

Information-theoretic limits of learning the structure of Bayesian networks

Asish Ghoshal and Jean Honorio
 Department of Computer Science
 Purdue University
 West Lafayette, IN - 47906
 {aghoshal, jhonorio}@purdue.edu

In this paper, we study the information-theoretic limits of learning the structure of Bayesian networks, on discrete as well as continuous random variables, from a finite amount of data. We show that under certain parameterizations of the Bayesian network — the minimum number of samples required to learn the “true” network grows as $\mathcal{O}(M)$ and $\mathcal{O}(S \log M)$ for non-sparse and sparse Bayesian networks respectively — where M is the number of variables in the network and S is the maximum number of parents of any node in the network. We study various commonly used Bayesian networks, such as Conditional Probability Table (CPT) based networks, Noisy-OR networks, Logistic regression (LR) networks, and Gaussian networks. We identify various important parameters of the conditional distributions that affect the complexity of learning such models, like the maximum inverse probability for CPT networks, the failure probability for Noisy-OR networks, the ℓ_2 norm of weight vectors for LR networks and the signal and noise parameters for Gaussian networks. We also show that an existing procedure called *SparsityBoost*, by Brenner and Sontag, for learning binary CPT networks is information-theoretically optimal in the number of variables.

1 Introduction

Motivation: Bayesian networks are a class of probabilistic graphical models that describe the conditional dependencies between a set of random variables as a directed acyclic graph (DAG). Over the past few decades, Bayesian networks have found wide-spread use in a number scientific disciplines for reasoning under uncertainty. Given a Bayesian network, the typical use case is to compute the value or distribution of some variables of interest given noisy observations of other variables — a process referred to as inference. However, in many problems of practical interest, the structure of the network is not known a priori and must be inferred from data.

The problem of Bayesian network structure learning has received increased attention in the past several years. Several algorithms have been developed for consistently recovering the best Bayesian network, in terms of “closeness” to the true distribution or true structure, for some measure of closeness, from a given sample of data — for a detailed survey of various algorithms see, e.g., [1] and [2]. However, an important question that has hitherto remained unanswered

is the information-theoretic limits of learning Bayesian networks, i.e. “What is the minimum number of samples required to recover the true structure of a Bayesian network irrespective of the particular procedure used?”. The answer to this question would help shed light on whether 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 Bayesian networks over M variables. We use Fano’s inequality as the main tool to obtain our results.

Comparison with Markov Random Fields: The dependence of sample complexity on the structure of the Bayesian network is captured through the sparsity index, S , of the Bayesian network, which is the maximum number of parents any node is allowed to have. The key difference between learning undirected graphical models like Markov random fields (MRF) and directed models like Bayesian networks is as follows. The complexity of learning MRFs is characterized in terms of parameters of the joint distribution over nodes, which in turn relates to the overall graph structure, while the complexity of learning Bayesian networks can be characterized by parameters of local conditional distributions of the nodes only. For CPT Bayesian networks, the limits on structure recovery depend on the maximum inverse probability, $1/\theta_{\min}$, across nodes and parent set assignments. For noisy-OR networks the complexity of structure recovery depends on the failure probability θ , while for Logistic regression Bayesian networks the complexity is characterized in terms of the ℓ_2 norm, or the dual of the S -support norm for the sparse case, of the weight vectors. Finally, for Gaussian Bayesian networks, the necessary conditions for structure recovery depend on the maximum signal, $\mu_{\max}^2$, and noise variance σ^2 .

Contribution: The main contributions of this paper are as follows:

- (I) We derive necessary conditions on the number of samples required to learn various widely used Bayesian networks like Conditional Probability Table (CPT) based networks, Noisy-OR Bayesian networks, Logistic regression networks and Gaussian Bayesian networks. For sparse networks the sample complexity grows as $\mathcal{O}(S \log M)$, while for non-sparse networks it is $\mathcal{O}(M)$.
- (II) The *SparsityBoost* algorithm developed by Brenner and Sontag [3] for learning binary CPT Bayesian networks has a sample complexity of $\mathcal{O}((1/\theta_{\min}) \log M)$, under certain conditions. We show that the rate approaches the information-theoretic limit of $\mathcal{O}\left(\frac{(\log(1/\theta_{\min}))S \log M}{1-\theta_{\min}}\right)$.
- (III) Finally, we demonstrate optimality of ℓ_1 -regularized logistic regression for support recovery. The necessary conditions on the sample complexity derived in this paper, $\mathcal{O}\left(\frac{\sqrt{S} \log M}{\max_l \|\mathbf{w}^l\|_{*,(S)}}\right)$, are complimentary to the sufficient conditions, $\mathcal{O}(S^2 \log M)$, obtained by Ravikumar [4].

Of lesser interest, yet interesting, is our comparison of the complexity of learning Gaussian Bayesian networks with Gaussian MRFs. We find that the information-theoretic limits are the same for the two in terms of the number of variables M and sparsity index S . However, for MRFs the necessary conditions depend on the minimum value entry of the partial correlation matrix of the joint distribution. For Bayesian networks the necessary conditions depend on the noise variance of the conditional distributions only.

Outline: The rest of the paper is organized as follows. In section 2 we discuss relevant prior work. Section 3 formalizes the problem of Bayesian network structure learning. In sections 4, we present our main result with a discussion of our overall solution approach. Then in section

5 we derive sample complexity results for commonly used Bayesian networks on discrete and continuous variables. Finally, we conclude by discussing potential avenues for future work in section 6.

2 Related Work

Some of the earliest sample complexity results for learning Bayesian networks were derived by Höffgen [5] and Friedman and Yakhini [6]. In both [5] and [6] the authors provide upper bounds on sample complexity of learning a Bayesian network 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. In other words, they do not recover the true structure of the network. Moreover, in [5] the authors consider boolean variables and each variable is allowed to have at most k parents. Abbeel et al. in [7] provide polynomial sample complexity results for learning likelihood consistent factor graphs.

Among sample complexity results for learning structure consistent Bayesian networks, 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 [3] provide upper bounds on the sample complexity of recovering the structure of sparse Bayesian networks. However, they consider binary valued variables only and the sample complexity grows as $\mathcal{O}(M^2)$.

There has been some prior work on information-theoretic limits of learning graphical models. However, those works have focused on a specific class of undirected graphical models called Markov Random Fields, while in this paper we focus on directed graphical models i.e. Bayesian networks. Santhanam and Wainwright [12] obtained necessary and sufficient conditions for learning pairwise binary Markov Random Fields. The necessary and sufficient conditions on the number of samples scaled as $\mathcal{O}(S^2 \log M)$ and $\mathcal{O}(S^3 \log M)$ respectively, where S is the maximum node degree and M is the number of nodes. Information theoretic limits of learning Gaussian MRFs was studied by Wang et al. [13] and for walk-summable Gaussian networks, by Anandkumar et al. [14]. In [15], Anandkumar et al. obtain a necessary condition of $\Omega(c \log M)$ for structure learning of Erdős-Rényi random Ising models, where c is the average node degree.

