Appendix A Comparison with Markov Random Fields.

While there has been a lot of prior work on determining the information-theoretic limits of structure recovery in Markov random fields (MRFs), which are undirected graphical models, characterizing the information-theoretic limits of learning BNs (directed models) is important in its own right for the following reasons. First, unlike MRFs where the undirected graph corresponding to a dependence structure is uniquely determined, multiple DAG structures can encode the same dependence structure in BNs. Therefore, one has to reason about Markov equivalent DAG structures in order to characterize the information-theoretic limits of structure recovery in BNs. Second, 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 BNs is characterized by parameters of local conditional distributions of the nodes. The latter presents a technical challenge, as shown in the paper, when the marginal or joint distribution of the nodes in a BN do not have a closed form solution.

A recurring theme in the available literature on information-theoretic limits of learning MRFs, is to construct ensembles of MRFs that are hard to learn and then use the Fano's inequality to lower bound the estimation error by treating the inference procedure as a communication channel. Santhanam and Wainwright [13] obtained necessary and sufficient conditions for learning pairwise binary MRFs. The necessary and sufficient conditions on the number of samples scaled as $\mathcal{O}\left(k^2 \log m\right)$ and $\mathcal{O}\left(k^3 \log m\right)$ respectively, where k is the maximum node degree. Information theoretic limits of learning Gaussian MRFs was studied by Wang et al. [14] and for walk-summable Gaussian networks, by Anandkumar et al. [17]. In [18], 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.

Appendix B Proofs of Main Results

Proof of Theorem 1 (Fano's inequality extension). Let,

$$E \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } X \neq \widehat{X} \\ 0 & \text{othewise} \end{cases},$$

and $p_{\text{err}} \stackrel{\text{def}}{=} \Pr\left\{X \neq \widehat{X}\right\}$. Then using the chain rule for entropy we can expand the conditional entropy $H(E, X|\widehat{X}, W)$ as follows:

$$H(E, X|\widehat{X}, W) = H(E|X, \widehat{X}, W) + H(X|\widehat{X}, W) \tag{15}$$
$$= H(X|E, \widehat{X}, W) + H(E|\widehat{X}, W) \tag{16}$$

Next, we bound each of the terms in (15) and (16). $H(E|X, \widehat{X}, W) = 0$ because E is a deterministic function of X and $\widehat{X}$. Moreover, since conditioning reduces entropy, we have that $H(E|\widehat{X}, W) \leq H(E) = H(p_{\text{err}})$. Using the same arguments we have the following upper bound on $H(X|E, \widehat{X}, W)$:

$$\begin{aligned} H(X|E, \widehat{X}, W) &= p_{\text{err}} H(X|E = 1, \widehat{X}, W) + (1 - p_{\text{err}}) H(X|E = 0, \widehat{X}, W) \\ &\leq p_{\text{err}} H(X|W) \end{aligned} \tag{17}$$

Next, we show that $I(\widehat{X}; X|W) \leq I(\widehat{X}; Y|W)$, which can be thought of as the conditional data processing inequality. Using the chain rule of mutual information we have that

$$\begin{aligned} I(Y, \widehat{X}; X|W) &= I(\widehat{X}; X|Y, W) + I(Y; X|W) \\ &= I(Y; X|\widehat{X}, W) + I(\widehat{X}; X|W). \end{aligned}$$

Since, conditioned on Y, X and $\widehat{X}$ are independent. We have $I(\widehat{X}; X|Y, W) = 0$.

$$\begin{aligned} &\implies I(Y; X|W) = I(Y; X|\widehat{X}, W) + I(\widehat{X}; X|W) \\ &\implies I(Y; X|W) \geq I(\widehat{X}; X|W). \end{aligned}$$

Therefore, we can bound $H(X|\widehat{X}, W)$ as follows:

$$H(X|\widehat{X}, W) = H(X|W) - I(\widehat{X}; X|W) \geq H(X|W) - I(Y; X|W). \tag{18}$$

Combining (15),(16),(17) and (18), we get:

$$\begin{aligned} & H(X|W) - I(Y;X|W) \leq H(p_{\text{err}}) + p_{\text{err}} H(X|W) \\ \implies & H(X|W) - I(Y;X|W) \leq \log 2 + p_{\text{err}} H(X|W) \\ \implies & p_{\text{err}} \geq 1 - \frac{I(Y;X|W) + \log 2}{H(X|W)} \end{aligned} \tag{19}$$

