Supp. file for "Inferning with High Girth Graphical Models"

May 18, 2014

This file contains proofs of the results in the main text. Throughout we use the following definitions (some repeated from the text). We make the following notational shortcuts for brevity:

- The graph underlying the model is denoted by (V, E) for vertices and edges. These are not always specified explicitly.
- The mutual information between X_i and X_j under the Ising model $p(x; \boldsymbol{\theta})$ is denoted by $I_{\boldsymbol{\theta}}(X_i; X_j)$.
- The dependence of model marginals on $\boldsymbol{\theta}$ is not always specified.
- A model is in $\mathcal{LG}_\epsilon$ if there exists an l that satisfies properties as in the definition of $\mathcal{LG}_\epsilon$ in the main text. This l of course depends on ϵ and the model itself, but we do not state this dependence directly. Furthermore, we routinely use the fact that this l will satisfy $l \leq \frac{g-1}{2}$ where g is the girth of the graph of $\mathcal{LG}_\epsilon$.

Definition 1. *$\boldsymbol{\theta} \in \mathcal{LG}_\epsilon$ if there exists an $l \in \mathbb{N}$ such that for all $i \in V$,and all $j \in nei(i)$,*

$$1 - \epsilon < \frac{p(x_i, x_j | x_{\partial B_l(ij)}; \boldsymbol{\theta})}{p(x_i, x_j; \boldsymbol{\theta})} < 1 + \epsilon \tag{0.1}$$

and $B_l(ij)$ is a tree.

Definition 2. *Given a model $p(x; \boldsymbol{\theta})$, we define α_{ij} for each $ij \in E$ as follows:*

$$\alpha_{ij} = p(x_i = 1, x_j = 1) - p(x_i = 1)p(x_k = 1) \tag{0.2}$$

Note that α_{ij} is a measure of the dependence between X_i and X_j (it is zero if they are independent).

In Section 6 we give some useful properties of α.

1 Proof of Lemma 1 in the main text

First, the fact that BP converges follows directly from the criterion in Tatikonda and Jordan (2002) and A_ϵ. In fact, Tatikonda and Jordan (2002) require a looser bound.[1] We require a smaller $\theta_{\max}$ for the parameter consistency results shown later.

We turn to show that the marginals $\tau_{ij}(x_i, x_j)$ obtained as the BP fixed points are within ϵ^2 of the true marginals.

Lemma 1. *Assume the model $(G, \boldsymbol{\theta})$ satisfies A_ϵ. Then running belief propagation on the model will converge to marginals $\tau_{ij}(x_i, x_j)$ such that for all $ij \in E$:*

$$|p(x_i, x_j; \boldsymbol{\theta}) - \tau_{ij}(x_i, x_j)| \leq \epsilon^2 \tag{1.1}$$

Proof: Since BP converges the $\boldsymbol{\tau}$ marginals are well defined. To calculate the conditional marginals $p(x_i, x_j | x_{\partial B_l(ij)})$, BP can be run on the tree $B_l(ij)$ with the given assignment for $x_{\partial B_l(ij)}$. The resulting conditional marginal will be exact since BP is exact on trees. The lemma then follows from applying Eq. 2.1 to all $x_{\partial B_l(ij)}$ assignments, taking their convex combination (according to the BP marginals on $x_{\partial B_l(ij)}$) and using the triangle inequality. ∎

[1] The bound in Tatikonda and Jordan (2002) is $\max_{i \in V} \sum_{j \in nei(i)} J_{ij} < 1$.

2 Proof of Lemma 2 in the main text

Lemma 2. *If $(G, \boldsymbol{\theta})$ satisfies the A_ϵ assumptions above, then the model is in $\mathcal{LG}_\epsilon$.*

Proof: Following the proof from Mossel and Sly (2008, Lemma 2.8) and Anandkumar et al. (2012, Proposition 1 in supp. file) with small modifications we have:

$$|p(x_i, x_j|x_{\partial B_l(ij)}) - p(x_i, x_j)| \leq \epsilon^2 \tag{2.1}$$

By Lemma 11 we have $p(x_i, x_j) \geq \epsilon$ and we conclude:

$$\begin{aligned} \frac{p(x_i, x_j|x_{\partial B_l(ij)})}{p(x_i, x_j)} - 1 &\leq \frac{\epsilon^2}{p(x_i, x_j)} \\ \frac{p(x_i, x_j|x_{\partial B_l(ij)})}{p(x_i, x_j)} &\leq 1 + \frac{\epsilon^2}{p(x_i, x_j)} \leq 1 + \epsilon \end{aligned}$$

The other direction follows in the same way. ∎

Corollary 1. *If $(G, \boldsymbol{\theta})$ satisfies the A_ϵ assumptions above, then*

$$1 - \epsilon \leq 1 - \frac{\epsilon^2}{p(x_i)} \leq \frac{p(x_i|x_{\partial B_l(ij)})}{p(x_i)} \leq 1 + \frac{\epsilon^2}{p(x_i)} \leq 1 + \epsilon \tag{2.2}$$

$$1 - 2\epsilon \leq \frac{1 - \epsilon^2}{1 + \epsilon^2} \leq \frac{p(x_i|x_j, x_{\partial B_l(ij)})}{p(x_i|x_j)} \leq \frac{1 + \epsilon}{1 - \epsilon} \leq 1 + 3\epsilon \tag{2.3}$$

Proof: Eq. (2.2) can be derived in exactly the same way as Lemma 2. This can be used to prove Eq. 2.3 as follows:

$$\begin{aligned} \frac{p(x_i|x_j, x_{\partial B_l(ij)})}{p(x_i|x_j)} &= \frac{p(x_i|x_j, x_{\partial B_l(ij)})p(x_j|x_{\partial B_l(ij)})}{p(x_i|x_j)p(x_j|x_{\partial B_l(ij)})} \\ &\leq \frac{p(x_i|x_j, x_{\partial B_l(ij)})p(x_j|x_{\partial B_l(ij)})}{p(x_i|x_j)p(x_j)(1-\epsilon)} = \frac{p(x_i, x_j|x_{\partial B_l(ij)})}{p(x_i, x_j)(1-\epsilon)} \\ &\leq \frac{1+\epsilon}{1-\epsilon} \leq 1 + 3\epsilon \end{aligned}$$

The last inequality is due to $\epsilon < \frac{1}{3}$. ∎

3 Proof of Structural Consistency

In what follows we always assume that assumptions A_ϵ holds, and hence $\boldsymbol{\theta} \in \mathcal{LG}_\epsilon$.