3 Problem Formulation

The key to using Fano's inequality to obtain necessary conditions on the number of samples required, is to carefully design a restricted ensemble of Bayesian networks which is amenable to rigorous mathematical analysis and at the same time provides tight and interesting bounds on the sample complexity. In this paper we consider restricted ensembles of Bayesian networks over M nodes where the graph structure decomposes into several layers, with nodes in each layer only allowed to have parents in the layer above it. This is a generalization over two layer Bayesian networks like the QMR-DT model [16]. The various ensembles considered in this paper are illustrated in Figure 1. Specifically, let $\mathcal{F}_{M,L}$ denote the family of Bayesian networks over M nodes organized into L layers. Let $\mathbf{X} = \{\{X_i^l\}_{i=1}^{M_l}\}_{l=1}^L$ denote all the nodes in the network with X_i^l being the i -th node in the l -th layer and M_l being the number of nodes in layer l . A

specific Bayesian network $f \in \mathcal{F}_{M,L}$ specifies the number of nodes M_l in each layer and the parent set for each node X_i^l , i.e. $f = \{(M_l, \pi_i^l) | \pi_i^l \subseteq \{1, \dots, M_{(l+1)}\}\}_{l=1}^L$, where $M_{L+1} = 0$. Let $\mathcal{P}_f$ denotes the family of distributions where the joint distribution $\Pr\{\mathbf{X}\}$ factorizes according to the Bayesian network f as follows:

$$\mathcal{P}_f \equiv \Pr\{\mathbf{X}\} = \prod_{l=1}^L \prod_{i=1}^{M_l} \Pr\{X_i^l | \mathbf{X}_{\pi_i^l}^{l+1}\}, \quad (1)$$

where $\mathbf{X}_{\pi_i^l}^{l+1} = \{X_i^{l+1} | i \in \pi_i^l\}$. To simplify notation, we drop the superscript $(l+1)$ and write $\mathbf{X}_{\pi_i^l}$ instead of $\mathbf{X}_{\pi_i^l}^{l+1}$, since it is understood that the parents of nodes in the l -th layer are in the $(l+1)$ -th layer. Finally we have that the nodes in the top most layer, i.e. layer L , have no parents i.e. $\Pr\{X_i^L | \mathbf{X}_{\pi_i^L}\} = \Pr\{X_i^L\}$. We also consider a family of sparse Bayesian networks $\mathcal{F}_{M,L,S}$ where for any node X_i^l , the number of parents is at most S i.e. $|\pi_i^l| \leq S$ and $S \ll \min(M_2, \dots, M_L)$. Finally, we consider two important special cases of both sparse and non-sparse Bayesian networks. One, where the number of nodes in each layer is the same and is given by M/L . We denote such networks by $\mathcal{F}_{M/L}$ ($\subset \mathcal{F}_{M,L}$) for the non-sparse case and by $\mathcal{F}_{M/L,S}$ ($\subset \mathcal{F}_{M,L,S}$) for the sparse case. The other special case being two layer networks with only one node in the first layer and $M-1$ nodes in the second layer. We denote such networks by $\mathcal{F}_{M-1}$ ($\subset \mathcal{F}_{M,L}$) for the non-sparse case and by $\mathcal{F}_{M-1,S}$ ($\subset \mathcal{F}_{M,L,S}$) for the sparse case.

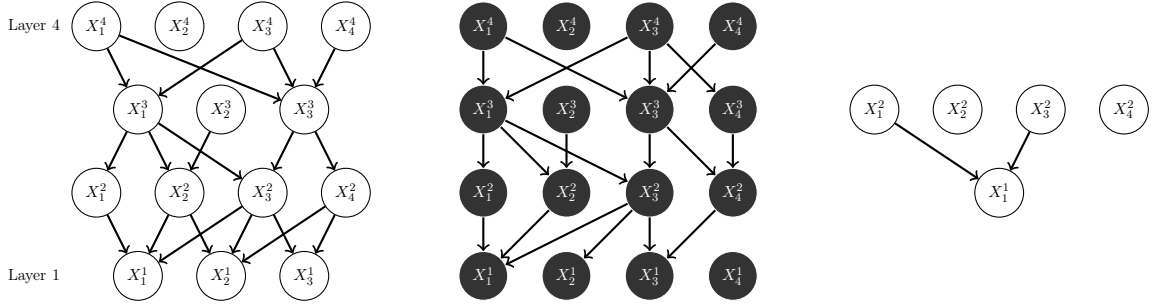


Figure 1: **(First)** An example of a Bayesian network in $\mathcal{F}_{M,L}$ with variable number of nodes in each layer. **(Middle)** An example of a Bayesian network in $\mathcal{F}_{M/L}$ where all the layers have the same number of nodes. **(Last)** An example of a Bayesian network in $\mathcal{F}_{M-1}$.

4 Main Results

In this section, we present our main result on the information-theoretic limits of learning Bayesian networks. The main tool from information theory that we use to obtain lower bounds on the sample complexity is the Fano's inequality. A key step in using the Fano's inequality, however, is to bound the maximum Kullback-Leibler (KL) divergence between the likelihood of the data computed using any two networks in the restricted ensemble. The following technical lemma provides such a bound.

Lemma 1 (KL bound). *Given a data set $\mathcal{S} = \{\mathbf{X}^{(n)}\}_{n=1}^N$, of N i.i.d observations drawn according to some Bayesian network in $\mathcal{F}_{M,L}$. The Kullback-Leibler (KL) divergence between the joint distribution of the data set specified by two Bayesian networks $f = \{(M_l, \pi_i^l) | \pi_i^l \subseteq$*

$\{1, \dots, M_{(l+1)}\}_{l=1}^L$ and $f' = \{(M_l, (\boldsymbol{\pi}')_i) | (\boldsymbol{\pi}')_i \subseteq \{1, \dots, M_{(l+1)}\}\}_{l=1}^L \in \mathcal{F}_{M,L}$ is bounded as follows:

$$\mathbb{KL}(\mathcal{P}_{S|f} \| \mathcal{P}_{S|f'}) = N \sum_{l=1}^{L-1} \sum_{m=1}^{M_l} \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\mathbb{KL} \left(\mathcal{P}_{X_i^l | \mathbf{X}_{\pi_i^l}} \| \mathcal{P}_{X_i^l | \mathbf{X}_{(\pi')_i^l}} \right) \right] \quad (2)$$

$$\leq N(M - M_L)K_{\max}, \quad (3)$$

where we have defined $K_{\max}$ as the maximum KL divergence across layers 1 to $L-1$ and across all nodes in those layers:

$$K_{\max} \equiv \max_{l=1}^{L-1} \max_{i=1}^{M_l} \left\{ \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\mathbb{KL} \left(\mathcal{P}_{X_i^l | \pi_i^l} \| \mathcal{P}_{X_i^l | (\pi')_i^l} \right) \right] \right\}.$$

Missing proofs of lemmas can be found in the Appendix. The following theorem describes the necessary conditions to recover the true structure of a Bayesian network from a finite sample of data.