Now if X and W are independent then $H(X|W) = H(X)$. Denoting the joint distribution over X, Y, W by $\mathcal{P}_{X,Y,W}$, the conditional distribution of X, Y given W by $\mathcal{P}_{X,Y|W}$ and so on; the final claim follows from bounding the term $I(Y;X|W)$ as follows:

$$\begin{aligned} I(Y;X|W) &= \mathbb{E}_{\mathcal{P}_{X,Y,W}} \left[\log \frac{\mathcal{P}_{X,Y|W}}{\mathcal{P}_{X|W} \mathcal{P}_{Y|W}} \right] = \mathbb{E}_{\mathcal{P}_W} \left[\mathbb{E}_{\mathcal{P}_{X,Y|W=w}} \left[\log \frac{\mathcal{P}_{X,Y|W=w}}{\mathcal{P}_{X|W=w} \mathcal{P}_{Y|W=w}} \right] \right] \\ &\leq \sup_{w \in \mathcal{W}} I(X;Y|W=w). \end{aligned}$$

□

Proof of Lemma 1. First, we briefly review Steinsky's method for counting essential DAGs, which in turn is based upon Robinson's [19] method for counting labeled DAGs. The main idea behind Steinsky's method is to split the set of essential DAGs into overlapping subsets with different terminal vertices — vertices with out-degree 0. Let $A_i \subset \widetilde{\mathcal{G}}_m$ be the set of essential DAGs where the i-th node is a terminal node. Using the inclusion-exclusion principle, the number of essential DAGs is given as follows:

$$\begin{aligned} c_m &\stackrel{\text{def}}{=} \left| \widetilde{\mathcal{G}}_m \right| = |A_1 \cup \ldots \cup A_m| \\ &= \sum_{s=1}^{m} (-1)^{(s+1)} \sum_{1 \leq i_1 \leq \ldots \leq i_s \leq m} |A_{i_1} \cap \ldots \cap A_{i_s}| \,. \end{aligned} \tag{20}$$

Now, consider the term $|A_1 \cap \ldots \cap A_{m-1}|$, i.e., number of essential DAGs where nodes $[m-1]$ are terminal nodes. The number of ways of adding the m-th vertex as a terminal vertex to an arbitrary essential DAG on the nodes $[m-1]$ is: $2^{m-1} - (m-1)$. The term $m-1$ needs to be subtracted to account for edges that are not protected. Therefore, c_m is given by the following recurrence relation:

$$c_m = \sum_{s=1}^{m} (-1)^{s+1} \binom{m}{s} (2^{m-s} - (m-s))^s c_{m-s}, \tag{21}$$

where $c_0 = 1$. Using Bonferroni's inequalities we can upper bound c_m as follows:

$$\begin{aligned} c_m &\leq m(2^{m-1} - (m-1)) c_{m-1} \leq m 2^{m-1} c_{m-1} \\ &\leq m! \, 2^{m(m-1)/2}. \end{aligned} \tag{22}$$

Using Bonferroni's inequalities to lower bound c_m produces recurrence relations that have no closed form solution. Therefore, we lower bound c_m in (20) as follows:

$$\begin{aligned} c_m &= |A_1 \cup \ldots \cup A_m| \geq \max_i |A_i| \\ &= (2^{m-1} - (m-1)) c_{m-1} \geq 2^{m-2} c_{m-1} \\ &\geq 2^{(m(m-3)/2)+1}. \end{aligned} \tag{23}$$

□

Proof of Lemma 2. In this case, A_i is the set of essential DAGs where the i-th node is a terminal node and all nodes have at most k parents. Once again, using the inclusion-exclusion principle, the number of essential DAGs with at most k parents is given as follows:

$$c_{m,k} \stackrel{\text{def}}{=} |\mathcal{G}_{m,k}| = \sum_{s=1}^{m} (-1)^{(s+1)} \sum_{1 \leq i_1 \leq \ldots \leq i_s \leq m} |A_{i_1} \cap \ldots \cap A_{i_s}| \tag{24}$$