3.1 Proof of Lemma 3 in the main text

Lemma 3. *Assume $(G, \boldsymbol{\theta})$ satisfies A_ϵ, for all $ij \in E$ and $k \notin B_l(ij)$ it holds that $I_{\boldsymbol{\theta}}(X_i; X_k) \leq \epsilon^2$.*

Proof: We can bound $p(x_i, x_k)$ as follows,

$$p(x_i, x_k) = \sum_{x_{\partial B_l(ij)}} p(x_i, x_k, x_{\partial B_l(ij)}) = \sum_{x_{\partial B_l(ij)}} p(x_i|x_{\partial B_l(ij)})p(x_k, x_{\partial B_l(ij)}) \leq (1 + \frac{\epsilon^2}{p(x_i)})p(x_i)p(x_k)$$

where the second equality is a result of $x_{\partial B_l(ij)}$ separating x_i and x_k in the graph, and the last inequality is Eq. (2.2).

Plugging this into the expression for mutual information, and using $\log(1 + \epsilon) \leq \epsilon$ gives the desired result:

$$\begin{aligned} I_{\boldsymbol{\theta}}(X_i; X_k) &= \sum_{x_i, x_k} p(x_i, x_k) \log \frac{p(x_i, x_k)}{p(x_i)p(x_k)} \leq \sum_{x_i, x_k} p(x_i, x_k) \log \frac{(1 + \frac{\epsilon^2}{p(x_i)})p(x_i)p(x_k)}{p(x_i)p(x_k)} \\ &= \sum_{x_i} p(x_i) \log \left(1 + \frac{\epsilon^2}{p(x_i)}\right) \leq \epsilon^2 \end{aligned}$$

3.2 Proof of Lemma 4 in the main text

The following is a key result in facilitating our greedy ECL procedure. It says that the information between any two disconnected variables X_i and X_j is strictly smaller than the information between any two variables on the path between X_i and X_j. This will later imply (see Theorem 1) that the greedy algorithm will not choose erroneous "shortcut" edges between such X_i and X_j.

Lemma 4. *Assume $(G, \boldsymbol{\theta})$ satisfies A_ϵ. Let $ij \notin E$ be two nodes whose distance in G is $q < \lfloor \frac{g-1}{2} \rfloor$. Let $P^{ij} = \{x_i = x_{p_1}, \ldots, x_{p_q} = x_j\}$ be a shortest path in the graph between i and j. Then,*

$$I_{\boldsymbol{\theta}}(X_i; X_j) + 3\epsilon < I_{\boldsymbol{\theta}}(X_{p_s}; X_{p_{s+1}}) \ \forall 1 \leq s \leq q-1 \tag{3.1}$$

For proving Lemma 4 we will need the following two lemmas. The first lemma will give a bound on the difference between a three variables Markov chain to three variables in a large girth model.

Lemma 5. *Let $k \in nei(i)$ be on the shortest path between i and j where $j \in B_l(i,k)$ in the graph G. Consider a new distribution on the three variables X_i, X_j, X_k defined as follows:*[2]

$$p_M(x_i, x_j, x_k) = p(x_i)p(x_k|x_i)p(x_j|x_k) \tag{3.2}$$

Denote by $I(X_i; X_j), I_M(X_i; X_j)$ the mutual informations under the models p and p_M respectively. Then it holds that:

$$|I(X_i; X_j) - I_M(X_i; X_j)| < 12\epsilon \tag{3.3}$$

Proof: We first relate $p(x_j|x_i)$ to $p_M(x_j|x_i)$ as follows:

$$\begin{aligned}
p(x_j|x_i) &= \frac{\sum_{x_k, x_{\partial B_l(ik)}} p(x_i, x_k, x_{\partial B_l(ik)}, x_j)}{p(x_i)} \\
&= \frac{\sum_{x_k, x_{\partial B_l(ik)}} p(x_k)p(x_{\partial B_l(ik)}|x_k)p(x_i|x_{\partial B_l(ik)}, x_k)p(x_j|x_k, x_{\partial B_l(ik)}, x_i)}{p(x_i)} \\
&\leq \frac{\sum_{x_k, x_{\partial B_l(ik)}} p(x_k)p(x_{\partial B_l(ik)}|x_k)p(x_i|x_k)(1+3\epsilon)p(x_j|x_k, x_{\partial B_l(ik)})}{p(x_i)} \qquad (3.4) \\
&= \frac{\sum_{x_k, x_{\partial B_l(ik)}} p(x_i, x_k)(1+3\epsilon)p(x_j, x_{\partial B_l(ik)}|x_k)}{p(x_i)} \\
&= \sum_{x_k} p(x_k|x_i)(1+3\epsilon)p(x_j|x_k) \\
&= (1+3\epsilon)p_M(x_j|x_i) \qquad (3.5)
\end{aligned}$$

where Eq. (3.4) is due to Eq. 2.3.

Similarly it can be shown that $(1-2\epsilon)p_M(x_j|x_i) \leq p(x_j|x_i)$ and $p_M(x_j)(1-2\epsilon) \leq p(x_j) \leq p_M(x_j)(1+3\epsilon)$. We can now write:

$$-H(X_j|X_i) = \sum_{x_i, x_j} p(x_i, x_j) \log p(x_j|x_i) \geq \log(1-2\epsilon) + \sum_{x_i, x_j} p(x_i, x_j) \log p_M(x_j|x_i)$$

Now because $\log p_M(x_j|x_i)$ is non-positive, and $p(x_i) = p_M(x_i)$ we have:

$$\begin{aligned}
-H(X_j|X_i) &\geq \log(1-2\epsilon) + (1+3\epsilon) \sum_{x_i} p(x_i)p_M(x_j|x_i) \log p_M(x_j|x_i) \\
&= \log(1-2\epsilon) - (1+3\epsilon)H_M(X_j|X_i) \\
&\geq -H_M(X_j|X_i) - 3\epsilon H_M(X_j|X_i) + \log(1-2\epsilon)
\end{aligned}$$

[2] In other words, p_M corresponds to the Markov chain $X_i \to X_k \to X_j$, with pairwise distributions inherited from $p(x; \boldsymbol{\theta})$

Next, use the fact that for $0 \leq \epsilon \leq 0.1$ it holds that $\log(1-2\epsilon) \geq -3\epsilon$, and the fact that conditioning reduces entropy to get:

$$-H(X_j|X_i) \geq -H_M(X_j|X_i) - \epsilon\left(3 + 3H_M(X_j)\right) \tag{3.6}$$

Similarly it can be shown that $H(X_j) \geq H_M(X_j) - \epsilon(3 + 3H_M(X_j))$. The result then follows by combining the above and using the fact that binary entropies are upper bounded by 1. ∎

We will now give a general lemma on Markov models which gives a bound on the difference between the information of edges and the information between non-edges.