Theorem 1 (Necessary conditions for learning the structure of Bayesian networks). *Assume there exists a true Bayesian network $\bar{f}$ picked uniformly at random from some family $\mathcal{F}$. Further assume that we are given a data set $\mathcal{S}$ of N i.i.d samples, for each of the M variables, drawn from the Bayesian network $\bar{f}$. The learner infers the network $\hat{f}$ from the sample data set $\mathcal{S}$ using any procedure. Then, if $N \leq g(M, L, K_{\max})$, with probability at least $1/2$, $\bar{f} \neq \hat{f}$. The function $g(M, L, K_{\max})$, for various classes of Bayesian networks, is given as follows:*

- (A) If $\mathcal{F} = \mathcal{F}_{M,L}$, then $g(M, L, K_{\max}) = \frac{(\sum_{l=1}^{L-1} M_l M_{l+1} - 2) \log 2}{2K_{\max} \sum_{l=1}^{L-1} M_l}$. For networks in $\mathcal{F}_{M/L}$, i.e. the networks where the number of nodes in each layer is M/L , the function $g(M, L, K_{\max}) = \mathcal{O}\left(\frac{M/L}{K_{\max}}\right)$. While, for networks in $\mathcal{F}_{M-1}$, the function $g(M, L, K_{\max}) = \mathcal{O}\left(\frac{M}{K_{\max}}\right)$.
- (B) If $\mathcal{F} = \mathcal{F}_{M,L,S}$ then $g(M, L, K_{\max}) = \frac{S \sum_{l=1}^{L-1} M_l \log(M_{l+1}/S) - 2 \log 2}{2K_{\max} \sum_{l=1}^{L-1} M_l}$. For networks in $\mathcal{F}_{M/L,S}$, the function $g(M, L, K_{\max}) = \mathcal{O}\left(\frac{S \log(M/L)}{K_{\max}}\right)$. While, for networks in $\mathcal{F}_{M-1,S}$, the function $g(M, L, K_{\max}) = \mathcal{O}\left(\frac{S \log M}{K_{\max}}\right)$.

Proof Sketch. In order to prove the above theorem, we compute the number of Bayesian networks in each of the different ensembles. Then we use an averaged pairwise KL-based bound from Yu [17] to bound the mutual information by the maximum KL divergence $K_{\max}$. Finally, we invoke the Fano's inequality to prove our claim. $\square$

Thus we conclude that for the family of deep Bayesian networks, $\mathcal{F}_{M/L}$, the minimum number of samples required grows as $\mathcal{O}(M/L)$ and for Bayesian networks with sparsity index, S , as $\mathcal{O}(S \log M/L)$. The dependence on the number of nodes per layer, M/L , instead of the total number of nodes, M , is a consequence of the fact that nodes in a given layer only depend on the layer above it. Such networks, clearly, are easier to learn than general Bayesian networks. The maximum KL divergence term, $K_{\max}$, in the sample complexity relates the complexity of learning a Bayesian network to the parameters of the conditional distributions. A high value of $K_{\max}$ indicates that parameters encode a lot of information about the sample data, and thus varying the parameters results in high variability of the likelihood of the data. Such Bayesian networks are again easier to learn than classes of networks with small $K_{\max}$. To derive sufficient conditions for specific Bayesian network families like CPT Bayesian networks, noisy-OR networks or LR

Bayesian networks we make use of the following technical lemma that relates the maximum KL divergence between (conditional) distributions, belonging to the exponential family, to the norm of the natural parameters.

Lemma 2 (KL bound for exponential family). *In Bayesian networks where the conditional distribution of nodes are described by distributions that belong to the exponential family, the maximum Kullback-Leibler divergence across all layers and nodes is given as follows:*

$$K_{\max} = 4 \max_{l=1}^{L-1} \max_{i=1}^{M_l} \max_{\pi_i^l} \left\{ \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\left\| \mathbb{E}_{\eta_{\pi_i^l}} [\mathbf{T}(X)] \right\|_* \left\| \eta_{\pi_i^l} \right\| \right] \right\},$$

where $\eta_{\pi_i^l}$ is the natural parameter for the conditional distribution on X_i^l , which in turn is a function of $\mathbf{X}_{\pi_i^l}^{l+1}$.

Proof Sketch. The main tools that we use to prove the above lemma are the Cauchy-Schwartz inequality, various properties of the exponential family of distributions and the mean value Theorem. $\square$

Note that, both the expected sufficient statistic $\mathbb{E}[\mathbf{T}(X)]$ and the natural parameter $\eta_{\pi_i^l}$, are functions of some subset of the parent nodes $\mathbf{X}^{(l+1)}$, and are therefore not independent in general. Therefore, the trick to coming up with a closed form expression for $K_{\max}$ is to bound the expected sufficient statistic by a constant which, as we show in subsequent sections, is often possible for discrete valued random variables. In the next section we instantiate Theorem 1 to obtain sufficient conditions on learning various commonly used Bayesian networks, the results of which are summarized in Table 1.

	CPT	Noisy-OR	Logistic	Gaussian
Non-Sparse	$\mathcal{O}\left(\frac{M \log(1/\theta_{\min})}{Lv\theta_{\max}}\right)$	$\mathcal{O}\left(\frac{M}{L \log \theta }\right)$	$\mathcal{O}\left(\frac{\sqrt{M}}{BL}\right)$	$\mathcal{O}\left(\frac{\sigma^2 M}{L(\mu_{\max}^2 + \sigma^2(L-1))}\right)$
Sparse	$\mathcal{O}\left(\frac{S \log(M/L) \log(1/\theta_{\min})}{v\theta_{\max}}\right)$	$\mathcal{O}\left(\frac{S \log(M/L)}{ \log \theta }\right)$	$\mathcal{O}\left(\frac{\sqrt{S} \log(M/L)}{\max_l \ \mathbf{w}^l\ _{*,(S)}}\right)$	$\mathcal{O}\left(\frac{\sigma^2 S \log(M/L)}{(\mu_{\max}^2 + \sigma^2(L-1))}\right)$

(a) Results for deep Bayesian networks: M variables are organized into L layers with each layer having M/L variables.

	CPT	Noisy-OR	Logistic	Gaussian
Non-Sparse	$\mathcal{O}\left(\frac{M \log(1/\theta_{\min})}{v\theta_{\max}}\right)$	$\mathcal{O}\left(\frac{M}{ \log \theta }\right)$	$\mathcal{O}\left(\frac{\sqrt{M}}{B}\right)$	$\mathcal{O}\left(\frac{M\sigma^2}{\mu_{\max}^2 + \sigma^2}\right)$
Sparse	$\mathcal{O}\left(\frac{S \log M \log(1/\theta_{\min})}{v\theta_{\max}}\right)$	$\mathcal{O}\left(\frac{S \log M}{ \log \theta }\right)$	$\mathcal{O}\left(\frac{\sqrt{S} \log M}{\ \mathbf{w}\ _{*,(S)}}\right)$	$\mathcal{O}\left(\frac{\sigma^2 S \log M}{\mu_{\max}^2 + \sigma^2}\right)$

(b) Results for shallow Bayesian networks: $M - 1$ variables in the second layer and one variable in the first layer.