Now, consider the term $|A_1, \cap \ldots \cap, A_s|$, i.e. number of essential DAGs where nodes $\{1, \ldots, s\}$ are terminal nodes. Let G_0 be an arbitrary essential DAG over nodes $\{s+1, \ldots, m\}$, where each node has at most k parents. Let $u \in \{1, \ldots, s\}$ and $v \in \{s+1, \ldots, m\}$ be arbitrary nodes. Let G_1 be the new graph, formed by connecting u to G_0. The edge $v \to u$ is not protected, or covered, in G_1 if $\pi_v(G_0) = \pi_u(G_1) \setminus \{v\}$. For each node $v \in \{s+1, \ldots, m\}$, there is exactly one configuration in which the edge $v \to u$ is covered, i.e., when we set the parents of u to be $\pi_v(G_0) \cup \{v\}$; unless $|\pi_v(G_0)| = k$, in which case $v \to u$ is always protected. Let $\kappa(G_0)$ be the number of nodes in G_0 that have less than k parents. Then the number of ways of adding a terminal vertex u to G_0 is: $\sum_{i=0}^{k} \binom{m-s}{i} - \kappa(G_0)$ when $(m-s) > k$, and $2^k - k$ when $(m-s) \leq k$. We can simply bound $\kappa(G_0)$ by: $0 \leq \kappa(G_0) \leq m - s$. This gives a lower bound on the number of ways to add a terminal vertex to G_0 as: $\sum_{i=0}^{k} \left\{ \binom{m-s}{i} - \frac{(m-s)}{k+1} \right\} \geq \sum_{i=0}^{k} \binom{m-s-1}{i}$. Using the fact that $\max_{i=1}^{m} |A_i| \leq c_{m,k} \leq \sum_{i=1}^{m} |A_i|$, we get the following recurrence relation for upper and lower bounds on the number of essential DAGs with at most k parents:

$$c_{m,k} \leq m \left(\sum_{i=0}^{k} \binom{m-1}{i} \right) c_{m-1,k} \tag{25}$$

$$c_{m,k} \geq \left(\sum_{i=0}^{k} \binom{m-2}{i} \right) c_{m-1,k}, \tag{26}$$

where from Lemma 1 we have that $2^{(k(k-3)/2)+1} \leq c_{k,k} \leq k!\, 2^{k(k-1)/2}$. Thus, we can upper bound $c_{m,k}$ as follows:

$$c_{m,k} \leq m!\, 2^{k(k-1)/2} \prod_{j=k+1}^{m-1} \left(\sum_{i=0}^{k} \binom{j}{i} \right) \tag{27}$$

Similarly, we can lower bound $c_{m,k}$ as follows:

$$c_{m,k} \geq 2^{(k(k-3)/2)+1} \prod_{j=k+1}^{m-1} \left(\sum_{i=0}^{k} \binom{j-1}{i} \right) \tag{28}$$

Finally, using (28), we lower bound $\log c_{m,k}$ as follows:

$$\log c_{m,k} \geq ((k(k-3)/2) + 1) \log 2 + \sum_{j=k+1}^{m-1} \log \left(\sum_{i=0}^{k} \binom{j-1}{i} \right)$$

The second term in the above equation is lower bounded as follows:

$$\begin{aligned} \sum_{j=k+1}^{m-1} \log \left(\sum_{i=0}^{k} \binom{j-1}{i} \right) &\geq \sum_{j=k}^{m-2} \log \left(\sum_{i=0}^{k} \binom{j}{i} \right) \geq \sum_{j=k}^{m-2} \log \binom{j}{k} \geq \sum_{j=k}^{m-2} \log \left(\frac{j}{k} \right)^k \\ &\geq k \left\{ \log(m-2)! - \log k! - (m-k-2) \log k \right\}. \end{aligned}$$

□

Proof Lemma 3. For layered non-sparse BNs, the number of possible choices for parents of a node in layer i is $2^{m_{i+1}}$. Therefore, the total number of non-sparse Bayesian networks is given as $\prod_{i=1}^{l-1} (2^{m_{i+1}})^{m_i}$. Similarly, for the sparse case, the number of possible choices for parents of a node in layer i is $\sum_{j=0}^{k} \binom{m_{i+1}}{j}$. Therefore, the total number of sparse Bayesian networks is $\prod_{i=1}^{l-1} \left[\sum_{j=0}^{k} \binom{m_{i+1}}{j} \right]^{m_i}$. □

