR networks, and logistic regression networks. Our lower bound of $\Omega\left(k^2 \log m\right)$ matches the upper bound on $\mathcal{O}\left(k^2 \log m\right)$, obtained by Ravikumar et al. [3] for ℓ_1-regularized logistic regression. We also show that the *SparsityBoost* algorithm developed by Brenner and Sontag [4] for learning binary CPT BNs, which has a sample complexity of $\mathcal{O}\left(m^2 ({}^{1}/_{\theta_{\min}})\right)$, is far from the information-theoretic limit of $\Omega\left(\frac{k \log m + {}^{k^2}/_{m}}{\log({}^{1}/_{\theta_{\min}})}\right)$, where $\theta_{\min}$, is the minimum

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

probability value in the conditional probability tables. An interesting corollary of our main result is that learning *layered BNs* — where the ordering of the nodes is known (upto layers) and the parent set of each variable is constrained to be in the layer above it — is as hard as learning general BNs in terms of their sample complexity. Lastly, of independent interest, are our extension of the Fano's inequality to the case where there are latent variables, and our upper-bound on the KL divergence between two exponential family distributions as the inner product of the difference between the natural parameter and the expected sufficient statistics.

2 Related Work

Höffgen [5], and Friedman and Yakhini [6] were among the first to derive sample complexity results for learning BNs. In both [5] and [6] the authors provide upper bounds on sample complexity of learning a BN that is likelihood consistent, i.e., the likelihood of the learned network is ϵ away from the likelihood of the true network in terms of the Kullback-Leibler (KL) divergence measure. Abbeel et al. [7] provide polynomial sample complexity results for learning likelihood consistent factor graphs.

Among sample complexity results for learning structure consistent BNs, where the structure of the learned network is close to the true network, Spirtes et al. [8] and Cheng et al. [9] provide such guarantees for polynomial-time test-based methods, but the results hold only in the infinite-sample limit. Chickering and Meek in [10] also provide a greedy hill-climbing algorithm for structure learning that is structure consistent in the infinite sample limit. Zuk et al. [11] show structure consistency of a single network and do not provide uniform consistency for all candidate networks, i.e., the bounds relate to the error of learning a specific wrong network having a score greater than the true network. Brenner and Sontag [4] provide upper bounds on the sample complexity of recovering the structure of sparse BNs. However, they consider binary valued variables only and the sample complexity grows as $\mathcal{O}\left(m^2\right)$.

Fano's method has also been used to obtain lower bounds on the sample complexity of undirected graphical model (Markov random fields or MRFs) selection. See Appendix A for results for MRFs and technical differences between learning BNs and MRFs.

3 Preliminaries

Let $\mathsf{X} = \{X_1, \ldots, X_m\}$ be a set of random variables, where $X_i \in \mathcal{X}_i, \forall i \in [m]$. Let $\mathcal{D} \stackrel{\text{def}}{=} \times_{i=1}^{m} \mathcal{X}_i$ be the domain in which the variables in X jointly take their values. A BN for X is a tuple $(G, \mathcal{P}(G, \Theta))$; where $G = (\mathsf{V}, \mathsf{E})$ is a directed acyclic graph (DAG) with $\mathsf{V} = [m]$ being the vertex set and $\mathsf{E} \subset [m] \times [m]$ being the set of directed edges, and $\mathcal{P}(G, \Theta)$ is a probability distribution over X that is parameterized by Θ and factorizes according to the DAG structure G. Particularly, $\forall \mathbf{x} \in \mathcal{D}$, $\mathcal{P}(\mathbf{x}; G, \Theta) = \prod_{i=1}^{m} \mathcal{P}_i(x_i; \pi_i(G), \Theta)$, where $\pi_i(G) \subseteq [m] \backslash \{i\}$ is the parent set of the i-th node in G, $\mathsf{X}_{\pi_i(G)} = \{X_j | j \in \pi_i(G)\}$, $\mathcal{P}_i(.) = \Pr\left\{x_i | \mathsf{X}_{\pi_i(G)}, \Theta_i\right\}$ is the conditional distribution of X_i given an assignment to its parent set, and Θ_i are the parameters for the i-th conditional distribution.

The DAG structure G of a BN specifies the conditional independence relationships that exist between different random variables in the set X. Different graph structures which make the same conditional independence assertions about a set of random variables are called Markov equivalent.

Definition 1 (Markov equivalence)**.** *Two DAGs* $G_1 = (\mathsf{V}, \mathsf{E}_1)$ *and* $G_2 = (\mathsf{V}, \mathsf{E}_2)$ *are Markov equivalent if for all disjoint subsets* $\mathsf{A}, \mathsf{B}, \mathsf{C} \subset \mathsf{V}$, $\mathsf{X}_\mathsf{A} \perp\!\!\!\perp \mathsf{X}_\mathsf{B} | \mathsf{X}_\mathsf{C}$ *in* $G_1 \iff \mathsf{X}_\mathsf{A} \perp\!\!\!\perp \mathsf{X}_\mathsf{B} | \mathsf{X}_\mathsf{C}$ *in* G_2.

The set of DAGs that are Markov equivalent to the DAG G is denoted by $[G]$.

An essential graph[1], consisting of both directed edges, which are called protected edges, and undirected edges, is a canonical representation of the (Markov) equivalence class of DAGs. The undirected edges can be oriented in either direction without changing the conditional independence relationships encoded by the graphs. We denote by G^* the essential graph for $[G]$.

4 Problem Formulation

Let $\mathcal{G}$ be an ensemble of DAGs. We denote by $\mathcal{G}_m$ the ensemble of DAGs over m nodes. Also, let $\boldsymbol{\varphi}(\mathcal{G})$ be some set of "parameter maps". A parameter map $\Theta \in \boldsymbol{\varphi}(\mathcal{G})$, maps a given DAG structure G to a specific instance of the conditional distribution parameters that are compatible with the DAG structure G, i.e., $\Theta(G)$. It is useful to think of Θ as a policy for setting the parameters of the conditional distributions, given a DAG G. For instance, for binary CPT networks, a particular policy Θ would consist of several candidate probability tables for each node, one for each possible number of parents the node can have (from 0 to $m-1$), with entries set to some specific values. Then, given a DAG structure G, $\Theta(G)$ assigns a probability table to each node (from the policy Θ) according to the number of parents of the node in G. This notion of parameter maps affords us the ability

[1] See Andersson et. al. [12] for a formal definition.