Table 1: Necessary conditions on the number of samples required to learn sparse and non-sparse Bayesian networks. The columns correspond to the probability distribution family that describes the conditional distributions of the variables in the network. For sparse networks, S is the sparsity index, i.e. the maximum number of parents of any node. For CPT, $\theta_{\max}$ and $\theta_{\min}$ are the maximum and minimum probabilities across all assignments to nodes and parents and v is the number values the random variable can take. For noisy-OR, θ is the failure probability, e.g. in the context of the QMR-DT model, of not observing a symptom even when the disease is present. For the Logistic case, B is the bound on the ℓ_2 norm of the weight vectors $\mathbf{w}^l$. $\|\mathbf{w}^l\|_{*,S}$ is the ℓ_2 norm of the S largest elements of $\mathbf{w}^l$. Finally, for the Gaussian case $\mu_{\max}$ and σ^2 are the maximum absolute prior mean and the noise variance respectively.

5 Implications for Commonly Used Bayesian Networks

In this section we obtain sufficient conditions for learning various widely used Bayesian networks like Conditional Probability Table (CPT) based networks, noisy-OR Bayesian networks, logistic regression networks and finally Gaussian Bayesian networks. The results for the different cases are summarized in Table 1. The key step for deriving the result for different networks is to bound the quantity $1/K_{\max}$.

5.1 Conditional Probability Table Bayesian Networks

Conditional Probability Table based Bayesian networks are perhaps the most widely used Bayesian networks, where conditional distribution of a node given its parents is described by probability tables. As is typically the case, we assume that the variables X_i^l take v possible values, i.e. $X_i^l \in \{1, \dots, v\}$. The conditional distribution of X_i^l is given by the following categorical distribution:

$$(\forall 1 \leq l \leq L-1) \quad \Pr \left\{ X_i^l | \mathbf{X}_{\pi_i^l}, \boldsymbol{\Theta}^{\pi_i^l} \right\} = \prod_{k=1}^v \left(\Theta_k^{\pi_i^l}(\mathbf{X}_{\pi_i^l}) \right)^{1[X_i^l=k]},$$

where $\boldsymbol{\Theta}^{\pi_i^l} : \{1, \dots, v\}^{|\pi_i^l|} \rightarrow \Delta_v$, is a conditional probability table that maps all possible assignments to the parent set $\mathbf{X}_{\pi_i^l}$ to the $(v-1)$ -dimensional probability simplex Δ_v . Also, $\Theta_k^{\pi_i^l}(\mathbf{X}_{\pi_i^l})$ represents the k -th entry of the v -dimensional vector $\boldsymbol{\Theta}^{\pi_i^l}(\mathbf{X}_{\pi_i^l})$. The probability tables for each node and all possible parent sets are assumed to be known and fixed beforehand. Finally, we also assume that the marginal distribution of X_i^L is parameterized by some fixed vector $\boldsymbol{\theta} \in \Delta_v$. The following lemma gives the upper bound on $K_{\max}$ for CPT distributions.

Lemma 3 (KL bound for Conditional Probability Table). *If the conditional distribution of nodes in the Bayesian network is described by conditional probability tables, then the maximum Kullback-Leibler (KL) divergence is bounded as follows:*

$$K_{\max} \leq v\theta_{\max} \log \left(\frac{1}{\theta_{\min}} \right), \quad \frac{1}{K_{\max}} \geq \frac{1}{v\theta_{\max} \log(1/\theta_{\min})},$$

where $\theta_{\max}$ and $\theta_{\min} > 0$ are the maximum and minimum probability values across all node and parent set assignments i.e.

$$\theta_{\max} \equiv \max_{l,i,\pi_i^l,\mathbf{x}_{\pi_i^l}} \max_{k=1}^v \Theta_k^{\pi_i^l}(\mathbf{x}_{\pi_i^l}) \quad \theta_{\min} \equiv \min_{l,i,\pi_i^l,\mathbf{x}_{\pi_i^l}} \min_{k=1}^v \Theta_k^{\pi_i^l}(\mathbf{x}_{\pi_i^l}),$$

and $\mathbf{x}_{\pi_i^l}$ is a particular assignment to the random variable $\mathbf{X}_{\pi_i^l}$.

Remarks: Note that for binary CPT networks, $\theta_{\max} = 1 - \theta_{\min}$, which leads to the lower bound of $\mathcal{O} \left(\frac{(S \log M) \log(1/\theta_{\min})}{1 - \theta_{\min}} \right)$ for sparse networks. The number of samples required by the *SparsityBoost* algorithm by Brenner and Sontag [3] scales as $\mathcal{O}(\max((\log M)m, M^2 \hat{m}_P^2))$, where m is defined in [3] as “the maximum inverse probability of an assignment to a separating set over all pairs of nodes”. That is an ensemble parameter and is $1/\theta_{\min}$ for the ensembles we consider. The parameter $\hat{m}_P^2$ is also defined as the maximum inverse probability of an assignment to a separating set but relates to the true graph, $\bar{f}$, that generated the data and can be $\ll 1/\theta_{\min}$. If *SparsityBoost* operates in the regime where the first term inside the max function dominates,

then that leads to a sufficient condition of $\mathcal{O}((1/\theta_{\min})(\log M))$ which is information-theoretically optimal in terms of number of variables. However, since the quantity $1/\theta_{\min}$ can be fairly large, the gap — as a function of the parameter $\theta_{\min}$ — between the performance of *SparsityBoost* and the fundamental limit, opens some interesting avenues for further research in learning CPT based Bayesian networks.

5.2 Noisy-OR Bayesian Networks

Noisy-OR Bayesian networks are another widely used class of Bayesian networks — a popular example being the two-layer QMR-DT network [16]. 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. In our case, we set $\theta_{ij} = \theta$ for all nodes i and j . The complete probabilistic specification is given in (4).

$$X_i^L \sim \text{Bernoulli}(\theta), \quad (\forall l < L) X_i^l | \mathbf{X}_{\pi_i^l} \sim \text{Bernoulli}(1 - \theta_i^l), \quad (4)$$

where $\theta_i^l = \theta \left(\prod_{j \in \pi_i^l} \theta^{X_j^{l+1}} \right)^{(1/|\pi_i^l|)}$ and $\theta \in (0, 0.5]$. In our parameterization it is implicit that $\theta_i^l = \theta$ if $\pi_i^l = \emptyset$. The following Lemma bounds the maximum KL divergence for noisy-OR networks.

