Contents

1 Introduction

We summarize Guy Bresler's paper [Bre], "Efficiently learning Ising models on arbitrary graphs." This is meant to be a reading guide for the paper itself.

1.1 Graphical models

For a graph G and vertex u, let ∂u denote the neighbors in u.

Definition 1.1: Let $G = (V, E)$ be a graph. A **Markov model** on G is a distribution on $\{-1, 1\}^V$ such that

$$\mathbb{P}[x_u | x_{V \setminus \{u\}}] = \mathbb{P}[x_u | x_{\partial u}],$$

i.e., x_u is independent of all other coordinates, given the values on the neighborhood of u. The neighbors "shield" x_u from the rest of the graph.

Recall that we can characterize all Markov models as follows.

Definition 1.2 (Gibbs distribution)**:** Let V be the set of coordinates, and $\mathcal{S}$ be a family of subsets of V. Let $x \in \{0, 1\}^V$. For each $S \in \mathcal{S}$, let ϕ_S be a function $\mathbb{R}^S \to \mathbb{R}$. A **Gibbs distribution** supported on $\mathcal{S}$ is a distribution in the form

$$\mathbb{P}(x) \propto \exp\left(\sum_{S \in \mathcal{S}} \phi_S(x)\right).$$

Theorem 1.3 (Clifford-Hammersley)**:** If $\mathbb{P}(x_1, \ldots, x_n) > 0$ for every configuration $x = (x_1, \ldots, x_n) \in \{0, 1\}^V$, then there exists a family of subsets $\mathcal{S}$, each set of which is a clique in

the dependency graph G, and such that the distribution is a Gibbs distribution supported on $\mathcal{S}$.

Conversely, a Gibbs distribution supported on $\mathcal{S}$ is a Markov model whose edges are the union of the cliques $S \subseteq \mathcal{S}$.

1.2 Ising model

To make learning easier, we focus on the special case of the Ising model.

Definition 1.4: Let $G = (V, E)$ be a graph. An **Ising model** is a probability distribution on $\{-1, 1\}^V$ where for $x \in \mathbb{R}^V$,

$$\mathbb{P}_\theta(x) = \exp\left(\sum_{ij \in E} \theta_{ij} x_i x_j + \sum_{i \in V} \theta_i x_i - \Phi(\theta)\right) \tag{1}$$

where $\theta \in \mathbb{R}^{E \cup V}$ and $\Phi(\theta)$ is a factor to make the probabilities sum to 1.

An Ising model is a Gibbs distribution where the cliques are just edges, and moreover the potential functions $\phi_{\{i,j\}}$ only depend on whether $x_i = x_j$.

Note that

1. if $\theta_{ij} > 0$ it is favorable for $x_i = x_j$,
2. if $\theta_{ij} < 0$ it is favorable for $x_i \neq x_j$,
3. if $\theta_i > 0$ it is favorable for $x_i = 1$,
4. if $\theta_i < 0$ it is favorable for $x_i = -1$.

We are concerned with the following problem.

Problem: Given m samples for the Ising model, learn the graph G exactly with probability $1 - \varepsilon$.

Learning the graph exactly is called **structure learning**.

We can't hope to have a uniform bound over θ because $\theta_{ij} \neq 0$ could be arbitrarily close to 0, and the difference in probabilities would be too small to determine if an edge exists or not. Thus we need to make the assumption that the θ_{ij} are bounded away from 0. Neither do we want them to be too large: if some probabilities are much greater than others, the small ones won't be noticed until a lot of samples are take.

1. Thus we define

$$\Omega_{\alpha,\beta,\gamma}(G) = \left\{\theta \in \mathbb{R}^{E \cup V} : \forall ij \notin E, \theta_{ij} = 0, \quad \forall ij \in E, |\theta_{ij}| \in [a, b], \quad \forall i \in V, |\theta_i| \leq h\right\},$$

and ask what the dependence of n is on $\alpha, \beta, h, \varepsilon$.

2. We also restrict to G having maximum degree $\leq d$. This is reasonable because in real life, each variable typically doesn't "depend" on more than a few other variables directly.

Here is the modified problem.

Problem 1.5: Suppose $\theta \in \Omega_{\alpha,\beta,\gamma}(G)$ where G has n vertices and maximum degree $\leq d$. Given m samples from the distribution given by (1), find G with probability $1-\varepsilon$ in a reasonable time.

What is the right dependence of m and the running time on $\alpha, \beta, h, \varepsilon, d$?

Bresler's main theorem is the following.