to generate a BN by sampling the DAG structure and the parameters independently of each other, which, as would be evident later, is a key technical simplification. Let $\mathcal{P}_{\mathcal{G}}$ and $\mathcal{P}_{\boldsymbol{\varphi}(\mathcal{G})}$ be probability measures on the set $\mathcal{G}$ and $\boldsymbol{\varphi}(\mathcal{G})$ respectively. Nature picks a graph structure G, according to $\mathcal{P}_{\mathcal{G}}$, and then samples a parameter map Θ, independently, according to $\mathcal{P}_{\boldsymbol{\varphi}(\mathcal{G})}$. Thereafter, nature generates a data set of n i.i.d observations, $\mathsf{S} = \{\mathbf{x}^{(i)}\}_{i=1}^{n}$, with $\mathbf{x}^{(i)} \in \mathcal{D}$, from the BN $(G, \mathcal{P}(G, \Theta(G)))$. The problem of structure learning in BNs concerns with estimating the graph structure $\widehat{G}$, up to Markov equivalence, from the data set S. In that context, we define the notion of a decoder. A decoder is any function $\zeta : \mathcal{D}^n \to \mathcal{G}$ that maps a data set of n observations to an estimated DAG $\widehat{G} \in \mathcal{G}$. The estimation error is defined as follows

$$p_{\text{err}} \overset{\text{def}}{=} \inf_{\zeta} \sup_{\mathcal{P}_{\boldsymbol{\varphi}(\mathcal{G})}} \sum_{G} \int_{\Theta \in \boldsymbol{\varphi}(\mathcal{G})} \Big(\mathcal{P}_{\boldsymbol{\varphi}(\mathcal{G})}(\Theta) \mathcal{P}_{\mathcal{G}}(G) \times \Pr\{\zeta(\mathsf{S}) \notin [G] | G, \Theta\} \Big), \tag{1}$$

where the probability $\Pr\{.|G, \Theta\}$ is computed over the data distribution $\mathcal{P}(G, \mathcal{P}(G, \Theta(G)))$ for a specific DAG structure G and parameters $\Theta(G)$.

Note that our definition of estimation error is stronger than what is typically used in the literature for structure recovery of MRFs (see e.g. [13] and [14]), since we focus on the maximum error across all measures $\mathcal{P}_{\boldsymbol{\varphi}(\mathcal{G})}$ over the parameter maps $\boldsymbol{\varphi}(\mathcal{G})$ which itself can be uncountable. Here, we are interested in obtaining necessary conditions for consistent structure recovery of BNs, i.e., we show that if the number of samples is less than a certain threshold, then any decoder ζ fails to recover the true graph structure with probability of error $p_{\text{err}} > 1/2$.

We emphasize that while our sample complexity results invariably depend on the parameter space $\boldsymbol{\varphi}(\mathcal{G})$ under consideration, the decoder only has access to the data set S. Apart from $\mathcal{G}_m$, we consider various other ensembles of DAGs in this paper, to fully characterize the fundamental complexity of learning different classes of BNs. Among the ensembles we consider, $\mathcal{G}_{m,k}$ denotes the family of DAGs, where each node is allowed to have at most k parents. We also consider generalizations of QMR-DT [15] type two-layered BNs, to multiple layers of nodes, with nodes in each layer only allowed to have parents in the layer above it.

Let $\mathcal{V} = \{\mathsf{V}_i\}_{i=1}^{l}$, define an ordering of m nodes into l layers where V_i is the set of nodes in the i-th layer. We have that $|\mathsf{V}_i| = m_i$ and $\sum_{i=1}^{l} |\mathsf{V}_i| = m$. $\mathcal{G}_m^l(\mathcal{V})$ denotes an ensemble of DAG structures where $\forall G = (\mathsf{V}, \mathsf{E}) \in \mathcal{G}_m^l$, $\mathsf{V} = \bigcup_{\mathsf{V}_i \in \mathcal{V}} \mathsf{V}_i$ and $\mathsf{E} = \{(u, v) | u \in \mathsf{V}_{i+1} \wedge v \in \mathsf{V}_i,\ i \in [l-1]\}$. We write $\mathcal{G}_m^l$ instead of $\mathcal{G}_m^l(\mathcal{V})$ to indicate that the members of $\mathcal{G}_m^l$ have some *known* layer-wise ordering of the m nodes, without making the ordering explicit. Finally, we consider another ensemble $\mathcal{G}_{m,k}^l \subset \mathcal{G}_m^l$ where the nodes are allowed to have at most k parents. Together, the ensembles $\mathcal{G}_m, \mathcal{G}_{m,k}, \mathcal{G}_m^l$ and $\mathcal{G}_{m,k}^l$, span a wide range of the sample complexity landscape of recovering the structure of BNs. In the following section we present our main result on the fundamental limits of learning BNs.

5 Main Results

Fano's inequality is one of the primary tools used for deriving necessary conditions on structure recovery of graphical models. The difficulty of recovering the DAG structure of a BN, however, depends both on the structural properties of the ensemble of DAG structures under consideration, as well as on the conditional distributions and their parameters. In order to obtain guarantees about structure recovery, we treat the parameters of the conditional distributions as latent variables — variables that we do not observe and are not interested in estimating. Given that the likelihood of the observed data depends, both on the structure and parameters of the BN that generated the data, it behooves us to ask: "*If we are only interested in recovering the structure of BNs, do the presence of unobserved parameters make structure estimation easier or harder?*" To rigorously answer this question, we extend the classic Fano's inequality, which is defined for a Markov chain, to a slightly more general setting as given below.

(a) (b)

Figure 1: Fano's inequality extension. In (a) the edge between W and X is undirected to indicate that the edge can be oriented in either direction.

Theorem 1 (Fano's inequality extension)**.** *Let W, X, and Y be random variables and let $\widehat{X}$ be any estimator of X. If the random variables are related according to the graphical model in Figure 1 (a), then*

$$\Pr\left\{X \neq \widehat{X}\right\} \geq 1 - \frac{I(Y; X|W) + \log 2}{H(X|W)}. \tag{2}$$

Moreover, if $W \in \mathcal{W}$, is independent of $X \in \mathcal{X}$ (Figure 1 (b)), and $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{P}_{\mathcal{W}}$ be any probability measures over $\mathcal{X}$ and $\mathcal{W}$ respectively, then,

$$\sup_{\mathcal{P}_{\mathcal{W}}} \sum_{x \in \mathcal{X}} \int_{w \in \mathcal{W}} \Pr\left\{x \neq \widehat{X} | X = x, W = w\right\} \mathcal{P}_{\mathcal{W}}(w) \mathcal{P}_{\mathcal{X}}(x)$$

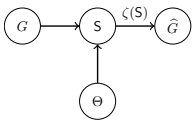

Figure 2: The DAG structure G and a parameter map Θ are sampled independently. Data set S of m samples is generated from the BN $(G, \mathcal{P}(G, \Theta(G))$. A decoder ζ then estimates the DAG structure $\widehat{G} = \zeta(\mathsf{S})$.

$$\geq 1 - \frac{\sup_{w \in \mathcal{W}} I(Y; X|W = w) + \log 2}{H(X)}. \tag{3}$$

Proofs of main results can be found in Appendix B.