Lemma 4 (KL bound for Noisy-OR). *If the conditional distribution of nodes in the Bayesian network is described by the noisy-OR distribution, then the maximum Kullback-Leibler (KL) divergence is bounded as follows:*

$$K_{\max} \leq 2 |\log \theta|, \quad \frac{1}{K_{\max}} \geq \frac{1}{2 |\log \theta|}$$

where $M_{\max}$ is the maximum number of nodes in any layer between layer 2 to layer L , i.e. $M_{\max} \equiv \max(M_2, \dots, M_L)$.

Remarks: Note that in our parameterization of the noisy-OR distribution in (4), we rescaled the conditional probability value by the number of parents of the node which gives rise to a sample complexity that is linear in the number of nodes, or for sparse networks, logarithmic in the number of nodes. Using the usual parameterization of noisy-OR distribution, gives a sample complexity that is $\mathcal{O}(1)$ in the number of nodes and $\mathcal{O}(|1/(\log \theta)|)$ in the parameter θ . Therefore, as θ increases, the noisy-OR Bayesian network becomes easier to learn — this is to be expected, because as the failure probability $\theta \rightarrow 0$, the noisy-OR becomes less “noisy” and in the limiting case of $\theta = 0$ becomes completely deterministic.

5.3 Logistic Regression Bayesian Networks

For logistic regression Bayesian networks the nodes are assumed to be binary valued. Each node in the network can be thought of as being classified as “0” or “1” depending on the values of its parents and some weight vector associated with the parent set. We assume that for each level $l \in [2, L]$, we have a weight vector $\mathbf{w}^l \in \mathbb{R}^{M_l}$ with bounded ℓ_2 norm, i.e. $\|\mathbf{w}^l\|_2 \leq B$, for some constant B . The marginal and conditional distribution of the nodes is given as:

$$X_i^L \sim \text{Bernoulli}(0.5), \quad (\forall l < L) X_i^l | \mathbf{X}_{\pi_i^l} \sim \text{Bernoulli} \left(\sigma \left(\left\langle \mathbf{X}_{\pi_i^l}, \mathbf{w}_{\pi_i^l}^{(l+1)} \right\rangle \right) \right), \quad (5)$$

where $\sigma(x) = 1/(1+\exp(-x))$, is the sigmoid function. Note that if $\pi_i^l = \emptyset$, then $\langle \mathbf{X}_{\pi_i^l}, \mathbf{w}_{\pi_i^l}^{(l+1)} \rangle = 0$ and therefore $X_i^l | \mathbf{X}_{\pi_i^l} \sim \text{Bernoulli}(0.5)$. The following lemma provides the upper bound on $K_{\max}$ for logistic regression networks.

Lemma 5 (KL bound for Logistic regression). *For Bayesian networks where the conditional distribution of nodes follows the logit model as given in (5), the maximum Kullback-Leibler divergence, $K_{\max}$, is given as:*

$$K_{\max} \leq \begin{cases} \max_{l=2}^L \sqrt{S} \|\mathbf{w}^l\|_{*,(S)} & (\text{if } \forall l, i \ |\pi_i^l| \leq S) \\ B\sqrt{M_{\max}} & (\text{otherwise}) \end{cases},$$

$$\frac{1}{K_{\max}} \geq \begin{cases} 1 / \max_{l=2}^L \sqrt{S} \|\mathbf{w}^l\|_{*,(S)} & (\text{if } \forall l, i \ |\pi_i^l| \leq S) \\ 1/B\sqrt{M_{\max}} & (\text{otherwise}) \end{cases},$$

where $\|\mathbf{w}\|_{*,(S)}$ is the dual of the S -support norm, which was introduced in [18], and is defined as $\|\mathbf{w}\|_{*,(S)} = \sqrt{\sum_{j=1}^S (|w_j^\downarrow|)^2}$.

Remarks: The results for the shallow case of LR Bayesian networks has implications for support recovery in logistic regression, where the nodes in the top layer can be thought of as predictors while the single node in the first layer can be thought of as the output variable. Ravikumar et al. [4] studied support recovery in logistic regression in the context of learning sparse Ising models. The upper bound of $\mathcal{O}(S^2 \log M)$ in Proposition 1 in [4], can be compared to our lower bound of $\mathcal{O}(\sqrt{S} \log M)$. If instead of bounding the ℓ_2 norm of the weight vectors we bound the ℓ_1 norm by the constant B , then the necessary number of samples scale as $\mathcal{O}(S \log M)$. The reason for an additional factor of S between the necessary and sufficient number of samples as derived in [4] remains unclear.

5.4 Gaussian Bayesian networks

In the case of Gaussian Bayesian networks, we assume that each node is the average of it's parents in the layer above it, corrupted by Gaussian noise of mean 0 and variance σ^2 . The complete probability model is described in 6.

$$X_i^L \sim \mathcal{N}(\mu_i, \sigma^2), \quad (\forall 1 \leq l < L) \ X_i^l | \mathbf{X}_{\pi_i^l} \sim \mathcal{N}(\mu_i^l, \sigma^2), \quad (6)$$

where $\mu_i^l = \frac{1}{|\pi_i^l|} \sum_{j \in \pi_i^l} X_j^{l+1}$. We assume that the nodes in layer L are independent. Once again, we first bound the maximum KL divergence in the following lemma which we then plugin in Theorem 1 to obtain the necessary conditions for learning Gaussian Bayesian networks.

Lemma 6 (KL bound for Gaussian). *For Bayesian networks, where the conditional distribution of the nodes is described by the Normal distribution as given in (6). The maximum Kullback-Leibler divergence, $K_{\max}$, is given as*

$$K_{\max} \leq \frac{2\mu_{\max}^2}{\sigma^2} + 2(L-1)^2, \quad \frac{1}{K_{\max}} \geq \frac{\sigma^2}{2\mu_{\max}^2 + 2\sigma^2(L-1)^2},$$

where we have defined $\mu_{\max} = \max_{i=1}^{M_L} |\mu_i|$.

Remarks: The fundamental limits of learning Gaussian Bayesian networks can be compared with those of learning Gaussian MRFs as developed by Wang et al. [13]. In [13], under certain assumptions on the partial correlation matrix of the joint distribution of the nodes, the necessary number of samples needed to learn degree bounded Gaussian MRFs scales as $\mathcal{O}(S \log(M/S))$, with S being the bound on the degree of all nodes. This scaling is exactly the same for Gaussian Bayesian networks. However, while the difficulty of learning Gaussian MRFs depend on partial correlation matrix of the joint distribution, for Bayesian networks the difficulty can be characterized by the noise variance (σ^2) and the maximum signal ($\mu_{\max}^2$) of local conditional distributions.