Lemma 6. *Let $X_1 - X_2 - X_3$ be a Markov chain with binary variables. Assume that $0 < \mu_i < 1$ for $i = 1, 2, 3$. Then for $k = 1, 3$ it holds that:*

$$\left(\frac{\mu_2(1-\mu_2)}{|\alpha_{k2}|}\right) I(X_1; X_3) \leq I(X_2; X_k) \tag{3.7}$$

Proof: Using the definition of α_{ij} it can be shown that if $\alpha_{23} < 0$ then $\frac{\partial H(X_3|X_1)}{\partial \alpha_{23}} \leq \frac{\partial H(X_3|X_2)}{\partial \alpha_{23}} \frac{|\alpha_{12}|}{\mu_2(1-\mu_2)}$ and if $\alpha_{23} > 0$ then $\frac{\partial H(X_3|X_1)}{\partial \alpha_{23}} > \frac{\partial H(X_3|X_2)}{\partial \alpha_{23}} \frac{|\alpha_{12}|}{\mu_2(1-\mu_2)}$(the derivative is negative).

Denote $H(X_3|X_1) = f(\alpha_{23})$, $H(X_3|X_2) = g(\alpha_{23})$ and $a = \frac{|\alpha_{12}|}{\mu_2(1-\mu_2)}$. Note that $f(0) = g(0) = H(X_3)$. By the fundamental theorem of calculus we have $f(x) - f(0) = \int_0^x f'(x)$ hence for $\alpha_{23} > 0$:

$$f(\alpha_{23}) - f(0) \quad = \quad \int_0^{\alpha_{23}} f'(y)dy \geq \int_0^{\alpha_{23}} a g'(y)dy = a \int_0^{\alpha_{23}} g'(y)dy$$

Since $g(x) - g(0) = \int_0^x g'(x)$ we have

$$\begin{aligned} f(\alpha_{23}) - f(0) &\geq a\left(g(\alpha_{23}) - g(0)\right) \\ \frac{1}{a}(f(\alpha_{23}) - f(0)) &\geq g(\alpha_{23}) - g(0) \\ \frac{1}{a} I(X_1; X_3) &\leq I(X_2; X_3) \end{aligned}$$

The other results follow similarly. ∎

Corollary 2. *If the condition of Lemma 6 holds and $I(X_1, X_3) > x$ then*

$$I(X_1, X_3) \leq I(X_1, X_2) - x\left(\frac{\mu_2(1-\mu_2)}{|\alpha_{k2}|} - 1\right) \tag{3.8}$$

Proof: From Lemma 6 we have

$$\begin{aligned} I(X_1, X_2) &\geq \left(\frac{\mu_2(1-\mu_2)}{|\alpha_{k2}|}\right) I(X_1, X_3) = I(X_1, X_3) + \left(\frac{\mu_2(1-\mu_2)}{|\alpha_{k2}|} - 1\right) I(X_1, X_3) \\ &\geq I(X_1, X_3) + \left(\frac{\mu_2(1-\mu_2)}{|\alpha_{k2}|} - 1\right) x \end{aligned}$$

switching sides and we have the result. ∎

Using Lemma 5 and Corollary 2 above we will prove Lemma 4 in the main text.

Proof: First note that the conditions of Lemma 5 hold for all i,j with distance less then $\lfloor \frac{g-1}{2} \rfloor$. Second, by assumption $I(X_i, X_j) > 13\epsilon$ and so using Lemma 5 we can conclude $I_M(X_i, X_j) > \epsilon$. Next we prove the desired result via induction on q, the length of the path.

For the base case $q = 3$, by Lemma 5 we have $I(X_{p_1}; X_{p_3}) \leq I_M(X_{p_1}; X_{p_3}) + 12\epsilon$. Note that α_{12} for both probabilities is equal, and we can use Lemma 10 to bound α_{12}. Now using Corollary 2 and the bound on $I_M(X_{p_1}, X_{p3})$ we have $I_M(X_{p_1}; X_{p_3}) \leq I_M(X_{p_1}; X_{p_2}) - 15\epsilon$. So we can conclude $I(X_{p_1}; X_{p_3}) \leq I_M(X_{p_1}; X_{p_2}) - 3\epsilon$. Since $I(X_{p_1}; X_{p_2}) = I_M(X_{p_1}; X_{p_2})$ we have $I(X_{p_1}; X_{p_3}) \leq I(X_{p_1}; X_{p_2}) - 3\epsilon$.

For the induction step, assume that $I(X_i, X_j) + 3\epsilon \leq I(X_{p_i}, X_{p_{i+1}})$ where $2 < i \leq q-1$ for all edges ij of length $q-1$. We next prove the result for length q. Note, that $X_{p_1} - X_{p_2} - X_{p_q}$ fulfill the conditions of Lemma 6. So

if we prove that $|\alpha_{2,q}| \leq |\alpha_{2,3}|$ by exactly the same argument as above we have $I(X_{p_1}, X_{p_q}) + 3\epsilon \leq I(X_{p_1}, I_{p_2})$ and $I(X_{p_1}, X_{p_q}) + 3\epsilon \leq I(X_{p_2}, I_{p_q})$ but since the distance between p_2 and p_q is $q-1$ we can use the induction assumption to complete the proof.