Theorem 1.6: Using

$$m = e^{O(\alpha^{-O(1)} e^{O(\beta d^2 + hd)} \ln \frac{n}{\varepsilon})}$$

samples, Algorithm 1.7 learns G exactly with probability $\geq 1-\varepsilon$.

The running time of the algorithm is $\widetilde{O}(n^2 m)$.

We remark the following.

1. The dependence on n is $O(\ln n)$ samples and $\widetilde{O}(n^2)$ running time (where the O hides a constant depending on $\alpha, \beta, d, \varepsilon$). The previous best algorithm that works for all G, θ as in the hypothesis takes time $\Omega(n^d)$. It goes as follows: to find the neighborhood of a node, guess all subsets of size d. For each guess S, verify whether u is independent of all other nodes given the coordinates in S (up to some error parameter).

   Thus in the regime of constant d and $n \to \infty$, this algorithm is better.

2. However, the dependence on d is doubly exponential. We'll talk about why this is later.

3. The dependence on d is at least exponential by information theory (the argument is involved; see [SW12]). Thus there is a gap between this algorithm and the lower bound.

4. There have been algorithms that work under stronger assumptions, for example, the **correlation decay property** ([Gam13]) which says that the correlation between $i, j \in V$ is at most $\rho^{d(i,j)}$ for some $\rho < 1$, where $d(i,j)$ is the graph distance between i and j. This allows quadratic time because to find all neighbors, we can choose some D so that every vertex farther than D will have small correlation with u. This restricts us to at most d^D vertices. Now run the $O(n^d)$ algorithm on these vertices to get a running time of $O(n(d^D)^{d+1} \ln n)$ per node and $\widetilde{O}(n^2)$ running time overall.

We omit the constants; see the original paper. Bresler defines the chain of variables in

Theorem 4.1:

$$\begin{aligned}
\delta &= \frac{1}{2}e^{-2(\beta d+h)} \\
\tau^* &= O\left(\frac{\alpha^2\delta^{O(d)}}{d\beta}\right) \\
\varepsilon^* &= O(\tau^*) \\
\ell^* &= O\left(\frac{1}{(\tau^*)^2}\right) \\
m &= O\left(\frac{\ell^*}{(\varepsilon^*)^2\delta^{\ell^*}\ln\left(\frac{p}{\varepsilon}\right)}\right);
\end{aligned}$$

unraveling these gives the bound for m in the theorem.

We present the algorithm.

1. Key Idea: Define the influence of i on u conditioned on S, $\nu_{u|i,S}$.
2. Key Theorem: Every vertex u has a neighbor with large influence on u (when conditioned on any $S \not\supseteq \partial u$).

Algorithm 1.7: alg:gm-b The following finds the neighborhood ∂u of a vertex u. (To find the graph, repeat for all vertices.)

The algorithm has 2 parts. Given a threshold parameter τ (we set it to τ^*),

1. Psueudo-neighborhood: Find $S \supseteq \partial u$ as follows.
   (a) Let $S = \phi$.
   (b) Repeat: While there exists i with large influence $\widehat{\nu}^{\text{avg}}_{u|i,S}$,
      i. choose i with the largest influence and set $S \leftarrow\!\shortmid S \cup \{i\}$.
2. Pruning
   (a) While there exists i with $\widehat{\nu}_{u|i,S\backslash\{i\}} < \tau^*$,
      i. Remove i from S, $S \leftarrow\!\shortmid S\backslash\{i\}$.

Output S. Note we only get a superset at first because a vertex that is not a neighbor, or in fact far away in graph distance, could have large influence on u ("long-range correlation"). This happens exactly because we're not assuming the correlation decay property. There could, for example, be many paths with positive θ_{jk}'s from i to u.

We go through the following steps in the analysis.

1. Bound the conditional probability of $u+, u-$ from below in terms of α, β, h.
2. Define the notion of influence.

3. Show that there always exists a neighbor of u which has large influence on u (also true if we condition on any subset not containing ∂u). To do this, show an anti-concentration bound.

4. Argue that $|S|$ is bounded indpendent of n (it would be bad if S grows to contain all the vertices!).

5. Use Chernoff's inequality to find how large n has to be so that the estimated probabilities/influences are close to the actual. This gives the number of samples we need.

Where does the double exponential (in d) come from?

1. The first exponential comes from the fact that in order to estimate $\nu_{u|i}$ we need to estimate $\mathbb{P}(x_S)$, which is bounded below by $\delta^{-|S|}$. Thus we need $\delta^{-O(|S|)}$ samples.