6 Conclusion

In this paper we derived fundamental lower bounds on the sample complexity of learning Bayesian networks. We showed that the minimum number of samples required grows linearly with the number of variables M . For sparse networks, the minimum number of samples required to learn the true network reduces to $\mathcal{O}(S \log M)$. For binary CPT Bayesian networks, we showed that the sample complexity of $\mathcal{O}((1/\theta_{\min}) \log M)$ for *SparsityBoost* [3] approaches the information-theoretic limit of $\mathcal{O}((1/(1 - \theta_{\min}))(\log(1/\theta_{\min}))S \log M)$, under certain conditions. We also showed that the sample complexity of $\mathcal{O}(S^2 \log M)$ for ℓ_1 regularized logistic regression approaches the information theoretic limit of $\mathcal{O}(\sqrt{S} \log M / \max_l \|\mathbf{w}^l\|_{*,(S)})$. In this paper we considered restricted ensembles where the parent set of a variable was restricted to the layer immediately above it. However, an interesting direction for future work is to obtain sample complexity results for the general Bayesian networks, where the parent set for a variable can be the entire set of remaining variables. The main challenge for the general case is to bound the Kullback-Leibler divergence for arbitrary DAGs. Obtaining sharp phase transitions for structure learning in Bayesian networks can also be of interest.

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] 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.
- [4] 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.
- [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] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] Bin Yu. *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, chapter Assouad, Fano, and Le Cam, pages 423–435. Springer New York, New York, NY, 1997.
- [18] Andreas Argyriou, Rina Foygel, and Nathan Srebro. Sparse Prediction with the k -Support Norm. *NIPS*, pages 1–14, 2012.

Appendix

Proofs for Main Results

Proof of Theorem 1 (Necessary conditions for learning the structure of Bayesian networks). We use an averaged pairwise KL-based bound from Yu [17] to bound the mutual information, $\mathbb{I}(\bar{f}, \mathcal{S})$, between the data set $\mathcal{S}$ and the true network $\bar{f}$. Let $\mathcal{P}_{\mathcal{S}|f}$ be the distribution of the sample data under the network f and $\mathcal{P}_{\mathcal{S}|f'}$ be the distribution of sample data under the network f' . Then the mutual information $\mathbb{I}(\bar{f}, \mathcal{S})$ is bounded as follows:

$$\mathbb{I}(\bar{f}, \mathcal{S}) \leq \frac{1}{k^2} \sum_{f \in \mathcal{F}} \sum_{f' \in \mathcal{F}} \mathbb{KL}(\mathcal{P}_{\mathcal{S}|f} \| \mathcal{P}_{\mathcal{S}|f'}). \quad (7)$$

Next we derive lower bounds for the number of Bayesian networks in the families $\mathcal{F}_{M,L}$ and $\mathcal{F}_{M,L,S}$. The number of Bayesian networks in the family $\mathcal{F}_{M,L}$ and $\mathcal{F}_{M,L,S}$ is given by the following equations respectively:

$$|\mathcal{F}_{M,L}| = \prod_{l=1}^{L-1} (2^{M_{l+1}})^{M_l}, \quad |\mathcal{F}_{M,L,S}| = \prod_{l=1}^{L-1} \left[\sum_{i=0}^S \binom{M_{l+1}}{i} \right]^{M_l}.$$

The number of Bayesian networks in $\mathcal{F}_{M,L,S}$ can be further lower bounded as:

$$|\mathcal{F}_{M,L,S}| \geq \sum_{l=1}^{L-1} M_l \log \binom{M_{l+1}}{S} \geq \sum_{l=1}^{L-1} M_l \log \left(\frac{M_{l+1}}{S} \right)^S \geq \sum_{l=1}^{L-1} M_l S \log \left(\frac{M_{l+1}}{S} \right).$$

Finally, we use the Fano's inequality, the pair wise KL bound on the mutual information (7), and Lemma 1 to compute the probability of failing to learn the true network as follows:

$$\Pr \{ \hat{f} \neq \bar{f} \} \geq 1 - \frac{\mathbb{I}(\bar{f}, \mathcal{S}) + \log 2}{\log k} \geq 1 - \frac{N(M - M_L)K_{\max} + \log 2}{\log k},$$

where k is the number of Bayesian networks in the particular ensemble under consideration. By setting the above equation to 1/2 we prove our claim. $\square$

Proof of Lemma 1 (KL bound). For two Bayesian networks f and f' in $\mathcal{F}_{M,L}$, the log likelihood ratio factorizes as follows

$$\begin{aligned} \log \frac{\Pr_f \{ \mathbf{X} \}}{\Pr_{f'} \{ \mathbf{X} \}} &= \log \left\{ \prod_{l=1}^{L-1} \prod_{i=1}^{M_l} \frac{\Pr \{ X_i^l | \mathbf{X}_{\pi_i^l} \}}{\Pr \{ X_i^l | \mathbf{X}_{(\pi')_i^l} \}} \Pr \{ X_i^L \} \right\} \\ &= \log \left\{ \prod_{l=1}^{L-1} \prod_{i=1}^{M_l} \frac{\Pr \{ X_i^l | \mathbf{X}_{\pi_i^l} \}}{\Pr \{ X_i^l | \mathbf{X}_{(\pi')_i^l} \}} \right\} \end{aligned}$$

Therefore the KL divergence between the distributions defined by the networks f and f' is given as

$$\mathbb{KL}(\mathcal{P}_{\mathbf{X}|f} \| \mathcal{P}_{\mathbf{X}|f'}) = \mathbb{E}_{\mathbf{X} \sim \mathcal{P}_f} \left[\log \frac{\Pr_f \{ \mathbf{X} \}}{\Pr_{f'} \{ \mathbf{X} \}} \right]$$

$$\begin{aligned}
&= \sum_{l=1}^{L-1} \sum_{i=1}^{M_l} \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\mathbb{E}_{X_i^l} \left[\log \frac{\Pr \{X_i^l | \mathbf{X}_{\pi_i^l}\}}{\Pr \{X_i^l | \mathbf{X}_{(\pi')_i^l}\}} \middle| \mathbf{X}^{(l+1)} \right] \right] \\
&= \sum_{l=1}^{L-1} \sum_{i=1}^{M_l} \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\mathbb{KL} \left(\mathcal{P}_{X_i^l | \mathbf{X}_{\pi_i^l}} \parallel \mathcal{P}_{X_i^l | \mathbf{X}_{(\pi')_i^l}} \right) \right] \\
&\leq (M - M_L) K_{\max}
\end{aligned}$$

From above and using the fact that the samples are i.i.d , we prove our claim. $\square$

Proof of Lemma 2 (KL bound for exponential family.) Consider the l -th layer. For clarity of exposition, let us denote the variable X_i^l as X and $\mathbf{X}^{(l+1)}$ as $\mathbf{Y}$. Then, we can rewrite $\mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\mathbb{KL} \left(\mathcal{P}_{X_i^l | \pi_i^l} \parallel \mathcal{P}_{X_i^l | (\pi')_i^l} \right) \right]$ as $\mathbb{E}_{\mathbf{Y}} \left[\mathbb{KL} \left(\mathcal{P}_{X | \mathbf{Y}_{\pi}} \parallel \mathcal{P}_{X | \mathbf{Y}_{\pi'}} \right) \right]$. The conditional distribution, $X | \mathbf{Y}_{\pi}$, belongs to the exponential family and is given as follows:

$$\Pr \{X | \mathbf{Y}_{\pi}\} = \exp \{ \boldsymbol{\eta}_{\pi}^T \mathbf{T}(X) - \psi(\boldsymbol{\eta}_{\pi}) \} h(X), \quad (8)$$

where, $\mathbf{T}(X)$ are the sufficient statistics of X , $\boldsymbol{\eta}_{\pi}$ is the natural parameter and $\psi(\boldsymbol{\eta}_{\pi})$ is the log-normalizer. We use $\boldsymbol{\eta}_{\pi}$ to indicate that the natural parameter is a function of $\mathbf{Y}_{\pi}$. Assuming $h(X) \neq 0$ for all X , the KL divergence is bounded as follows:

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}} \left[\mathbb{KL} \left(\mathcal{P}_{X | \mathbf{Y}_{\pi}} \parallel \mathcal{P}_{X | \mathbf{Y}_{\pi'}} \right) \right] &= \mathbb{E}_{\mathbf{Y}} \left[\mathbb{E}_X \left[\log \frac{\exp \{ \boldsymbol{\eta}_{\pi}^T \mathbf{T}(X) - \psi(\boldsymbol{\eta}_{\pi}) \} h(X)}{\exp \{ \boldsymbol{\eta}_{\pi'}^T \mathbf{T}(X) - \psi(\boldsymbol{\eta}_{\pi'}) \} h(X)} \middle| \mathbf{Y} \right] \right] \\
&= \mathbb{E}_{\mathbf{Y}} \left[\mathbb{E}_{\boldsymbol{\eta}_{\pi}} [\mathbf{T}(X)]^T \boldsymbol{\eta}_{\pi} - \mathbb{E}_{\boldsymbol{\eta}_{\pi'}} [\mathbf{T}(X)]^T \boldsymbol{\eta}_{\pi'} + \psi(\boldsymbol{\eta}_{\pi'}) - \psi(\boldsymbol{\eta}_{\pi}) \right] \\
&\leq \mathbb{E}_{\mathbf{Y}} \left[\|\mathbb{E}_{\boldsymbol{\eta}_{\pi}} [\mathbf{T}(X)]\|_* \|\boldsymbol{\eta}_{\pi}\| + \|\mathbb{E}_{\boldsymbol{\eta}_{\pi'}} [\mathbf{T}(X)]\|_* \|\boldsymbol{\eta}_{\pi'}\| \right] + \\
&\quad \mathbb{E}_{\mathbf{Y}} [\psi(\boldsymbol{\eta}_{\pi'}) - \psi(\boldsymbol{\eta}_{\pi})] \\
&\leq \mathbb{E}_{\mathbf{Y}} [T_{\max} (\|\boldsymbol{\eta}_{\pi}\| + \|\boldsymbol{\eta}_{\pi'}\|) + \psi(\boldsymbol{\eta}_{\pi'}) - \psi(\boldsymbol{\eta}_{\pi})]
\end{aligned}$$

, where $\mathbb{E}_{\boldsymbol{\eta}} [\cdot]$ indicates taking expectation with respect to the distribution parameterized by $\boldsymbol{\eta}$ and we have defined $T_{\max} \equiv \max_{\boldsymbol{\eta}} \|\mathbb{E}_{\boldsymbol{\eta}} [\mathbf{T}(X)]\|$. Now, from the mean value theorem we have that

$$\begin{aligned}
\psi(\boldsymbol{\eta}_{\pi'}) - \psi(\boldsymbol{\eta}_{\pi}) &= \nabla \psi(\boldsymbol{\eta}_{\hat{\pi}})^T [\boldsymbol{\eta}_{\pi'} - \boldsymbol{\eta}_{\pi}] \\
&\leq \|\mathbb{E}_{\boldsymbol{\eta}_{\hat{\pi}}} [\mathbf{T}(X)]\|_* (\|\boldsymbol{\eta}_{\pi}\| + \|\boldsymbol{\eta}_{\pi'}\|) \\
&\leq T_{\max} (\|\boldsymbol{\eta}_{\pi}\| + \|\boldsymbol{\eta}_{\pi'}\|)
\end{aligned}$$

, where $\boldsymbol{\eta}_{\hat{\pi}} = \theta \boldsymbol{\eta}_{\pi'} + (1 - \theta) \boldsymbol{\eta}_{\pi}$ for some $0 \leq \theta \leq 1$. As a result, $\boldsymbol{\eta}_{\hat{\pi}}$ lies in the natural parameter space. Therefore, we have that

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}} \left[\mathbb{KL} \left(\mathcal{P}_{X | \mathbf{Y}_{\pi}} \parallel \mathcal{P}_{X | \mathbf{Y}_{\pi'}} \right) \right] &\leq 2 \mathbb{E}_{\mathbf{Y}} [T_{\max} \|\boldsymbol{\eta}_{\pi}\| + \|\boldsymbol{\eta}_{\pi'}\|] \\
&\leq 4 \max_{\pi} \mathbb{E}_{\mathbf{Y}} [T_{\max} \|\boldsymbol{\eta}_{\pi}\|]
\end{aligned}$$

$\square$

Proofs for Commonly Used Bayesian Networks

Proof of Lemma 3 (KL bound for Conditional Probability Table). The sufficient statistics and the natural parameter for the categorical distribution is given as follows:

$$\mathbf{T}(X_i^l) = \left(\mathbf{1} [X_i^l = k] \right)_{k=1}^v \quad \eta_{\pi_i^l} = \left(\log \Theta_k^{\pi_i^l}(\mathbf{X}_{\pi_i^l}) \right)_{k=1}^v.$$

Now, the norm of the expected sufficient statistic is bounded as follows:

$$\begin{aligned} \left\| \mathbb{E}_{\eta_{\pi_i^l}} [\mathbf{T}(X_i^l)] \right\|_{\infty} &= \left\| \Theta^{\pi_i^l}(\mathbf{X}_{\pi_i^l}) \right\|_{\infty} \\ \max_{i,l,\pi_i^l} \left\| \mathbb{E}_{\eta_{\pi_i^l}} [\mathbf{T}(X_i^l)] \right\|_{\infty} &\leq \max_{i,l,\pi_i^l} \left\| \Theta^{\pi_i^l}(\mathbf{X}_{\pi_i^l}) \right\|_{\infty} \leq \theta_{\max}. \end{aligned}$$

The expected value of the natural parameter can be written as:

$$\begin{aligned} \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\left\| \eta_{\pi_i^l} \right\|_1 \right] &= \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\sum_{k=1}^v \left| \log \Theta_k^{\pi_i^l}(\mathbf{X}_{\pi_i^l}) \right| \right] \\ &= \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\sum_{k=1}^v -\log \Theta_k^{\pi_i^l}(\mathbf{X}_{\pi_i^l}) \right] \\ &= \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\sum_{k=1}^v \log \left(\frac{1}{\Theta_k^{\pi_i^l}(\mathbf{X}_{\pi_i^l})} \right) \right] \end{aligned}$$