We are left to prove $|\alpha_{2,q}| \leq |\alpha_{2,3}|$. The conditional $p(x_{p_q}|x_{p_2})$ can be treated exactly as in Eq. 3.5 resulting in $p(x_{p_q}|x_{p_2})p(x_{p_2}) \leq (1+3\epsilon)\sum_{x_{p_3}} p(x_{p_3}|x_{p_2})p(x_{p_q}|x_{p_3})$. The same argument can be repeated for $x_{p_4}, \ldots x_{q-1}$. so we have

$$p(x_{p_q}|x_{p_2})p(x_{p_2}) \leq (1+3\epsilon)^{q-3} \sum_{x_{p_3},\ldots x_{p_{q-1}}} \prod_{s=3}^{q-1} p(x_{p_{s+1}}|x_{p_s})p(x_{p_3}|x_{p_2})p(x_{p_2}) = (1+3\epsilon)^{q-3} p_M(x_{p_2}, x_{p_q}) \tag{3.9}$$

Looking at how α is calculated in Markov chain, one can see that each factor $p(x_{p_{s+1}}|x_{p_s})$ contributes to α_{p_2,p_q} a factor of $\frac{\alpha_{s+1,s}}{\mu_s(1-\mu_s)}$ so we can write:

$$|\alpha_{p_2,p_q}| \leq (1+\epsilon)^{q-3} \prod_{s=3}^{q-1} \frac{|\alpha_{p_s,p_{s+1}}|}{\mu_{p_s}(1-\mu_{p_s})} |\alpha_{p_2,p_3}| \leq |\alpha_{p_2,p3}| \tag{3.10}$$

where the last inequality is by Lemma 10 we can conclude $\frac{\mu_{p_s}(1-\mu_{p_s})}{|\alpha_{p_s,p_{s+1}}|} > 1+3\epsilon$. ∎

3.3 Proof of Theorem 1 in the main text - Structure Consistency for Infinite Data

The next theorem deals with the infinite data case, where the empirical mutual information is equal to the true information $I_{\boldsymbol{\theta}}(X_i; X_j)$.

Theorem 1. *Let $(G^*, \boldsymbol{\theta}^*)$ be an Ising model satisfying assumptions A_ϵ. Denote the model graph by $G = (V^*, E^*)$. Then running ECL with mutual informations $I_{\boldsymbol{\theta}^*}(X_i; X_j)$ calculated from $p(x; \boldsymbol{\theta}^*)$ and the true girth of G, will result in a set of edges E such that:*

- *If $ij \in E^*$ then $ij \in E$. Namely, E contains all edges in E^*.*
- *If $ij \in E \setminus E^*$ then $I_{\boldsymbol{\theta}^*}(X_i; X_j) \leq 13\epsilon$. Namely, E contains no redundant edges except possibly those with mutual information less than 13ϵ.*

Proof: We prove by induction on n, the number of edges in E after the k^{th} step of the ECL procedure. For $n=1$ we will need to prove that the pair ij with the maximum information $I_{\boldsymbol{\theta}}(X_i; X_j)$[3] satisfies $ij \in E^*$. We will prove by contradiction. By the assumption of contradiction $I_{\boldsymbol{\theta}}(X_i; X_j) > \epsilon$, and hence by Lemma 3 we conclude that $j \in B_l(ik)$. Let k be a neighbor of i in the path to j ($j \neq k$ by contradiction). But by Lemma 4 we know that $I(X_i; X_k) > I(X_i; X_j)$ in contradiction to the optimality of $I(X_i, X_j)$.

By the induction assumption we assume that all edges in E with information greater or equal to the information on the $n-1$ edge are in the graph. If the information of the n^{th} edge is less than 13ϵ the result follows. Otherwise let ij be the next edge to be added.

By the algorithm definition ij is a legal edge (i.e., there is no path in E shorter than $g-1$ between i, j) with the maximal information from all legal edges not in E. Now assume $ij \notin E^*$. From Lemma 3 and the fact that $I(X_i; X_j) > 13\epsilon$ we can conclude that $j \in B_l(ik)$. Hence, there is a path in E^* with length q smaller then $q \leq \lfloor \frac{g-1}{2} \rfloor$. The edge ij satisfies the condition of Lemma 4, and hence we have $I(X_i; X_j) + 3\epsilon \leq I(X_{p_s}; X_{p_{s+1}})$ for all $s = 1, \ldots, q-1$ where X_{p_s} are the vertices along the shortest path between i and j. So all edges in the path have information greater then $I(X_i, X_j)$. But note that one edge in the path is not in E from the legality of ij. We reach a contradiction to the fact that ij is the edge with maximal information from all legal edges not in E.
∎

3.4 Proof of Theorem 2 in the main text - Structure Consistency for Finite Samples

Here we consider the case where the mutual informations used by ECL are calculated from a finite sample. Intuitively, as more data becomes available the information estimates should improve and we should be able to find the correct structure with high probability. The following theorem quantifies the number of samples.

[3] In other words $I_{\boldsymbol{\theta}}(X_i; X_j) \geq I_{\boldsymbol{\theta}}(X_k; X_l)\ \forall k, l \in V$ (including of course pairs that are not in E^*).

Theorem 2. *Assume an IID sample of size n is generated from a model $(G^*, \boldsymbol{\theta}^*)$ satisfying A_ϵ. For any $\delta > 0$ let n satisfy $n > N_0$ where:*

$$N_0 = \frac{C_0 \log p \log \frac{1}{\delta}}{\epsilon^3} \tag{3.11}$$

and C_0 is a constant. Then with probability greater than $1 - \delta$ ECL will recover E^* as in Theorem 1.

Proof: First, recall that the Chernoff bound (Mitzenmacher and Upfal, 2005, Theorem 4.4,4.5) with relative error can be written as:

$$P\left((1-\gamma) < \frac{\bar{\mu}_{ij}(x_i, x_j)}{\mu_{ij}(x_i, x_j)} < (1+\gamma)\right) \leq \exp\{\frac{-\gamma^2 \mu_{ij}(x_i, x_j) n}{3}\} \tag{3.12}$$

We start by bounding the empirical error of the mutual information. Denote by $\bar{\mu}$ the empirical marginals, and μ the true marginals. Given a γ such that $(1-\gamma) \leq \frac{\bar{\mu}_{ij}(x_i,x_j)}{\mu_{ij}(x_i,x_j)} \leq (1+\gamma)$ we can write:

$$\begin{aligned}
I_{\bar{\boldsymbol{\mu}}}(X_i; X_j) &= \sum_{x_i, x_j} \bar{\mu}_{ij}(x_i, x_j) \log \frac{\bar{\mu}_{ij}(x_i, x_j)}{\bar{\mu}_i(x_i)\bar{\mu}_j(x_j)} \\
&\leq \sum_{x_i, x_j} \bar{\mu}_{ij}(x_i, x_j) \log \frac{\mu_{ij}(x_i, x_j)(1+\gamma)}{\mu_i(x_i)\mu_j(x_j)(1-\gamma)^2} \\
&\leq \sum_{x_i, x_j} \bar{\mu}_{ij}(x_i, x_j) \log \frac{\mu_{ij}(x_i, x_j)}{\mu_i(x_i)\mu_j(x_j)} + 4\gamma \\
&\leq \sum_{x_i, x_j} \mu_{ij}(x_i, x_j)(1+\gamma) \log \frac{\mu_{ij}(x_i, x_j)}{\mu_i(x_i)\mu_j(x_j)} + 4\gamma \\
&\leq I_{\boldsymbol{\mu}}(X_i; X_j)(1+\gamma) + 4\gamma
\end{aligned}$$

Using the fact that the information in the binary case is upper bounded by 1 we conclude that $|I_{\bar{\boldsymbol{\mu}}_{ij}}(X_i; X_j) - I_{\boldsymbol{\mu}}(X_i; X_j)| < 5\gamma$.