Remark 1. *Theorem 1 can be seen an extension of Fano's inequality to the case where there are latent variables W that influence Y, while we are interested in only estimating X. If in Figure 1 we have that $W \to X$, then $I(X;Y|W) \geq I(X;Y)$. Further, since $H(X|W) \leq H(X)$, we conclude that presence of the latent variable W, reduces the estimation error p_{err}.*

Remark 2. *When X and W are independent, we get $I(X;Y|W) = I(X;Y)$, and (2) reduces to the well known Fano's inequality. However, the conditional mutual information $I(X;Y|W = w)$ in (3) can be computed easily as compared to $I(X;Y)$ when we have access to the conditional distribution of Y given X and W. Also note that we do not need $\mathcal{W}$ to be countable.*

Theorem 1 serves as our main tool for lower bounding the estimation error, and subsequently obtaining necessary conditions on the number of samples. In order to obtain sharp lower bounds on p_{err}, we assume that DAG structures and parameters maps (and by extension parameters) are sampled independently. Figure 2 shows the schematics of the inference procedure.

Henceforth, we will use the terms "parameter maps" and "parameters" interchangeably, since given a DAG structure G, the parameter map Θ, maps G to a specific parameterization of the BN. Then, given any ensemble of DAG structures $\mathcal{G}$, the main steps involved in using Theorem 1 to lower bound p_{err} are: (a) obtaining lower bounds on $H(G) = \log|\mathcal{G}|$, which follows from our assumption that G is sampled uniformly from $\mathcal{G}$, and (b) computing the mutual information between the data set and DAG structures over all possible parameter choices, i.e., $\sup_{\Theta \in \boldsymbol{\varphi}(\mathcal{G})} I(\mathsf{S}; G|\Theta)$. To accomplish each of the above objectives, we consider restricted ensembles $\widetilde{\mathcal{G}}_m \subset \mathcal{G}_m$ and $\widetilde{\mathcal{G}}_{m,k} \subset \mathcal{G}_{m,k}$, of size-one equivalence classes, i.e., $\forall G \in \widetilde{\mathcal{G}}_m \vee G \in \widetilde{\mathcal{G}}_{m,k}$, $|[G]| = 1$. Note that for any graph G in $\mathcal{G}_m^l$ or $\mathcal{G}_{m,k}^l$, we have that $|[G]| = 1$, since edges are constrained to go from layer $(i+1)$ to i. Thus, the ensembles $\widetilde{\mathcal{G}}_m, \widetilde{\mathcal{G}}_{m,k}, \mathcal{G}_m^l$ and $\mathcal{G}_{m,k}^l$ can be thought of as consisting of essential DAGs, where all edges are protected. In the following section, we bound the number of essential DAGs in each of the restricted ensembles.

Enumerating DAGs. Essential DAGs, i.e., Markov equivalent classes of DAGs of size 1, was first enumerated by Steinsky [16]. However, the number of DAGs is given as a recurrence relation, for which a closed form solution is difficult to compute. Therefore, we compute tight bounds on the number of essential DAGs in the following paragraphs. In the following lemmas we bound the number of DAGs in each of the restricted ensembles introduced previously.

Lemma 1. *The size of the restricted ensemble $\widetilde{\mathcal{G}}_m$ is bounded as follows:*

$$2^{(m(m-3)/2)+1} \leq \left|\widetilde{\mathcal{G}}_m\right| \leq m!\, 2^{m(m-1)/2}, \tag{4}$$

and $\log\left|\widetilde{\mathcal{G}}_m\right| \geq ((m(m-3)/2) + 1)\log 2$.

Note that the lower bound in Lemma 1 is asymptotically tight. Now we bound the number of essential DAGs where each node is allowed to have at most k parents.

Lemma 2. *Assuming $k > 1$ and $m > 2$, the size of the restricted ensemble $\widetilde{\mathcal{G}}_{m,k}$ is bounded as follows:*

$$2^{(k(k-3)/2)+1} \prod_{j=k+1}^{m-1} \left(\sum_{i=0}^{k} \binom{j-1}{i}\right) \leq \left|\widetilde{\mathcal{G}}_{m,k}\right| \leq m!\, 2^{k(k-1)/2} \prod_{j=k+1}^{m-1} \left(\sum_{i=0}^{k} \binom{j}{i}\right), \tag{5}$$

and $\log\left|\widetilde{\mathcal{G}}_{m,k}\right| \geq k\{\log(m-2)! - (m-k-2)\log k - \log k!\} + \{(k(k-3)/2) + 1\}\log 2$.

Note that using Stirling's factorial formula, the above lemma gives the following lower bound on the number of sparse essential DAGs: $\log\left|\widetilde{\mathcal{G}}_{m,k}\right| = \Omega(km\log m)$. Further, a little calculation shows that $\log\left|\widetilde{\mathcal{G}}_{m,k}\right| = \mathcal{O}(km\log m)$ for large enough m. Thus our bounds for the number of sparse essential DAGs is tight. The following lemma bounds the number of "layered" essential DAGs.

Lemma 3. *The number of BNs in the family $\mathcal{G}_m^l$ and $\mathcal{G}_{m,k}^l$ is as follows:*

$$\left|\mathcal{G}_m^l\right| = \prod_{i=1}^{l-1} (2^{m_{i+1}})^{m_i},\ \left|\mathcal{G}_{m,k}^l\right| = \prod_{i=1}^{l-1} \left[\sum_{j=0}^{k} \binom{m_{i+1}}{j}\right]^{m_i}. \tag{6}$$

Further, $\log\left|\mathcal{G}_m^l\right|$ *and* $\log\left|\mathcal{G}_{m,k}^l\right|$ *are given as follows:*

$$\log\left|\mathcal{G}_m^l\right| = (\log 2)\sum_{i=1}^{l-1}(m_{i+1})(m_i),$$

$$\log \left|\mathcal{G}^l_{m,k}\right| \geq k \sum_{l=1}^{l-1} m_i \log\left(\frac{m_{i+1}}{k}\right).$$

Next, we compute bounds on the mutual information between the data set and DAG structures.