Which results in the following bound for the maximum expected norm of the natural parameter:

$$\begin{aligned} \max_{i,l,\pi_i^l} \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\left\| \eta_{\pi_i^l} \right\|_1 \right] &\leq \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[v \log \left(\max_{i,l,\pi_i^l} \frac{1}{\Theta_k^{\pi_i^l}(\mathbf{X}_{\pi_i^l})} \right) \right] \\ &\leq v \log \left(\frac{1}{\theta_{\min}} \right) \end{aligned}$$

□

Proof of Lemma 4 (KL bound for Noisy-OR). The sufficient statistics and natural parameter for the Bernoulli distribution is given as:

$$\mathbf{T}(X_i^l) = X_i^l \quad \eta_{\pi_i^l} = \log \frac{1 - \theta_i^l}{\theta_i^l}$$

Therefore, the expected sufficient statistic is bounded as follows:

$$\left| \mathbb{E}_{\eta_{\pi_i^l}} [\mathbf{T}(X_i^l)] \right| = \theta_i^l = \theta \left(\theta \left\{ \frac{1}{|\pi_i^l|} \sum_{j \in \pi_i^l} X_j^{l+1} \right\} \right) \leq 1.$$

Also, the natural parameter is bounded as follows:

$$\mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\left| \eta_{\pi_i^l} \right| \right] = \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\left| \log \left(\frac{1}{\theta_i^l} - 1 \right) \right| \right]$$

$$\begin{aligned}
&\leq \log \left(\mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\frac{1}{\theta_i^l} \right] \right) \\
&= \log \left(\mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\theta^{-\left\{ \frac{1}{|\pi_i^l|} \sum_{j \in \pi_i^l} X_j^{(l+1)} + 1 \right\}} \right] \right) \\
&\leq \log (\theta^{-2}) = 2 |\log \theta|
\end{aligned}$$

, where in the second line we used the fact that $\theta_i^l \leq \theta \leq 1/2$ followed by Jensen's inequality. $\square$

Proof of Lemma 5 (KL bound for Logistic regression). The expected sufficient statistic is bounded as follows:

$$\mathbb{E}_{X_i^l} [\mathbf{T}(X_i^l)] = \sigma \left(\left\langle \mathbf{X}_{\pi_i^l}, \mathbf{w}_{\pi_i^l}^{(l+1)} \right\rangle \right) \leq 1.$$

Also, the natural parameter is bounded as follows

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\left| \eta_{\pi_i^l} \right| \right] &= \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\log \frac{\sigma \left(\left\langle \mathbf{w}_{\pi_i^l}^{(l+1)}, \mathbf{X}_{\pi_i^l} \right\rangle \right)}{1 - \sigma \left(\left\langle \mathbf{w}_{\pi_i^l}^{(l+1)}, \mathbf{X}_{\pi_i^l} \right\rangle \right)} \right] \\
&= \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\left\langle \mathbf{w}_{\pi_i^l}^{(l+1)}, \mathbf{X}_{\pi_i^l} \right\rangle \right] \\
&\leq \left\| \mathbf{w}_{\pi_i^l}^{(l+1)} \right\|_1 \leq \sqrt{|\pi_i^l|} \left\| \mathbf{w}_{\pi_i^l}^{(l+1)} \right\|_2
\end{aligned}$$

Using the above bounds, we can then derive the constant $K_{\max}$ as follows:

$$\begin{aligned}
K_{\max} &= \max_{i,l,\pi_i^l} \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\left| \mathbb{E}_{X_i^l} [\mathbf{T}(X_i^l)] \right| \left| \eta_{\pi_i^l} \right| \right] \\
&\leq B \sqrt{M_{\max}}
\end{aligned}$$

If the number of parents for any node is at most S , then the above reduces to:

$$K_{\max} = \max_{l=2}^L \sqrt{S} \left\| \mathbf{w}^l \right\|_{*,(S)}.$$

$\square$

Proof of Lemma 6 (KL bound for Gaussian). Let $\mathbf{A}_i^l$ be the set of top-level ancestors of the node X_i^l , i.e. $\mathbf{A}_i^l = \{j | X_i^l \text{ is a descendant of } X_j^L\}$. Also denote by $\boldsymbol{\mu}_{\mathbf{A}_i^l}$, the set of mean parameters of the nodes in $\mathbf{A}_i^l$, i.e. $\boldsymbol{\mu}_{\mathbf{A}_i^l} = \{\mu_j | j \in \mathbf{A}_i^l\}$. We can then recursively expand the node X_i^l as follows:

$$\begin{aligned}
X_i^l &= \frac{1}{|\pi_i^l|} \sum_{j \in \pi_i^l} X_j^{l+1} + \varepsilon \\
&= \boldsymbol{\alpha}^T \boldsymbol{\mu}_{\mathbf{A}_i^l} + (L - l + 1) \varepsilon
\end{aligned}$$

, where $\boldsymbol{\alpha}$ is some convex combination i.e. $\sum_j \alpha_j = 1$. Therefore, the marginal distribution of the node X_i^l is given as:

$$X_i^l \sim \mathcal{N} \left(\boldsymbol{\alpha}^T \boldsymbol{\mu}_{\mathbf{A}_i^l}, (L - l + 1)^2 \sigma^2 \right).$$

Let

$$\mu_i^l \equiv \frac{1}{|\pi_i^l|} \sum_{j \in \pi_i^l} X_j^{l+1} \text{ and } \nu_i^l \equiv \frac{1}{|(\pi')_i^l|} \sum_{j \in (\pi')_i^l} X_j^{l+1}.$$

Then we have

$$\begin{aligned} \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[\mathbb{KL} \left(\mathcal{P}_{X_i^l | \mathbf{X}_{\pi_i^l}} \left\| \mathcal{P}_{X_i^l | \mathbf{X}_{(\pi')_i^l}} \right\| \right) \right] &= \frac{1}{2\sigma^2} \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[(\mu_i^l - \nu_i^l)^2 \right] \\ &\leq \frac{2}{\sigma^2} \max_{\pi_i^l} \left\{ \mathbb{E}_{\mathbf{X}^{(l+1)}} \left[(\mu_i^l)^2 \right] \right\} \\ &\leq \frac{2}{\sigma^2} \max_{\pi_i^l} \max_{j=1}^{M_{l+1}} \mathbb{E}_{X_j^{l+1}} \left[(X_j^{l+1})^2 \right] \\ &= \max_{j=1}^{M_{l+1}} \frac{2}{\sigma^2} \left\{ (\boldsymbol{\alpha}^T \boldsymbol{\mu}_{\mathbf{A}_j^{l+1}})^2 + (L-l)^2 \sigma^2 \right\} \\ &\leq \max_{j=1}^{M_{l+1}} \frac{2}{\sigma^2} \left\{ \|\boldsymbol{\alpha}\|_1^2 \left\| \boldsymbol{\mu}_{\mathbf{A}_j^{l+1}} \right\|_\infty^2 + (L-l)^2 \sigma^2 \right\} \\ &= \frac{2\mu_{\max}^2}{\sigma^2} + 2(L-l)^2. \end{aligned}$$

□