From Theorem 1 we know that as long as the relative order of the mutual informations is preserved, the algorithm will recover the correct E. Now by Lemma 4 we know that $I(X_i; X_j) + 3\epsilon < I(X_{p_s}; X_{p_{s+1}})$ for all $1 \leq s \leq q-1$ so if the mistake is less than ϵ the relative order of the informations will be preserved with high probability. Thus the ECL algorithm will recover the true graph (as in Theorem 1) despite the noisy marginals.

To ensure that the statistical error in the information is smaller than ϵ, we need to have at most $\gamma = \frac{\epsilon}{5}$. By Eq. (3.12)

$$P\left(\frac{\bar{\mu}_{ij}(x_i, x_j)}{\mu_{ij}(x_i, x_j)} > (1 - \frac{\epsilon}{5})\right) \leq \exp\left\{\frac{-\epsilon^3 n}{75}\right\} \leq \delta \tag{3.13}$$

where we use Lemma 11 for replacing μ with ϵ.

The above guaranteed ϵ accurate mutual informations for a single marginal. To get the bound for all edges and vertices a union bound can be used resulting in:

$$N_0 > \frac{a \log(p) \log \frac{1}{\delta}}{\epsilon^3} \tag{3.14}$$

for some constant a and the result follows.

∎

4 Parameter Consistency

Here we prove that the parameters learned by ECL achieve a likelihood that is ϵ close to optimal.

The result relies on the connection between the true partition function $Z(\boldsymbol{\theta})$ and its approximation calculated using the Bethe variation approximation $Z_B(\boldsymbol{\theta})$. Recall that the Bethe approximation of the partition function is given by (e.g., see Heinemann and Globerson, 2011):

$$\log Z_B(\boldsymbol{\theta}) = \max_{\boldsymbol{\mu} \in \mathcal{M}_L} \boldsymbol{\mu} \cdot \boldsymbol{\theta} + H_B(\boldsymbol{\mu}) \tag{4.1}$$

where $\mathcal{M}_L$ is the *local marginal polytope* which is the set of consistent pairwise marginals, and $H_B(\boldsymbol{\theta})$ is the Bethe entropy given by:

$$H_B(\boldsymbol{\mu}) = \sum_i H(X_i) - \sum_{ij \in E} I(X_i; X_j) \tag{4.2}$$

The entropy and informations above are calculated using the singleton and pairwise marginals in $\boldsymbol{\mu}$. Similarly we can define the *Bethe likelihood* by replacing the exact partition function with its Bethe approximation:

$$L_B(\boldsymbol{\theta}) = \boldsymbol{\mu} \cdot \boldsymbol{\theta} + \log Z_B(\boldsymbol{\theta}) \tag{4.3}$$

Lemma 7. *Assume a model with parameters $\boldsymbol{\theta}$ satisfies A_ϵ. Then*

$$|\log Z(\boldsymbol{\theta}) - \log Z_B(\boldsymbol{\theta})| < (|E| - p)\epsilon^2 \tag{4.4}$$

Proof: We know (e.g., see Sudderth et al., 2008) that $\log Z(\boldsymbol{\theta}) - \log Z(\boldsymbol{\theta}^c(\boldsymbol{\tau})) = \log Z_B(\boldsymbol{\theta})$. Where $\boldsymbol{\tau}$ are the Bethe marginals for parameters $\boldsymbol{\theta}$. Consider the case where $\boldsymbol{\theta}$ is a tree. In this case $\log Z(\boldsymbol{\theta}^c(\boldsymbol{\tau})) = 0$ and indeed the Bethe partition function is exact.

Now consider a model identical to $\boldsymbol{\theta}$ except the ij edge is removed. Denote the corresponding parameter by $\boldsymbol{\theta}^{ij}$, and its exact marginals by p^{ij}. Furthermore let $Z^{ij}(x_i, x_j)$ denote the partition function of $\boldsymbol{\theta}^{ij}$ when the variables i, j are assigned x_i, x_j. By the definition of the partition function we have

$$\begin{aligned}
Z(\boldsymbol{\theta}^c(\boldsymbol{\tau})) &= \sum_{x_i, x_j} Z^{ij}(x_i, x_j) \left(\frac{\tau_{ij}(x_i, x_j)}{\tau_i(x_i)\tau_j(x_j)} \right) \\
&= Z^{ij}(\boldsymbol{\theta}^c(\boldsymbol{\tau})) \sum_{x_i, x_j} p^{ij}(x_i, x_j) \left(\frac{\tau_{ij}(x_i, x_j)}{\tau_i(x_i)\tau_j(x_j)} \right) \\
&= Z^{ij}(\boldsymbol{\theta}^c(\boldsymbol{\tau})) \sum_{x_i, x_j} p^{ij}(x_i, x_j) \left(1 + \frac{(-1)^{x_i - x_j}\, \alpha_{ij}}{\tau_i(x_i)\tau_j(x_j)} \right) \\
&= Z^{ij}(\boldsymbol{\theta}^c(\boldsymbol{\tau})) \left(1 + \sum_{x_i, x_j} \frac{(-1)^{x_i - x_j}\, p^{ij}(x_i, x_j)\alpha_{ij}}{\tau_i(x_i)\tau_j(x_j)} \right)
\end{aligned}$$

Using arguments as in Section 2 it can be shown that $|p^{ij}(x_i, x_j) - \tau_i(x_i)\tau_j(x_j)| \leq 2\epsilon^2$. Plugging this into the above yields:

$$\frac{Z(\boldsymbol{\theta}^c(\boldsymbol{\tau}))}{Z^{ij}(\boldsymbol{\theta}^c(\boldsymbol{\tau}))} \leq 1 + \sum_{x_i, x_j} \frac{\epsilon^2 \alpha_{ij}}{\tau_i(x_i)\tau_j(x_j)} \leq 1 + \epsilon^2 \tag{4.5}$$

where the last inequality is due to Lemma 10.
Repeating this argument recursively for all edges, until a tree is reached, we have the result.
∎

Recall the parameters we are learning are $\boldsymbol{\theta}^c$. In order to prove approximation results, we will use the fact that $\boldsymbol{\theta}^c \in \mathcal{LG}_\epsilon$, as the following result states.