Proof of Lemma 4. Let $c \stackrel{\text{def}}{=} |\mathcal{G}|$, for some ensemble of DAGs $\mathcal{G}$. Denoting the conditional distribution of the data given a specific instance of the parameters Θ by $\mathcal{P}_{\mathsf{S}|\Theta}$, we have:

$$\sup_{\Theta \in \boldsymbol{\varphi}(\mathcal{G})} I(\mathsf{S}; G|\Theta) = \sup_{\Theta \in \boldsymbol{\varphi}(\mathcal{G})} \frac{1}{c} \sum_{G \in \mathcal{G}} \mathbb{KL} \left(\mathcal{P}_{\mathsf{S}|G,\Theta} \middle\| \mathcal{P}_{\mathsf{S}|\Theta} \right), \tag{29}$$

where in $\mathbb{KL}\left(\mathcal{P}_{\mathsf{S}|G,\Theta}\middle\|\mathcal{P}_{\mathsf{S}|\Theta}\right)$, Θ and G are specific instances and not random variables. For any distribution $\mathcal{Q}$ over S, we can rewrite $\mathbb{KL}\left(\mathcal{P}_{\mathsf{S}|G,\Theta}\middle\|\mathcal{P}_{\mathsf{S}|\Theta}\right)$ as follows:

$$\begin{aligned}\mathbb{KL}\left(\mathcal{P}_{\mathsf{S}|G,\Theta}\middle\|\mathcal{P}_{\mathsf{S}|\Theta}\right) &= \mathbb{E}_{\mathsf{S}}\left[\log\frac{\mathcal{P}_{\mathsf{S}|G,\Theta}}{\mathcal{Q}}\frac{\mathcal{Q}}{\mathcal{P}_{\mathsf{S}|\Theta}}\right]\\ &= \mathbb{KL}\left(\mathcal{P}_{\mathsf{S}|G,\Theta}\middle\|\mathcal{Q}\right) - \mathbb{E}_{\mathsf{S}}\left[\log\frac{\mathcal{P}_{\mathsf{S}|\Theta}}{\mathcal{Q}}\right], \end{aligned} \tag{30}$$

where the expectation $\mathbb{E}_{\mathsf{S}}[.]$ is with respect to the distribution $\mathcal{P}_{\mathsf{S}|G,\Theta}$. Now, $\mathbb{E}_{\mathsf{S}}\left[\log\frac{\mathcal{P}_{\mathsf{S}|\Theta}}{\mathcal{Q}}\right]$ can be written as follows:

$$\begin{aligned}\sum_{G\in\mathcal{G}}\mathbb{E}_{\mathsf{S}}\left[\log\frac{\mathcal{P}_{\mathsf{S}|\Theta}}{\mathcal{Q}}\right] &= \sum_{G\in\mathcal{G}}\sum_{\mathsf{S}}\Pr\{\mathsf{S}|G,\Theta\}\log\frac{\mathcal{P}_{\mathsf{S}|\Theta}}{\mathcal{Q}}\\ &= c\sum_{\mathsf{S}}\sum_{G\in\mathcal{G}}\Pr\{G\}\Pr\{\mathsf{S}|G,\Theta\}\log\frac{\mathcal{P}_{\mathsf{S}|\Theta}}{\mathcal{Q}}\\ &= c\mathbb{KL}\left(\mathcal{P}_{\mathsf{S}|\Theta}\middle\|\mathcal{Q}\right), \end{aligned} \tag{31}$$

where, once again, we emphasize that in $\mathbb{KL}\left(\mathcal{P}_{\mathsf{S}|\Theta}\middle\|\mathcal{Q}\right)$, Θ is a particular instance of the parameters and not a random variable. Combining (29), (30) and (31), and using the fact that $\mathbb{KL}\left(\mathcal{P}_{\mathsf{S}|\Theta}\middle\|\mathcal{Q}\right) > 0$, we get

$$\sup_{\Theta\in\boldsymbol{\varphi}(\mathcal{G})} I(\mathsf{S};G|\Theta) \leq \sup_{\Theta\in\boldsymbol{\varphi}(\mathcal{G})}\frac{1}{c}\sum_{G\in\mathcal{G}}\mathbb{KL}\left(\mathcal{P}_{\mathsf{S}|G,\Theta}\middle\|\mathcal{Q}\right) \tag{32}$$