Mutual Information Bounds. The mutual information $I(\mathsf{S}; G|\Theta)$ cannot be computed exactly, in general. Therefore, we use the following lemma to bound the mutual information from above.

Lemma 4. *Let $\mathcal{P}_{\mathsf{S}|G,\Theta}$ be the distribution of S conditioned on a specific DAG G and specific parameters Θ, and let $\mathcal{Q}$ be any distribution over S. Then we have*

$$\sup_{\Theta \in \boldsymbol{\varphi}(\mathcal{G})} I(\mathsf{S}; G|\Theta) \leq \sup_{\Theta \in \boldsymbol{\varphi}(\mathcal{G})} \frac{1}{|\mathcal{G}|} \sum_{G \in \mathcal{G}} \mathbb{KL}\left(\mathcal{P}_{\mathsf{S}|G,\Theta} \middle\| \mathcal{Q}\right). \tag{7}$$

Assuming that $\mathcal{X}_i = \mathcal{X}, \forall i \in [m]$, we chose the distribution $\mathcal{Q}$ to be the product distribution $\mathcal{Q} = \mathcal{Q}_0^{mn}$, where $\mathcal{Q}_0 = \mathcal{P}_i(\varnothing, \Theta_i)$. In other words, the distribution $\mathcal{Q}$ is chosen to be the distribution encoded by a DAG with no edges.

The main hurdle in using (7) to bound the mutual information $I(\mathsf{S}, G|\Theta)$, is computing the KL divergence $\mathbb{KL}\left(\mathcal{P}_{\mathsf{S}|G,\Theta} \middle\| \mathcal{Q}\right)$. *Often times, in BNs characterized by local conditional distributions, coming up with a closed form solution for the joint distribution over all nodes or even the marginal distribution of an arbitrary node is not possible*; unless we assume that the marginal distribution of the parents of a node form a *conjugate prior* for the conditional distribution of the node — an assumption which is quite restrictive. Therefore, to tackle the above problem, we derive the following upper bound on the KL divergence for exponential family distributions which is easy to compute.

Lemma 5 (KL Divergence Bound for Exponential Family Distributions)**.** *Let $X \in \mathbb{R}^d$ be any random variable. Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be distributions over X, belonging to the exponential family, with natural parameters $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ respectively, i.e.*

$$\mathcal{P}_1(x) = \exp(\boldsymbol{\eta}_1^T \mathbf{T}(x) - \psi(\boldsymbol{\eta}_1))h(x),$$

where $\mathbf{T}(X)$ is the sufficient statistics (similarly for $\mathcal{P}_2$). Assuming $h(x) \neq 0\ \forall x \in \mathbb{R}^d$, we have

$$\mathbb{KL}\left(\mathcal{P}_1 \| \mathcal{P}_2\right) \leq \Delta(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2), \tag{8}$$

$$\Delta(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) \stackrel{\text{def}}{=} (\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2)^T (\mathcal{T}(\boldsymbol{\eta}_1) - \mathcal{T}(\boldsymbol{\eta}_2)), \tag{9}$$

where $\mathcal{T}(\boldsymbol{\eta}_1) \stackrel{\text{def}}{=} \mathbb{E}_X\left[\mathbf{T}(x)|\boldsymbol{\eta}_1\right]$ is the expected sufficient statistic of X as computed by the distribution parameterized by $\boldsymbol{\eta}_1$ (similarly for $\mathcal{T}(\boldsymbol{\eta}_2)$).

Note that even though $\mathbb{KL}\left(\mathcal{P}_1 \| \mathcal{P}_2\right)$ is not symmetric, its upper bound $\Delta(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ is symmetric. Given the fact that S is sampled i.i.d from $\mathcal{P}(G, \Theta(G))$, which in turn factorizes as a product of conditional distributions $\mathcal{P}_i$, we then have the following result for the mutual information $I(\mathsf{S}; G|\Theta)$.

Lemma 6 (Mutual Information Bound)**.** *For any ensemble of DAG structures $\mathcal{G}$, we have*

$$\sup_{\Theta \in \boldsymbol{\varphi}(\mathcal{G})} I(\mathsf{S}; G|\Theta) \leq \frac{n}{|\mathcal{G}|} \sum_{G \in \mathcal{G}} \sum_{i=1}^{m} \sup_{\Theta \in \boldsymbol{\varphi}(\mathcal{G})} \mathbb{E}_{\mathsf{X}_{\pi_i}}\left[\mathbb{KL}\left(\mathcal{P}_i(\pi_i(G), \Theta) \| \mathcal{Q}_0\right)\right], \tag{10}$$

where $\mathcal{P}_i(\pi_i(G), \Theta)$ is the conditional distribution of the i-th node. Further, if we have that, $\forall i \in [m], X_i \in \mathcal{X}$ and $\mathcal{P}_i(\pi_i(G), \Theta_i)$ belongs to the exponential family with natural parameter $\boldsymbol{\eta}_i \stackrel{\text{def}}{=} \boldsymbol{\eta}(\mathsf{X}_{\pi_i}, \Theta_i)$ and $\mathcal{Q}_0$ belongs to the exponential family with natural parameter $\boldsymbol{\eta}_0$; then,

$$\sup_{\Theta \in \boldsymbol{\varphi}(\mathcal{G})} I(\mathsf{S}; G|\Theta) \leq \frac{n}{|\mathcal{G}|} \sum_{G \in \mathcal{G}} \sum_{i=1}^{m} \sup_{\Theta \in \boldsymbol{\varphi}(\mathcal{G})} \mathbb{E}_{\mathsf{X}_{\pi_i}}\left[\Delta(\boldsymbol{\eta}_i, \boldsymbol{\eta}_0)\right].$$

Remark 3. *In the above lemma, $\Delta(\boldsymbol{\eta}_i, \boldsymbol{\eta}_0)$ is a random variable because the natural parameter $\boldsymbol{\eta}_i$ depends on the parents X_{π_i}. The quantity $\Delta(\boldsymbol{\eta}_i, \boldsymbol{\eta}_0)$ in the above lemma is non-negative and measures how far the conditional distribution of a variable with parents $\pi_i(G)$ is from the distribution of the variable with no parents, as a function of the difference between the expected sufficient statistics and the natural parameters. The mutual information between the data set S and the DAG structure G is then a sum of the expected "distances" of the conditional distributions from the distribution of a variable with no parents.*

With the exception of Gaussian BNs, where we can write the joint and marginal distributions of the variables in closed form, it is in general difficult to compute the expectation of $\Delta(\boldsymbol{\eta}_i, \boldsymbol{\eta}_0)$. Therefore, we bound the mutual information by bounding $\Delta(\boldsymbol{\eta}_i, \boldsymbol{\eta}_0)$, which can be easily done for bounded random variables. From the above lemma, we then get the following mutual information bound for layered BNs.