2. The second exponential comes from the fact that the maximum size of the set S is bounded in terms of τ^*, so $|S| = \delta^{-O(d)}$.

First we justify that pruning works.

1. If $i \in \partial u$, then the key theorem will tell us that $\nu^{\text{avg}}_{u|i,S}$ is large, so we won't remove it.

2. If $i \notin \partial u$, then i is independent of u given $\partial u \subseteq S$, and this implies $\nu^{\text{avg}}_{u|i,S} = 0$, so we will remove it.

We don't get access to the actual influences, only to the estimated influence (from the sample), but the algorithm will work fine as long as the estimated influences are close enough $\varepsilon^* = \frac{\tau^*}{2}$. We'll address this in the last step.

2 Bounding conditional probabilities from below

We will often write indices as shorthand for coordinates; for example, $\mathbb{P}[u + |x_S]$ means $\mathbb{P}[X_u = +1 | X_S = x_S]$.

Lemma 2.1: lem:gm-p-not-too-small For any $u \in V, S \subseteq V \backslash \{u\}, x_S \in \{-1, 1\}^S$,

$$\mathbb{P}[u \pm |x_S] \geq \delta = \frac{1}{2} e^{-2(\beta d + h)}$$

For a set U,

$$\mathbb{P}[x_U | x_S] \geq \delta^{|U|}.$$

Proof. We just show the lemma for a singleton; the case for general $\mathcal{U}$ is similar.

$$\text{eq:gm-2} \mathbb{P}[u + |x_{V \backslash u}] = \frac{e^{\sum_{i \in \partial u} \theta_{ui} x_i + \theta_u}}{e^{\sum_{i \in \partial u} \theta_{ui} x_i + \theta_u} + e^{-\sum_{i \in \partial u} \theta_{ui} x_i - \theta_u}} = \frac{1}{1 + e^{-2(\sum_{i \in \partial u} \theta_{ui} x_i + \theta_u)}} \geq \delta. \quad (2)$$

□

3 Influence

First, a little motivation. The original definition of influence is as follows: for a boolean function $f : \{0,1\}^n \to \{0,1\}$, the influence is the expected change in f if we flip the kth bit, and the total influence is the sum of influences.

$$I_k(f) = \mathbb{E}_x|\Delta_k f| = \mathbb{E}_x|f(x+e_k) - f(x)|$$
$$I(f) = \sum_k I_k(f).$$

This is well-understood (ex. KKL theorem: there is a coordinate with influence $\gtrsim \frac{\ln n}{n}$, etc.).

The definition of influence is different here, though it also measures the effect of a bit flip. It measures how a conditional probability of $u+$ changes when x_i is flipped. We take the expectation with respect to a different measure, as well.

Definition 3.1: Define

$$\nu_{u|i;x_S} = \mathbb{P}[u + |i+, x_S] - \mathbb{P}[u + |i-, x_S]$$
$$\lambda_i(x_S) = 2\mathbb{P}[i + |x_S]\mathbb{P}[i - |x_S]$$
$$\nu^{\text{avg}}_{u|i;x_S} = \mathbb{E}[|\nu_{u|i;X_S}|\lambda_i(X_S)].$$

Let's focus on the special case where $S = \phi$. We will only sketch proofs when $S = \phi$ (i.e., the first step, when we're trying to find a single neighbor of u). The proofs basically carry over by changing the probabilities to probabilities conditioned on x_S. In this case,

$$\nu_{u|i} = \mathbb{P}[u + |i+] - \mathbb{P}[u + |i-]$$
$$\lambda_i = 2\mathbb{P}[i+]\mathbb{P}[i-]$$
$$\nu^{\text{avg}}_{u|i} = |\nu_{u|i;X_S}|\lambda_i(X_S).$$

Why are the weights are given by λ_i?

1. We shouldn't weight the summands in the $\mathbb{E}$ equally, because some some of the events are more likely than others. We discount the events where $\mathbb{P}[i+]$ or $\mathbb{P}[i-]$ is very small because we rarely get data from them in sampling. Think of i as "almost fixed given x_S." Thus those events shouldn't affect the influence much.

   (Note the definition of I_k is an $\mathbb{E}$ assuming the uniform distribution on $\{0,1\}^n$, which we don't have here.)

2. When we're trying to sum the $\nu_{u|i}$ over neighbors, the factor λ_i is exactly what we multiply by to give the sum a nice interpretation as an expected value we can bound using anticoncentration inequalities.

The key theorem that makes the algorithm work is the following.