□

Proof of Lemma 5 (KL bound for exponential family).

$$\mathbb{KL}\left(\mathcal{P}_1\|\mathcal{P}_2\right) = \boldsymbol{\eta}_1^T\mathbb{E}_X\left[\mathbf{T}(x)|\boldsymbol{\eta}_1\right] - \psi(\boldsymbol{\eta}_1) - \boldsymbol{\eta}_2^T\mathbb{E}_X\left[\mathbf{T}(x)|\boldsymbol{\eta}_1\right] + \psi(\boldsymbol{\eta}_2), \tag{33}$$

where for computing the expected sufficient statistic, $\mathbb{E}_X\left[\mathbf{T}(x)|\boldsymbol{\eta}_1\right]$, we take the expectation with respect to the distribution parameterized by $\boldsymbol{\eta}_1$. Now, from the mean value theorem we have that

$$\begin{aligned}\psi(\boldsymbol{\eta}_2) - \psi(\boldsymbol{\eta}_1) &= \nabla\psi(\alpha\boldsymbol{\eta}_2 + (1-\alpha)\boldsymbol{\eta}_1)^T[\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1]\\ &= (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1)^T\mathbb{E}_X\left[\mathbf{T}(x)|\alpha\boldsymbol{\eta}_2 + (1-\alpha)\boldsymbol{\eta}_1\right]\end{aligned}$$

for some $\alpha\in[0,1]$. Then we have that,

$$\begin{aligned}\mathbb{KL}\left(\mathcal{P}_1\|\mathcal{P}_2\right) &= \mathcal{T}(\boldsymbol{\eta}_1)^T[\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2] + (\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1)^T\mathcal{T}(\alpha\boldsymbol{\eta}_2 + (1-\alpha)\boldsymbol{\eta}_1)\\ &= (\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2)^T\{\mathcal{T}(\boldsymbol{\eta}_1) - \mathcal{T}(\alpha\boldsymbol{\eta}_2 + (1-\alpha)\boldsymbol{\eta}_1)\}\end{aligned} \tag{34}$$

Since the function $\mathcal{T}$ is the gradient of the convex function ψ, it is monotonic. Therefore, the function $\mathcal{T}(\alpha\boldsymbol{\eta}_2 + (1-\alpha)\boldsymbol{\eta}_1)$ takes the maximum value at the end points $\alpha = 0$ or at $\alpha = 1$. Assuming $\mathcal{T}$ is maximized at $\alpha = 0$, the i-th KL divergence term can be upper bound using (34) as:

$$0 \leq \mathbb{KL}\left(\mathcal{P}_1\|\mathcal{P}_2\right) \leq (\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2)^T\{\mathcal{T}(\boldsymbol{\eta}_1) - \mathcal{T}(\boldsymbol{\eta}_2)\}$$

On the other hand, assuming $\mathcal{T}$ is maximized at $\alpha = 1$, the i-th KL divergence term can be upper bound using (34) as:

$$0 \leq \mathbb{KL}\left(\mathcal{P}_1\|\mathcal{P}_2\right) \leq 0.$$

Therefore, clearly, $\mathcal{T}(\alpha\boldsymbol{\eta}_2 + (1-\alpha)\boldsymbol{\eta}_1)$ is maximized at $\alpha = 0$. □

Proof of Theorem 2. Setting the measure $\mathcal{P}_{\mathcal{G}}$ to be the uniform over $\mathcal{G}$, and using the Fano's inequality from Theorem 1 and the mutual information bound from Lemma 6, combined with our Assumption 1, we can bound the estimation error as follows:

$$p_{\text{err}} \geq 1 - \frac{nm\Delta_{\max} + \log 2}{\log|\mathcal{G}|}.$$

Then by using the lower bounds on the number of DAG structures in each of the ensembles from Lemmas 1, 2 and 3, and setting p_{err} to $1/2$, we prove our claim. □