Corollary 1 (Mutual Information Bound for Layered BNs)**.** *If $\mathcal{G} = \mathcal{G}^l_m(\mathcal{V}) \vee \mathcal{G} = \mathcal{G}^l_{m,k}(\mathcal{V})$, then*

$$\sup_{\Theta \in \boldsymbol{\varphi}(\mathcal{G})} I(\mathsf{S}; G|\Theta) \leq \frac{(m - m_l)n}{|\mathcal{G}|} \sum_{G \in \mathcal{G}} \Big\{ \max_{i \in \mathsf{V} \setminus \mathsf{V}_l} \sup_{\Theta \in \boldsymbol{\varphi}(\mathcal{G})} \mathbb{E}_{\mathsf{X}_{\pi_i}}\left[\mathbb{KL}\left(\mathcal{P}_i(\pi_i(G), \Theta) \| \mathcal{Q}_0\right)\right] \Big\},$$

where we recall that $\mathcal{V} = \{\mathsf{V}_j\}_{j=1}^l$ is an ordering of nodes into l layers, V_j is the set of nodes in the j-th

layer and $m_j = |\mathsf{V}_j|$. Further, for exponential family conditional distributions, we have

$$\sup_{\Theta\in\boldsymbol{\varphi}(\mathcal{G})} I(\mathsf{S};G|\Theta) \leq \frac{(m-m_l)n}{|\mathcal{G}|}\sum_{G\in\mathcal{G}}\Big\{\max_{i\in\mathsf{V}\setminus\mathsf{V}_l}\sup_{\Theta\in\boldsymbol{\varphi}(\mathcal{G})} \mathbb{E}_{\mathsf{X}_{\pi_i}}\left[\Delta(\boldsymbol{\eta}_i,\boldsymbol{\eta}_0)\right]\Big\}.$$

In order to obtain tight sample complexity results, we need to create "difficult instances" of BNs that are hard to learn. Intuitively speaking, inferring the parents of a node will be hard if the conditional distribution of a node with many parents is close to that of a node with no parents. Therefore, we make the following crucial assumption about the conditional distributions specified by the BN $(G,\mathcal{P}(G,\Theta(G)))$.

Assumption 1. *The KL divergence between the conditional distributions $\mathcal{P}_i(\pi_i(G),\Theta)$ and $\mathcal{Q}_0 \stackrel{\text{def}}{=} \mathcal{P}_i(\varnothing,\Theta)$ over the variable X_i is bounded by a constant, which for exponential family distribution translates to:*

$$\sup_{\Theta\in\boldsymbol{\varphi}(\mathcal{G})} \mathbb{E}_{\mathsf{X}_{\pi_i}}\left[\Delta(\boldsymbol{\eta}_i,\boldsymbol{\eta}_0)\right] \leq \Delta_{\max}, \forall i\in[m],$$

where $\Delta_{\max} > 0$ is a constant.

We show that under certain parameterizations, Assumption 1 holds for many commonly used BNs under very mild restrictions on the parameter space $\boldsymbol{\varphi}(\mathcal{G})$.

Theorem 2. *Let $\mathsf{X} = \{X_1,\ldots,X_m\}$ be a set of random variables with $X_i \in \mathcal{X}$. Let G be a DAG structure over X drawn uniformly at random from some family $\mathcal{G}$ of DAG structures. Further assume that we are given a data set S of n i.i.d samples drawn from a BN $(G,\mathcal{P}(G,\Theta(G)))$, where the parameter map Θ, is drawn from some family $\boldsymbol{\varphi}(\mathcal{G})$. Assuming that the condition in Assumption 1 holds for the conditional distributions $\mathcal{P}_i$, then If $n \leq \mathcal{L}(\mathcal{G})$, than any decoder ζ fails to recover the true DAG structure G with probability $p_{\text{err}} \geq 1/2$, where p_{err} is defined according to (1). The necessary number of samples $\mathcal{L}(\mathcal{G})$, for various classes of BNs, is given as follows:*

$$\mathcal{L}(\widetilde{\mathcal{G}}_m) = \frac{\log 2}{\Delta_{\max}}\left(\frac{m-3}{2} - \frac{1}{m}\right),$$

$$\mathcal{L}(\widetilde{\mathcal{G}}_{m,k}) = \frac{1}{2m\Delta_{\max}}\Big(k\Big\{\log(m-2)! - \log k! - (m-k-2)\log k\Big\} + \left\{\binom{k(k-3)}{2} - 1\right\}\log 2\Big),$$

$$\mathcal{L}(\mathcal{G}^l_m) = \frac{\log 2}{2(m-m_l)\Delta_{\max}}\left(\sum_{i=1}^{l-1} m_{i+1}m_i - 2\right),$$

$$\mathcal{L}(\mathcal{G}^l_{m,k}) = \frac{k\sum_{i=1}^{l-1} m_i \log\left(\frac{m_{i+1}}{k}\right) - 2\log 2}{2(m-m_l)\Delta_{\max}}.$$

Since the difficulty of learning a class of BNs is determined by the difficulty of learning the most difficult subset within that class, we immediately get the following corollary on the fundamental limits of learning non-sparse and sparse BNs.

Corollary 2 (Fundamental limits of BN structure learning)**.** *The necessary number of samples required to learn BNs on m variables and sparse BNs on m variables and maximum number of parents k is as follows:*

$$\mathcal{L}(\mathcal{G}_m) = \frac{\log 2}{\Delta_{\max}}\left(\frac{m-3}{2} - \frac{1}{m}\right),$$

$$\mathcal{L}(\mathcal{G}_{m,k}) = \frac{1}{2\Delta_{\max}}\Big(k\log(m-2) + \frac{k(k-3)\log 2}{2m} - R(m,k)\Big),$$

where $R(m,k) = \binom{k}{m}\Big\{(m-2)+2\log(m-2)+\log k!+(m-k-2)\log k\Big\} + \binom{\log 2}{m}$.