Lemma 8. *Assume a model $\boldsymbol{\theta}$ satisfies A_ϵ. Now assume that $\boldsymbol{\theta}^c$ are calculated using the marginals of $p(x; \boldsymbol{\theta})$. Then $\boldsymbol{\theta}^c \in \mathcal{LG}_\epsilon$.*

Proof: The result follows by deriving the Ising model form of $\boldsymbol{\theta}^c$, and showing that its maximal J_{ij} implies that $\boldsymbol{\theta}^c \in \mathcal{LG}_\epsilon$. ∎

4.1 Proof of Theorem 3 in the main text

We assume as in Section 3.3 that the data size goes to infinity and thus we measure the true information $I_{\boldsymbol{\theta}}(X_i; X_j)$ on all edges.

Theorem 3. *Assume a sample is generated IID from a model $(G^*, \boldsymbol{\theta}^*)$ satisfying A_ϵ where $\epsilon < \frac{1}{|E^*|-p}$. Then as $n \to \infty$ ECL will return $(G, \boldsymbol{\theta})$ such that*

$$L^*(\theta^*, E^*) - L^*(\theta, E) < 2\epsilon \tag{4.6}$$

where L^* is the generalization likelihood.

Proof: Using Lemma 8 and Section 2 we have $\boldsymbol{\theta}^c, \boldsymbol{\theta} \in \mathcal{LG}_\epsilon$. Now by Lemma 7 we can bound $|\log Z(\boldsymbol{\theta}) - \log Z_B(\boldsymbol{\theta})| < (|E^*| - p)\epsilon^2$. By assumption $\epsilon < \frac{1}{|E*|-p}$ so that $|\log Z(\boldsymbol{\theta}) - \log Z_B(\boldsymbol{\theta})| < \epsilon$ and we conclude,

$$\begin{aligned}
L^*(\boldsymbol{\theta}, E) &= \boldsymbol{\theta}_{E^*} \cdot \bar{\boldsymbol{\mu}} - \log Z(\boldsymbol{\theta}_{E^*}) \\
&\leq \boldsymbol{\theta}_{E^*} \cdot \bar{\boldsymbol{\mu}} - \log Z_B(\boldsymbol{\theta}_{E^*}) + \epsilon \\
&\leq \boldsymbol{\theta}^c_{E^*} \cdot \bar{\boldsymbol{\mu}} - \log Z_B(\boldsymbol{\theta}^c_{E^*}) + \epsilon \\
&\leq \boldsymbol{\theta}^c_{E} \cdot \bar{\boldsymbol{\mu}} - \log Z_B(\boldsymbol{\theta}^c_{E}) + \epsilon \\
&\leq \boldsymbol{\theta}^c_{E} \cdot \bar{\boldsymbol{\mu}} - \log Z(\boldsymbol{\theta}^c_{E}) + 2\epsilon \\
&= L^*(\boldsymbol{\theta}^c, E) + 2\epsilon
\end{aligned}$$

The second inequality is due to the optimality of the canonical parameters in maximizing the Bethe likelihood (see Heinemann and Globerson (2011)) when $\bar{\boldsymbol{\mu}}$ is learnable (as in our case). The third inequality is due to the consistency of ECL and the fact that additional edges only increase the Bethe likelihood and we have the result. ∎

Theorem 4. *Assume an IID sample of size n is generated from a model $(G^*, \boldsymbol{\theta}^*)$ satisfying A_ϵ where $\epsilon < \frac{1}{|E^*|-p}$. For any $\alpha > 0, \delta > 0$, let n satisfy $n > N_1$ where:*

$$N_1 = \frac{C_1 |E^*| \log \frac{1}{\delta}}{\epsilon^3 \alpha^2} \ . \tag{4.7}$$

and C_1 is a constant. Then ECL will return a model that satisfies the following with probability greater than $1-\delta$:

$$L^*(\theta^*, E^*) - L^*(\theta, E) < 3\epsilon + \alpha \ , \tag{4.8}$$

where L^* is the generalization likelihood.

Proof: We want to show that the bound on the ratio between the empirical marginals $\bar{\boldsymbol{\mu}}$ and the exact marginals $\boldsymbol{\mu}$ results in a bound on the likelihood. Simple calculation shows that if $(1 - \frac{\alpha}{32(|E^*|+p)}) \leq \frac{\bar{\mu}_{ij}(x_i,x_j)}{\mu_{ij}(x_i,x_j)} \leq (1 + \frac{\alpha}{32(|E^*|+p)})$ then $|\theta^c(\bar{\mu}_{ij}) - \theta^c(\mu_{ij})| \leq \frac{\alpha}{8(|E^*|+p)}$. This is true for all parameters (singleton and pairwise). Summing over all parameters yields $|\boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}) - \boldsymbol{\theta}^c(\boldsymbol{\mu})| \leq \frac{\alpha}{2}$. Remember that the Lipschitz constant of the likelihood of bounded parameter model is 2 (see lemma 19 in Honorio, 2012). We thus have $|L^*(\boldsymbol{\theta}^c(\boldsymbol{\mu}), E^*) - L^*(\boldsymbol{\theta}^c(\bar{\boldsymbol{\mu}}, E^*)| < \alpha$. Using Theorem 3 we see that if $(1 - \frac{\alpha}{32(|E^*|+p)}) \leq \frac{\bar{\mu}_{ij}(x_i,x_j)}{\mu_{ij}(x_i,x_j)} \leq (1 + \frac{\alpha}{32(|E^*|+p)})$ we guarantee Eq. 4.8. To obtain the desired sample complexity we now use Eq. (3.12):

$$P\left((1 - \frac{\alpha}{32(|E^*|+p)}) \leq \frac{\bar{\mu}_{ij}(x_i,x_j)}{\mu_{ij}(x_i,x_j)} \leq (1 + \frac{\alpha}{32(|E^*|+p)})\right) \leq \exp\left\{\frac{\alpha^2 \epsilon n}{C_2(|E^*|+p)^2}\right\} \leq \delta \tag{4.9}$$

were we use Lemma 11 for bounding $\bar{\mu}_{ij}(x_i, x_j)$.
Using the union bound for the structure learning part and the above, and solving for n gives the result. ∎