Appendix C Proofs of Results for Commonly Used Bayesian Networks

Proof of Lemma 7 (Mutual Information bound for CPT networks). For CPT, the mutual information bound is representative of the case when we do not have a closed form solution for the marginal and joint distributions; yet, we can easily bound $\Delta(\boldsymbol{\eta}_i, \boldsymbol{\eta}_0)$ through a simple application of the Cauchy-Schwartz inequality, and obtain tighter bounds than the naive $\mathcal{O}(mn \log v)$ bound on the mutual information $I(\mathsf{S}; G|\Theta)$. The sufficient statistics and the natural parameter for the categorical distribution is given as follows:

$$\mathbf{T}(x) = (\mathbf{1}[x = j])_{j=1}^{v} \qquad \boldsymbol{\eta}_i(\mathsf{X}_{\pi_i}, \Theta_i) = (\log \theta_{ij}(\mathsf{X}_{\pi_i}))_{j=1}^{v}.$$

Therefore, the expected sufficient statistic $\mathcal{T}(\boldsymbol{\eta}_i) = \Theta_i(\mathsf{X}_{\pi_i})$. From that we get the following upper bound

$$\begin{aligned}
\sup_{\Theta \in \boldsymbol{\varphi}(G)} \mathbb{E}_{\mathsf{X}_{\pi_i}}[\Delta(\boldsymbol{\eta}_i, \boldsymbol{\eta}_0)] &= \sup_{\Theta \in \boldsymbol{\varphi}(G)} \mathbb{E}_{\mathsf{X}_{\pi_i}}\left[(\boldsymbol{\eta}_i - \boldsymbol{\eta}_0)^T \{\mathcal{T}(\boldsymbol{\eta}_i) - \mathcal{T}(\boldsymbol{\eta}_0)\}\right] \\
&\leq \sup_{\Theta \in \boldsymbol{\varphi}(G)} \mathbb{E}_{\mathsf{X}_{\pi_i}}\left[\|\boldsymbol{\eta}_i - \boldsymbol{\eta}_0\|_\infty \|\mathcal{T}(\boldsymbol{\eta}_i) - \mathcal{T}(\boldsymbol{\eta}_0)\|_1\right] \\
&\leq 4 \log(1/\theta_{\min}),
\end{aligned}$$

where in the above we used the Cauchy-Schwartz inequality followed by the fact $\|\Theta_i(\mathbf{x})\|_1 = 1$, $\forall \mathbf{x} \in \mathcal{X}$. □

Proof of Lemma 8 (Mutual Information bound for Gaussian). This exemplifies the case where we have closed form solutions for the joint and marginal distributions, which in this case is Gaussian, and we can compute the expected value of $\Delta(\boldsymbol{\eta}_i, \boldsymbol{\eta}_0)$. The sufficient statistics and natural parameter for the i-th conditional distribution are given as follows:

$$\mathbf{T}(X_i) = \frac{X_i}{\sigma/\sqrt{2}}, \qquad \boldsymbol{\eta}_i = \frac{\mu_i}{\sigma/\sqrt{2}}.$$

Also note that, $\forall i \in [m]$, the marginal expectation $\mathbb{E}[X_i] \leq \mu \|\mathbf{w}_i\|_2$. Therefore, we have that $\mathbb{E}_{\mathsf{X}_{\pi_i}}[\Delta(\boldsymbol{\eta}_i, \boldsymbol{\eta}_0)] = \mathbb{E}_{\mathsf{X}_{\pi_i}}\left[2(\mu_i - \mu)^2/\sigma^2\right] = 2(\mathbb{E}_{\mathsf{X}_{\pi_i}}[(\mu_i - \mu)]^2 + \mathrm{Var}_{\mathsf{X}_{\pi_i}}[\mu_i])/\sigma^2 \leq 2(\mu^2(\|\mathbf{w}_i\|_2 - 1)^2 + \mathrm{Var}_{\mathsf{X}_{\pi_i}}[\mu_i])/\sigma^2$. Hence, we need to upper bound $\mathrm{Var}_{\mathsf{X}_{\pi_i}}[\mu_i]$ in order to upper bound $\mathbb{E}_{\mathsf{X}_{\pi_i}}[\Delta(.)]$. Let $(i)_G \in [m]$ be the i-th node in the topological order defined by the graph G. We use the shorthand notation (i), when it is clear from context that the i-th node in the topological ordering is intended. Now, from the properties of the Gaussian distribution we know that if the conditional distributions are all Gaussian, then the joint distribution over any subset of X is Gaussian as well. Let $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times m}$ be the covariance matrix for the joint distribution over X, and similarly $\boldsymbol{\Sigma}_{(i)} \in \mathbb{R}^{i \times i}$ denote the covariance matrix for the joint distribution over variables $\{X_{(1)}, \ldots, X_{(i)}\}$. Let $\overline{\mathbf{w}}_{(i)} \in \mathbb{R}^{i-1}$ be the weight vector defined as follows:

$$\forall j \in [i-1],\ (\overline{w}_{(i)})_j = \begin{cases} 0 & \text{if } j \notin \pi_{(i)}, \\ (w_{(i)})_j & \text{otherwise.} \end{cases}$$