In the above corollary, the lower bounds for non-sparse and sparse BNs come from the results for the restricted ensembles $\widetilde{\mathcal{G}}_m$ and $\widetilde{\mathcal{G}}_{m,k}$ respectively.

Remark 4. *For sparse BNs, if k is constant with respect to m, then the reminder term in the lower bound is $R(m,k) \leq 1$ for sufficiently large m and the sample complexity grows as $\Omega(\log m)$. In this regime, our lower bound for learning sparse Bayesian networks matches the analogous lower bound of $\mathcal{O}(\log m)$ for learning bounded degree Ising models as obtained by Santhanam and Wainwright [13]. In general, however, $R(m,k) = \mathcal{O}(k\log k)$; therefore the number of samples necessary to learn sparse BNs grows as $\Omega\left(k\log m + \binom{k^2}{m}\right)$.*

Remark 5. *Note that since the number of samples required to learn both general essential DAGs ($\mathcal{L}(\widetilde{\mathcal{G}}_m)$) and layered networks ($\mathcal{L}(\mathcal{G}^l_m)$) grows as $\Omega(m)$. This, combined with the fact that our lower bounds on the number of DAG structures in the ensemble $\widetilde{\mathcal{G}}_m$ was tight, leads us to the conclusion that the ordering of variables does not add much to the difficulty of learning BNs in terms of sample complexity.*

Theorem 2 provides a simple recipe for obtaining necessary conditions on the sample complexity of learning any exponential family BN, as we demonstrate in the next section.

6 Implications for Commonly Used Bayesian Networks

In this section, we instantiate Theorem 2 for specific conditional distributions, to derive fundamental limits

	Non-Sparse	Sparse
CPT	$\frac{m}{\log(1/\theta_{\min})}$	$\frac{k\log m+k^2/m}{\log(1/\theta_{\min})}$
Gaussian	$\frac{\sigma_{\min}^2 m}{\sigma_{\min}^2+2\mu_{\max}^2(w_{\max}^2+1)}$	$\frac{\sigma_{\min}^2(k\log m+k^2/m)}{\sigma_{\min}^2+2(w_{\max}^2+1)w_{\max}^2}$
Noisy-OR	$\frac{m}{\left\|\log(\hat{\theta}/1-\hat{\theta})\right\|}$	$\frac{k\log m+k^2/m}{\left\|\log(\hat{\theta}/1-\hat{\theta})\right\|}$
Logistic	$\frac{m}{w_{\max}^1}$	$\frac{k\log m+k^2/m}{w_{\max}^1}$

Table 1: Fundamental limits of learning the structure of various types of BNs from Corollary 2, where the entries of the tables are lower bounds, i.e., $\Omega(.)$. For CPT BNs, $\theta_{\min}$ is the minimum entry in the conditional probability tables. For Gaussian BNs, $w_{\max}, \mu_{\max}$ and $\sigma_{\min}$ are the maximum ℓ_2 norm of the weight vectors, maximum absolute mean and minimum conditional variance respectively. For noisy-OR networks, $\hat{\theta} \in (0,1)$ is the failure probability. Lastly, for logistic regression, $w_{\max}^1$ is the maximum ℓ_1 norm of the weight vectors.

of learning various widely used BNs. This also allows us to highlight the role of parameters of the conditional distributions in the sample complexity of learning the DAG structure of BNs. Table 1 summarizes our results for various commonly used BNs. Proofs of results derived in this section can be found in Appendix C.

Conditional Probability Table BNs. CPT BNs are perhaps the most widely used BNs, where the conditional distribution of a node given its parents is described by probability tables. As is typically the case, we assume that the support of $X_i \in \mathcal{X} = [v]$ for all i. The conditional distribution of X_i is given by the following categorical distribution:

$$\mathcal{P}_i(x_i; \pi_i(G), \Theta) = \prod_{j=1}^{v} \left(\theta_{ij}(\mathbf{x})\right)^{\mathbf{1}[X_i=j]},$$

where $\Theta(G) \in \boldsymbol{\varphi}(\mathcal{G})$ are the set of conditional probability tables for the variables $\{X_1, \ldots, X_m\}$ compatible with the DAG structure G. Let us denote $\Theta^G \stackrel{\text{def}}{=} \Theta(G)$. The conditional probability table for the i-th random variable $\Theta_i^G : [v]^{|\pi_i|} \to \Delta_v$, maps all possible assignments to the parent set X_{π_i} to the $(v-1)$-dimensional probability simplex Δ_v, and $\theta_{ij}^G(.)$ represents the j-th entry of the v-dimensional vector $\Theta_i^G(.)$. The following lemma gives the upper bound on the mutual information $I(\mathsf{S}; G|\Theta)$ for CPT BNs.

Lemma 7 (Mutual Information bound for CPT networks)**.** *For CPT BNs we have*

$$\Delta_{\max} \leq 4\log(1/\theta_{\min}),$$
$$\sup_{\Theta\in\boldsymbol{\varphi}(\mathcal{G})} I(\mathsf{S}; G|\Theta) \leq 4nm\log(1/\theta_{\min}),$$

where $\theta_{\min} > 0$ is minimum probability value in a probability table across all node and parent set assignments i.e.,

$$\theta_{\min} \stackrel{\text{def}}{=} \inf_{\Theta\in\boldsymbol{\varphi}(\mathcal{G})} \min_{G\in\mathcal{G}} \min_{i=1}^{m} \min_{\mathbf{x}\in\mathcal{X}^{|\pi_i(G)|}} \min_{j=1}^{v} \theta_{ij}^G(\mathbf{x}).$$

Remark 6. *The necessary number of samples required to learn dense and sparse CPT BNs is $\Omega\left(\frac{m}{\log(1/\theta_{\min})}\right)$ and $\Omega\left(\frac{k\log m+k^2/m}{\log(1/\theta_{\min})}\right)$, respectively. In the regime that $\theta_{\min} \geq \exp(-1/m)$, the sample complexity for learning dense and sparse BNs is $\Omega\left(m^2\right)$ and $\Omega\left(km\log m + k^2\right)$, respectively.*