5 Proof of Theorem 5 in the main text - Learning Bounded Parameters

Since we are interested in bounding the Ising interaction parameters J_{ij} we need to understand how they are related to canonical parameters $\boldsymbol{\theta}^c$. Assume we have such canonical parameters calculated from marginals with given $\mu_i, \mu_j, \alpha_{ij}$. Then it can be shown that the Ising parameter J_{ij} is given by:

$$J_{ij} = \frac{1}{4}\left(\theta^c_{ij}(1,1) + \theta^c_{ij}(0,0) - \theta^c_{ij}(1,0) - \theta^c_{ij}(0,1)\right) \tag{5.1}$$

For a given value of μ_i and μ_j, we write J_{ij} as a function of α_{ij} via:

$$J_{ij} = \psi(\alpha_{ij}; \mu_i, \mu_j) = \frac{1}{4}\log\left(1 + \frac{\alpha_{ij}}{(\mu_i(1-\mu_j) - \alpha_{ij})((1-\mu_i)\mu_j - \alpha_{ij})}\right) \tag{5.2}$$

For our parameter update, we will need a function mapping from J_{ij}, μ_i, μ_j to α_{ij}. We define:

$$\phi(J_{ij}, \mu_i, \mu_j) \equiv \psi^{-1}(J_{ij}; \mu_i, \mu_j) \tag{5.3}$$

Namely $\phi(J_{ij}, \mu_i, \mu_j)$ returns an α_{ij} such that $J_{ij} = \psi(\alpha_{ij}; \mu_i, \mu_j)$. In other words, α_{ij} is such that the canonical parameters θ^c calculated from $\mu_i, \mu_j, \alpha_{ij}$ have an Ising interaction parameter of J_{ij}.

Although ϕ is not given in closed form, for any given μ_i, μ_j it is an inverse of a scalar function and can thus be evaluated numerically with high precision.

Theorem 5. *Assume we are given a set of empirical marginals $\bar{\mu}$, and a tree structured graph $G = (V, E)$. Then the parameters $\boldsymbol{\theta}^*$ that maximize the likelihood under constraints $|J_{ij}| \leq \zeta$ are as follows. Define:*

$$\begin{aligned}
\bar{\alpha}_{ij} &= \bar{\mu}_{ij}(1,1) - \bar{\mu}_i(1)\bar{\mu}_j(1) \\
\lambda^+_{ij} &= 4\max\{0, \bar{\alpha}_{ij} - \phi(\zeta, \bar{\mu}_i, \bar{\mu}_j)\} \\
\lambda^-_{ij} &= 4\max\{0, -\bar{\alpha}_{ij} + \phi(-\zeta, \bar{\mu}_i, \bar{\mu}_j)\}
\end{aligned}$$

where ϕ is the function defined in Section 5 of the supplementary file. Use these to define a new set of marginals $\tilde{\bar{\mu}}(x_i, x_j)$ given as follows:[4]

$$\begin{cases}
\bar{\mu}_{ij}(x_i, x_j) & |\alpha_{ij}| \leq \phi(\zeta, \bar{\mu}_i\bar{\mu}_j) \\
\bar{\mu}_{ij}(x_i, x_j) - \frac{1}{4}\lambda^+_{ij} & x_i = x_j \text{ and } \alpha_{ij} > \phi(\zeta, \bar{\mu}_i, \bar{\mu}_j) \\
\bar{\mu}_{ij}(x_i, x_j) + \frac{1}{4}\lambda^+_{ij} & x_i \neq x_j \text{ and } \alpha_{ij} > \phi(\zeta, \bar{\mu}_i, \bar{\mu}_j) \\
\bar{\mu}_{ij}(x_i, x_j) + \frac{1}{4}\lambda^-_{ij} & x_i = x_j \text{ and } \alpha_{ij} < \phi(-\zeta, \bar{\mu}_i, \bar{\mu}_j) \\
\bar{\mu}_{ij}(x_i, x_j) - \frac{1}{4}\lambda^-_{ij} & x_i \neq x_j \text{ and } \alpha_{ij} < \phi(-\zeta, \bar{\mu}_i, \bar{\mu}_j)
\end{cases} \tag{5.4}$$

Then $\boldsymbol{\theta}^*$ are the canonical parameters calculated from marginals $\tilde{\bar{\mu}}$, namely $\boldsymbol{\theta}^* = \boldsymbol{\theta}^c(\tilde{\bar{\mu}})$.

Proof: Given parameters $J_{ij}(x_i, x_j)$ the interaction parameter in the equivalent Ising model is given by Eq. (5.1). Therefore we would like to solve:

$$\begin{aligned}
\max \quad & L(\bar{\boldsymbol{\mu}}, \boldsymbol{\theta}) \\
\text{s.t} \quad & |\frac{1}{4}(\theta_{ij}(1,1) + \theta_{ij}(0,0) - \theta_{ij}(1,0) - \theta_{ij}(0,1))| < \zeta
\end{aligned}$$

The Lagrangian can be written as

$$\begin{aligned}
\mathcal{L}(\bar{\boldsymbol{\mu}}, \boldsymbol{\theta}, \boldsymbol{\lambda}^+, \boldsymbol{\lambda}^-) = L(\bar{\boldsymbol{\mu}}, \boldsymbol{\theta}) \quad &- \quad \sum_{ij} \lambda^+_{ij}(\frac{1}{4}(\theta_{ij}(1,1) + \theta_{ij}(0,0) - \theta_{ij}(1,0) - \theta_{ij}(0,1)) - \zeta) \\
&+ \quad \sum_{ij} \lambda^-_{ij}(\frac{1}{4}(\theta_{ij}(1,1) + \theta_{ij}(0,0) - \theta_{ij}(1,0) - \theta_{ij}(0,1)) + \zeta)
\end{aligned}$$

It can be verified that the proposed solution indeed satisfies the KKT conditions for this problem and is thus optimal.[5] Taking the derivative respect to $\theta_{ij}(x_i, x_j)$ will result in

$$\bar{\mu}_{ij}(x_i, x_j) - \tau^{\boldsymbol{\theta}}_{ij}(x_i, x_j) \mp \frac{1}{4}\lambda^+ \pm \frac{1}{4}\lambda^- \tag{5.5}$$

Where $\tau^{\boldsymbol{\theta}}$ are the Bethe marginals (we know that we have a unique solution). and $\pm$ is $+$ for $x_i = x_j$ and $-$ for $x_i \neq x_j$ and the other way around for $\mp$.