Note that $\left\|\overline{\mathbf{w}}_{(i)}\right\|_2 = \left\|\mathbf{w}_{(i)}\right\|_2 \leq 1/\sqrt{2(i-1)}$. Then, we have that $\mathrm{Var}\left[\mu_{(i)}\right] = \overline{\mathbf{w}}_{(i)_G}^T \boldsymbol{\Sigma}_{(i-1)_G} \overline{\mathbf{w}}_{(i)_G}$ and $\mathrm{Var}\left[X_{(i)}\right] = \mathrm{Var}\left[\mu_{(i)}\right] + \sigma^2/2$. Also, for any $j \in [i-1]$, we have that $\mathrm{Cov}\left[X_{(i)} X_{(j)}\right] = \overline{\mathbf{w}}_{(i)_G}^T (\boldsymbol{\Sigma}_{(i-1)_G})_{*,j}$, where $(\boldsymbol{\Sigma}_{(i-1)_G})_{*,j}$ is the j-th column of the matrix $\boldsymbol{\Sigma}_{(i-1)}$. Therefore, the covariance matrix $\boldsymbol{\Sigma}_{(i)_G}$ can be written as follows:

$$\boldsymbol{\Sigma}_{(i)_G} = \begin{bmatrix} \boldsymbol{\Sigma}_{(i-1)_G} & \boldsymbol{\Sigma}_{(i-1)_G} \overline{\mathbf{w}}_{(i)_G} \\ \overline{\mathbf{w}}_{(i)_G}^T \boldsymbol{\Sigma}_{(i-1)_G} & \overline{\mathbf{w}}_{(i)_G}^T \boldsymbol{\Sigma}_{(i-1)_G} \overline{\mathbf{w}}_{(i)_G} + \sigma^2/2 \end{bmatrix},$$

where $\boldsymbol{\Sigma}_{(1)} \in \mathbb{R}^{1 \times 1} = [[\sigma^2/2]]$. Since $\boldsymbol{\Sigma}_{(i)}$ is positive definite, we have that $\lambda_{\max}(\boldsymbol{\Sigma}_{(i)}) \leq \lambda_{\max}(\boldsymbol{\Sigma}_{(i-1)}) + \overline{\mathbf{w}}_{(i)}^T \boldsymbol{\Sigma}_{(i-1)} \overline{\mathbf{w}}_{(i)} + \sigma^2/2$. Next, we prove, by induction, that $\lambda_{\max}(\boldsymbol{\Sigma}_{(i)}) \leq i\sigma^2$. First, note that the base case holds, i.e., $\lambda_{\max}(\boldsymbol{\Sigma}_{(1)}) = \sigma^2/2 \leq \sigma^2$. Now assume, that $\lambda_{\max}(\boldsymbol{\Sigma}_{(i-1)}) \leq (i-1)\sigma^2$. Then, we have:

$$\begin{aligned}
\lambda_{\max}(\boldsymbol{\Sigma}_{(i)}) &\leq \lambda_{\max}(\boldsymbol{\Sigma}_{(i-1)}) + \overline{\mathbf{w}}_{(i)}^T \boldsymbol{\Sigma}_{(i-1)} \overline{\mathbf{w}}_{(i)} + \frac{\sigma^2}{2} \\
&\leq (i-1)\sigma^2 + \left\|\overline{\mathbf{w}}_{(i)}\right\|_2^2 (i-1)\sigma^2 + \frac{\sigma^2}{2} \\
&\leq (i-1)\sigma^2 + \frac{1}{2(i-1)}(i-1)\sigma^2 + \frac{\sigma^2}{2}
\end{aligned}$$

$$\leq i\sigma^2.$$

Therefore, we can bound the variance of μ_i as follows:

$$\begin{aligned}\text{Var}\left[\mu_{(i)}\right] &= \overline{\mathbf{w}}_{(i)}^T \boldsymbol{\Sigma}_{(i-1)} \overline{\mathbf{w}}_{(i)} \leq \left\|\overline{\mathbf{w}}_{(i)}\right\|_2^2 \lambda_{\max}(\boldsymbol{\Sigma}_{(i-1)}) \\ &\leq \frac{1}{2(i-1)}(i-1)\sigma^2 = \frac{\sigma^2}{2}.\end{aligned}$$