Remark 7. *The sample complexity of* SparsityBoost *algorithm by Brenner and Sontag [4] for recovering the structure of binary-valued, sparse, CPT BNs grows as $\mathcal{O}\left(\max((\log m)\mu, m^2\widehat{\mu}_P^2)\right)$, where μ is defined in [4] as "the maximum inverse probability of an assignment to a separating set over all pairs of nodes", and is $1/\theta_{\min}$ for the ensembles we consider. The parameter $\widehat{\mu}_P^2$ is also defined as the maximum inverse probability of an assignment to a separating set but relates to the true graph, G, that generated the data and can be $\ll 1/\theta_{\min}$. If* SparsityBoost *operates in the regime where the second term inside the* max *function dominates, which the authors believe to be the case, then that leads to a sufficient condition of $\mathcal{O}\left((1/\theta_{\min})m^2\right)$, which is quite far from the information-theoretic limit.*

Gaussian BNs. In this case, we assume that the support $X_i \in \mathcal{X} = \mathbb{R}$ for all i, the parameters of the i-th node $\Theta_i(G) = (\mathbf{w}_i^G, \mu, \sigma^2)$, and the conditional distributions are described by the following linear Gaussian model:

$$\mathcal{P}_i(\pi_i(G), \Theta) = \mathcal{N}(\mu_i, \sigma^2/2), \tag{11}$$
$$\mu_i = \begin{cases} (\mathbf{w}_i^G)^T \mathsf{X}_{\pi_i} & \pi_i(G) \neq \varnothing, \\ \mu & \text{otherwise} \end{cases}, \tag{12}$$

where $t_G : [m] \to [m]$ is a function that maps a node to its "topological order" as defined by the graph G. We assume that

$$\mathbf{w}_i^G \in \mathbb{B}_i^G \stackrel{\text{def}}{=} \{\mathbf{w} \in \mathbb{R}^{|\pi_i(G)|} | \left\|\mathbf{w}\right\|_2 \leq 1/\sqrt{2(t_G(i)-1)}\},$$
$$\mu \in [\mu_a, \mu_b],\ \sigma \in [\sigma_{\min}, \sigma_{\max}],\ \text{and}$$
$$\Theta(G) \in \left(\times_{i=1}^m \mathbb{B}_i^G\right) \times [\mu_a, \mu_b] \times [\sigma_{\min}, \sigma_{\max}],$$

where $-\infty < \mu_a \leq \mu_b < \infty$, $0 < \sigma_{\min} \leq \sigma_{\max} < \infty$. Accordingly, we have that $\mathcal{Q}_0 = \mathcal{P}_i(\varnothing, \Theta) = \mathcal{N}(\mu, \frac{\sigma^2}{2})$, and $\mu_i \leq \mu \left\|\mathbf{w}_i^G\right\|_2$. Once again, we first bound the mutual information $I(\mathsf{S}; G|\Theta)$ in the following lemma which we then plugin in Theorem 2 to obtain the necessary conditions for learning Gaussian BNs.

Lemma 8 (Mutual Information bound for Gaussian networks)**.** *For Gaussian BNs we have:*

$$\Delta_{\max} \leq 1 + \frac{2\mu_{\max}^2(w_{\max}^2+1)}{\sigma_{\min}^2},$$

$$\sup_{\Theta\in\boldsymbol{\varphi}(\mathcal{G})} I(\mathsf{S};G|\Theta)\leq nm\left(1+\frac{2\mu_{\max}^2(w_{\max}^2+1)}{\sigma_{\min}^2}\right),$$

where $\mu_{\max}\stackrel{\text{def}}{=}\max(|\mu_a|,|\mu_b|)$ *and* $w_{\max}$ *is the maximum* ℓ_2 *norm of the weight vectors, i.e.*

$$w_{\max}\stackrel{\text{def}}{=}\sup_{\Theta\in\boldsymbol{\varphi}(\mathcal{G})}\max_{G\in\mathcal{G}}\max_{i=1}^{m}\left\|\mathbf{w}_i^G\right\|_2.$$

Remark 8. *Invoking Theorem 2 for a two layer ordering of nodes, where there are* $m-1$ *nodes in the top layer and 1 node in the bottom layer, we recover the information-theoretic limits of linear regression. Specifically, we have that the necessary number of samples for linear regression and sparse linear regression scale as* $\Omega\left(\frac{\sigma_{\min}^2 m}{\sigma_{\min}^2+2\mu_{\max}^2 w_{\max}^2}\right)$ *and* $\Omega\left(\frac{\sigma_{\min}^2 k\log m}{\sigma_{\min}^2+2\mu_{\max}^2 w_{\max}^2}\right)$ *respectively.*

Remark 9. *The sample complexity of learning the structure of degree bounded Gaussian MRFs scales as* $\Omega\left(k\log(m/k)/\log(1+k\lambda)\right)$ *[14], where* λ *is the minimum correlation between pairs of nodes that are connected by an edge. The corresponding result for sparse BNs is* $\Omega\left(\frac{\sigma_{\min}^2(k\log m+k^2/m)}{\sigma_{\min}^2+2\mu_{\max}^2 w_{\max}^2}\right)$*, which is slightly stronger than the corresponding lower bound for learning Gaussian MRFs, with respect to sparsity index* k*.*

Noisy-OR BNs. Noisy-OR BNs are another widely used class of BNs — a popular example being the two-layer QMR-DT network [15]. They are usually parameterized by failure probabilities θ_{ij}, which in the context of the QMR-DT network of diseases and symptoms can be interpreted as the probability of not observing the i-th symptom given that the j-th disease is present. More formally, we have binary valued random variables, i.e., $X_i\in\mathcal{X}=\{0,1\}$ for all i, the i-th conditional distribution is given by the Bernoulli distribution ($\mathcal{B}$) with parameter $\Theta_i=\theta\in(0,1)$:

$$\mathcal{P}_i(\pi_i(G),\Theta_i)=\mathcal{B}(1-\theta_i)\quad\mathcal{P}_i(\varnothing,\Theta_i)=\mathcal{B}(\theta),\quad(13)$$

where $\theta_i=\theta\Big(\prod_{j\in\pi_i}\theta^{X_j}\Big)^{1/|\pi_i|}$. The following lemma bounds the mutual information for noisy-OR networks.

Lemma 9 (Mutual Information bound for Noisy-OR)**.** *For Noisy-OR BNs we have:*

$$\Delta_{\max}\leq 2\left|\log\left(\hat{\theta}/(1-\hat{\theta})\right)\right|$$
$$\sup_{\Theta\in\boldsymbol{\varphi}(\mathcal{G})} I(\mathsf{S};G|\Theta)\leq 2nm\left|\log\left(\hat{\theta}/(1-\hat{\theta})\right)\right|,$$

where $\hat{\theta}\stackrel{\text{def}}{=}\operatorname{argmax}_{\theta\in\boldsymbol{\varphi}(\mathcal{G})}|\log(\theta/(1-\theta))|$.