[4] Note that these will be consistent and feasible by construction.

[5] See Yang and Ravikumar (2011) for a related derivation.

When the interaction is not "too strong", namely $|\alpha_{ij}| \leq \phi(\zeta, \bar{\mu}_i, \bar{\mu}_j)$, we have $\lambda^+ = \lambda^- = 0$ and $\tau_{ij}(x_i, x_j) = \bar{\mu}_{ij}(x_i, x_j)$ so indeed the derivative equals zero. Next we consider the case $|\alpha_{ij}| > \phi(\zeta, \bar{\mu}_i, \bar{\mu}_j)$ and show that indeed the derivative is zero in the $x_i = x_j = 1$ and $\alpha_{ij} > 0$ case. The other cases follow similarly.

$$\begin{aligned} &= \bar{\mu}_{ij}(1,1) - \tau^{\boldsymbol{\theta}}_{ij}(1,1) - \frac{1}{4}\lambda^+ + \frac{1}{4}\lambda^- \\ &= \bar{\mu}_{ij}(1,1) - \tilde{\bar{\mu}}_{ij}(1,1) - \frac{1}{4}\lambda^+ \\ &= \bar{\mu}_{ij}(1,1) - \bar{\mu}_{ij}(1,1) + \frac{1}{4}\lambda^+ - \frac{1}{4}\lambda^+ \\ &= 0 \end{aligned}$$

∎

6 Auxiliary Lemmas

Our first result provides useful bounds on the α measure.

Lemma 9. *Let $p(x;\boldsymbol{\theta})$ be an Ising model with maximum interaction parameter $J_{max} = \max_{ij\in E} |J_{ij}|$. Denote the maximum magnitude of α_{ij} for this model by $\alpha_{max} = \max_{ij\in E} |\alpha_{ij}|$. Then:*

$$\alpha_{max} \leq \frac{1}{4} \tanh J_{max} \tag{6.1}$$

Proof: Following (Anandkumar and Valluvan, 2013, Fact 1 supplementary), we can write

$$\alpha_{ij} = \frac{\sinh(2J_{ij})}{2(e^{J_{ij}}\cosh(h_i + h_j) + e^{-J_{ij}}\cosh(h_i - h_j))^2}$$

The denominator is maximized when $h_i = h_j = 0$ so that

$$\begin{aligned} |\alpha_{ij}| &\leq \frac{e^{2|J_{ij}|} - e^{-2|J_{ij}|}}{4(e^{|J_{ij}|} + e^{-|J_{ij}|})^2} \\ &= \frac{e^{|J_{ij}|} - e^{-|J_{ij}|}}{4(e^{|J_{ij}|} + e^{-|J_{ij}|})} \\ &= \frac{1}{4}\tanh(|J_{ij}|) \end{aligned}$$

∎

Lemma 10. *Let $\boldsymbol{\theta}$ satisfy A_ϵ then for all $ij \in E$*

$$\alpha_{ij} < \frac{1}{32}\min\{(1-\mu_i)(1-\mu_j), \mu_i(1-\mu_j), (1-\mu_i)\mu_j, \mu_i\mu_j, \mu_i(1-\mu_i), \mu_j(1-\mu_j)\} \tag{6.2}$$

Proof: By Lemma 9 and the bound on θ_{max} we have

$$\alpha_{ij} \leq \frac{1}{4}\tanh\theta_{max} \leq \frac{1}{8d_{max}}\eta^2 \leq \frac{1}{16}\eta^2 \tag{6.3}$$

Using the definition of η and the bound on h_i, it is easy to show that $\eta \leq \mu_i \leq 1-\eta$ for all i. Insert it to the above equation and we have the result. ∎

Lemma 11. *Let $\boldsymbol{\theta}$ satisfy A_ϵ then for all $ij \in E$*

$$p(x_i, x_j) \geq \epsilon \tag{6.4}$$

Proof: Using the minimal representation for $p(x_i, x_j)$ we can write it as a function of the singleton and α.

$$p(x_i, x_j) \geq \eta^2 - \alpha \geq \frac{15}{16}\eta^2 \geq \epsilon \tag{6.5}$$

Where the second inequality is Lemma 10 and the third is by the definition of η. ∎

References

Anandkumar, A., V. Y. Tan, F. Huang, and A. S. Willsky (2012). High-dimensional structure estimation in Ising models: Local separation criterion. *The Annals of Statistics 40*(3), 1346–1375.

Anandkumar, A. and R. Valluvan (2013). Learning loopy graphical models with latent variables: Efficient methods and guarantees. *The Annals of Statistics 41*(2), 401–435.

Heinemann, U. and A. Globerson (2011). What cannot be learned with Bethe approximations. In *Proceedings of the 27th Conference Annual Conference on Uncertainty in Artificial Intelligence*, pp. 319–326. AUAI Press.

Honorio, J. (2012). Lipschitz parametrization of probabilistic graphical models. *arXiv preprint arXiv:1202.3733*.

Mitzenmacher, M. and E. Upfal (2005). *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press.

Mossel, E. and A. Sly (2008). Rapid mixing of gibbs sampling on graphs that are sparse on average. In *Proceedings of the Nineteenth Symposium on Discrete Algorithms*, Philadelphia, PA, USA, pp. 238–247. Society for Industrial and Applied Mathematics.

Sudderth, E. B., M. J. Wainwright, and A. S. Willsky (2008). Loop series and Bethe variational bounds in attractive graphical models. In *Advances in Neural Information Processing Systems 20*, pp. 1425–1432.

Tatikonda, S. C. and M. I. Jordan (2002). Loopy belief propagation and gibbs measures. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pp. 493–500. Morgan Kaufmann Publishers Inc.

Yang, E. and P. K. Ravikumar (2011). On the use of variational inference for learning discrete graphical model. In L. Getoor and T. Scheffer (Eds.), *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, New York, NY, USA, pp. 1009–1016. ACM.