Thus, we have that $\text{Var}\left[\mu_i\right] \leq {}^{\sigma^2}/_2$, $\mathbb{E}_{\mathsf{X}_{\pi_i}}\left[\Delta(\boldsymbol{\eta}_i, \boldsymbol{\eta}_0)\right] \leq 1 + {}^{2\mu^2(\|\mathbf{w}_i\|_2 - 1)^2}/_{\sigma^2}$, and $\Delta_{\max} \leq 1 + {}^{(2\mu_{\max}^2(w_{\max}^2+1))}/_{\sigma_{\min}^2}$.

□

Proof of Lemma 9 (Mutual Information bound for Noisy-OR). The expected sufficient statistics and natural parameter for the Bernoulli distribution is given as:

$$\mathcal{T}(\boldsymbol{\eta}_i) = 1 - \theta_i, \qquad\qquad \boldsymbol{\eta}_i = \log\frac{\theta_i}{1-\theta_i}.$$

Also, $\mathcal{T}(\boldsymbol{\eta}_0) = \theta$ and $\boldsymbol{\eta}_0 = \log({}^{(1-\theta)}/_{\theta})$. Using the fact that $\theta^2 \leq \theta_i \leq \theta$, we can bound $\mathbb{E}_{\mathsf{X}_{\pi_i}}\left[\Delta(\boldsymbol{\eta}_i, \boldsymbol{\eta}_0)\right]$ as follows:

$$\begin{aligned}\Delta(\boldsymbol{\eta}_i, \boldsymbol{\eta}_0) &= \left(\log\frac{\theta_i}{1-\theta_i} - \log\frac{1-\theta}{\theta}\right)(1-\theta_i-\theta) \\ &\leq \left|\log\frac{\theta_i\theta}{(1-\theta_i)(1-\theta)}\right| \\ &\leq 2\left|\log\frac{\theta}{1-\theta}\right|\end{aligned}$$

Therefore, we have that $\Delta_{\max} \leq 2\left|\log({}^{\hat{\theta}}/_{(1-\hat{\theta})})\right|$.

□

Proof of Lemma 10 (MI bound for Logistic regression networks). The expected sufficient statistics and the natural parameter are given as follows:

$$\mathcal{T}(\boldsymbol{\eta}_i) = \sigma(\mathbf{w}_i^T \mathsf{X}_{\pi_i}), \qquad\qquad \boldsymbol{\eta}_i = \log\frac{\sigma(\mathbf{w}_i^T\mathsf{X}_{\pi_i})}{1-\sigma(\mathbf{w}_i^T\mathsf{X}_{\pi_i})} = \mathbf{w}_i^T\mathsf{X}_{\pi_i}.$$

From the above, we also have that $\mathcal{T}(\boldsymbol{\eta}_0) = {}^1/_2$ and $\boldsymbol{\eta}_0 = 0$. Then, $\Delta_{\max}$ is bounded as follows:

$$\begin{aligned}\Delta_{\max} &= \mathbb{E}_{\mathsf{X}_{\pi_i}}\left[\mathbf{w}_i^T\mathsf{X}_{\pi_i}(\sigma(\mathbf{w}_i^T\mathsf{X}_{\pi_i}) - {}^1/_2)\right] \\ &\leq \frac{1}{2}\mathbb{E}_{\mathsf{X}_{\pi_i}}\left[\mathbf{w}_i^T\mathsf{X}_{\pi_i}\right] \leq \frac{1}{2}\mathbb{E}_{\mathsf{X}_{\pi_i}}\left[\|\mathbf{w}_i\|_1 \|\mathsf{X}_{\pi_i}\|_\infty\right] \\ &\leq \frac{\|\mathbf{w}_i\|_1}{2} \leq \frac{w_{\max}^1}{2}.\end{aligned}$$

□