Remark 10. *From the above lemma we notice that recovering the structure of noisy-OR networks becomes more difficult as the failure probability* θ *moves farther away from* $1/2$*. That is because as* $\theta\to 1$*, the noisy-OR network becomes more "noisy". While, as* $\theta\to 0$*, the top level nodes (nodes with no parents) take the value 1 with low probability, in which case the child nodes take values 0 with high probability.*

Logistic Regression BNs. For logistic regression BNs, the nodes are assumed to be binary valued, i.e., $X_i\in\mathcal{X}=\{0,1\}$ for all i. Each node in the network can be thought of as being classified as "0" or "1" depending on some linear combination of the values of its parents. The parameters for the i-th conditional distribution are $\Theta_i=\mathbf{w}_i$, where the vectors $\mathbf{w}_i$ are assumed to have bounded ℓ_1 norm, i.e., $\mathbf{w}_i\in\mathbb{R}^{|\pi_i|}\wedge\|\mathbf{w}_i\|_1\leq w_{\max}^1$, for some constant $w_{\max}^1$. The conditional distribution of the nodes are given as:

$$\mathcal{P}_i(\pi_i(G),\Theta_i)=\mathcal{B}\left(\sigma\left(\langle\mathsf{X}_{\pi_i},\mathbf{w}_i\rangle\right)\right),\ \mathcal{P}_i(\varnothing,\Theta_i)=\mathcal{B}(1/2),\quad(14)$$

where $\mathcal{B}$ is the Bernoulli distribution and $\sigma(x)=(1+e^{-x})^{-1}$ is the sigmoid function. The following lemma upper bounds the mutual information for logistic regression BNs.

Lemma 10 (Mutual Information bound for Logistic regression networks)**.** *For Logistic regression BNs we have:*

$$\Delta_{\max}\leq w_{\max}^1/2,\qquad\sup_{\Theta\in\boldsymbol{\varphi}(\mathcal{G})} I(\mathsf{S},G|\Theta)\leq (nmw_{\max}^1)/2,$$

where $w_{\max}^1\stackrel{\text{def}}{=}\sup_{\Theta\in\boldsymbol{\varphi}(\mathcal{G})}\max_{i\in[m]}\|\Theta_i(G)\|_1$.

Remark 11. *Once again, we can instantiate Theorem 2 for the two-layer case and obtain necessary number of samples for support recovery in logistic regression. We have that the number of samples needed for support recovery in logistic regression scales as* $\Omega\left(k\log(m)/w_{\max}^1\right)$*. In the regime that* $w_{\max}^1\leq 1/k$*, the necessary number of samples scales as* $\Omega\left(k^2\log m\right)$*. Ravikumar et al. [3] studied support recovery in logistic regression in the context of learning sparse Ising models. The upper bound of* $\mathcal{O}\left(k^2\log m\right)$ *in Proposition 1 in [3], is thus information-theoretically optimal.*

Concluding Remarks An important direction for future work is to study the information-theoretic limits of both structure and parameter recovery of BNs. However, the analysis for that situation is complicated by the fact that one has to come up with an appropriate joint distribution on the structures and parameters of the ensembles. While it is possible to do so for BNs with specific conditional distributions, we anticipate that coming up with general results for BNs would be hard, if at all possible. Also of complimentary interest is the problem of obtaining sharp thresholds for structure learning of Bayesian networks. However, such analysis might also need to be done on a case-by-case basis for specific BNs.

References

[1] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[2] Cassio P De Campos and Qiang Ji. Efficient structure learning of Bayesian networks using constraints. *The Journal of Machine Learning Research*, 12:663–689, 2011.

[3] Pradeep Ravikumar, Martin Wainwright, and John Lafferty. High-dimensional Ising model selection using L1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[4] Eliot Brenner and David Sontag. Sparsityboost: A new scoring function for learning Bayesian network structure. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI-13)*, pages 112–121, Corvallis, Oregon, 2013. AUAI Press.

[5] Klaus-U Höffgen. Learning and robust learning of product distributions. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 77–83. ACM, 1993.

[6] Nir Friedman and Zohar Yakhini. On the sample complexity of learning Bayesian networks. In *Proceedings of the Twelfth international conference on Uncertainty in Artificial Intelligence*, pages 274–282. Morgan Kaufmann Publishers Inc., 1996.

[7] P. Abbeel, D. Koller, and A. Ng. Learning factor graphs in polynomial time and sample complexity. *UAI*, 2005.

[8] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*, volume 81. MIT press, 2000.

[9] J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. Liu. Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence Journal*, 2002.

[10] D. Chickering and C. Meek. Finding optimal Bayesian networks. *UAI*, 2002.

[11] O. Zuk, S. Margel, and E. Domany. On the number of samples needed to learn the correct structure of a Bayesian network. *UAI*, 2006.

[12] Steen A Andersson, David Madigan, and Michael D Perlman. On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs. *Scandinavian Journal of Statistics*, 24(1):81–102, 1997.

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

[14] Wei Wang, Martin J Wainwright, and Kannan Ramchandran. Information-theoretic bounds on model selection for Gaussian Markov random fields. In *ISIT*, pages 1373–1377. Citeseer, 2010.

[15] Michael A Shwe, B Middleton, DE Heckerman, M Henrion, EJ Horvitz, HP Lehmann, and GF Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. *Methods of information in Medicine*, 30(4):241–255, 1991.

[16] Bertran Steinsky. Enumeration of labelled chain graphs and labelled essential directed acyclic graphs. *Discrete Mathematics*, 270(1-3):267–278, 2003.

[17] Animashree Anandkumar, Vincent YF Tan, Furong Huang, and Alan S Willsky. High-dimensional Gaussian graphical model selection: Walk summability and local separation criterion. *The Journal of Machine Learning Research*, 13(1):2293–2337, 2012.

[18] Animashree Anandkumar, Vincent YF Tan, Furong Huang, Alan S Willsky, et al. High-dimensional structure estimation in Ising models: Local separation criterion. *The Annals of Statistics*, 40(3):1346–1375, 2012.

[19] R W Robinson. Counting unlabeled acyclic digraphs. *Combinatorial Mathematics V*, 622:28–43, 1977.