Theorem 3.2: thm:gm-v-large For $\theta \in \Omega_{\alpha,\beta,\gamma}(G)$, $\partial u \not\subseteq S$, there is $i \in \partial u \backslash S$ with $\nu^{\text{avg}}_{u|i;S} \geq 2\tau^*$.

For simplicity, we'll just sketch the proof when $S = \phi$, but the proof carries over by conditioning on x_S.

To show there is $i \in \partial u$ with $\nu_{u|i}$ large, we show there is that the sum over $i \in \partial u$ is large. Let $\mathcal{U} = N(u)$. We will prove the following bound:

$$\sum_{i \in \mathcal{U}} \theta_{ui} \underbrace{\lambda_i \nu_{u|i}}_{\nu_{u|i}^{\text{avg}}} \geq \|\theta_{u\mathcal{U}}\|_1 \, 2\tau^*.$$

The fact that one summand is large will follow directly.

Why the weights θ_{ui}? Recall that $\theta_{ui} > 0$ means (roughly) that $x_i = x_u$ is more likely, so we would expect $\nu_{u|i}$ to be positive, and if $\theta_{ui} < 0$ we would expect $\nu_{u|i}$ to be negative. Thus we multiply by the θ_{ui} to try to make each summand positive (on the average).

We now calculate.

$$\begin{aligned}
\nu_{u|i} &= \mathbb{P}[u + |i+] - \mathbb{P}[u + |i-] \\
&= \sum_{x_{\mathcal{U}\setminus i}} \mathbb{P}[u + |i+, x_{\mathcal{U}\setminus i}] \mathbb{P}[x_{\mathcal{U}\setminus i}|i+] + \sum_{x_{\mathcal{U}\setminus i}} \mathbb{P}[u + |i-, x_{\mathcal{U}\setminus i}] \mathbb{P}[x_{\mathcal{U}\setminus i}|i-] \\
&= \sum_{x_{\mathcal{U}\setminus i}} \mathbb{P}[u + |i+, x_{\mathcal{U}\setminus i}] \frac{\mathbb{P}[x_{\mathcal{U}\setminus i}, i+]}{\mathbb{P}[i+]} + \sum_{x_{\mathcal{U}\setminus i}} \mathbb{P}[u + |i-, x_{\mathcal{U}\setminus i}] \frac{\mathbb{P}[x_{\mathcal{U}\setminus i}, i-]}{\mathbb{P}[i-]} \\
&= \sum_{x_{\mathcal{U}}} \mathbb{P}[u + |x_{\mathcal{U}}] \frac{x_i}{\mathbb{P}[x_i]} \mathbb{P}[x_{\mathcal{Y}}]
\end{aligned}$$

At this point, we notice that $\frac{x_i}{\mathbb{P}[x_i]}$ is ugly because it has $\mathbb{P}[x_i]$ in the denominator. It would be much nicer to work with simply the random variable x_i:

$$\begin{aligned}
\frac{x_i}{\mathbb{P}(x_i)} &\in \left\{ \frac{1}{\mathbb{P}(i+)}, \frac{1}{\mathbb{P}(i-)} \right\} \\
x_i &\in \{-1, 1\}.
\end{aligned}$$

Define the average and half-difference of $\frac{x_i}{\mathbb{P}(x_i)}$:

$$\begin{aligned}
s_i &= \frac{1}{2} \left(\frac{1}{\mathbb{P}(i+)} - \frac{1}{\mathbb{P}(i-)} \right) \\
t_i &= \frac{1}{2} \left(\frac{1}{\mathbb{P}(i+)} + \frac{1}{\mathbb{P}(i-)} \right) = \frac{1}{\lambda_i}.
\end{aligned}$$

Note that $\frac{s_i}{t_i} = \mathbb{P}(i-) - \mathbb{P}(i+) = -\mathbb{E}X_i$. Subtracting the average and dividing by the half-difference gives a ± 1 random variable:

$$\begin{aligned}
\frac{1}{t_i} \left(\frac{x_i}{\mathbb{P}(x_i)} - s_i \right) &= x_i \\
\implies \lambda_i \frac{x_i}{\mathbb{P}(x_i)} &= x_i + \frac{s_i}{t_i} = x_i - \mathbb{E}x_i.
\end{aligned}$$

This suggests multiplying by λ_i. Note $\mathbb{P}[u + |X_{\partial u}]$ is given by (2); $2\mathbb{P}[u + |X_{\partial u}]$ is given by tanh which is why we make the transformation below.

$$\begin{aligned}
\lambda_i \nu_{u|i} &= \mathbb{E}_{X_{\partial u}} \mathbb{P}[u + |X_{\partial u}](X_i - \mathbb{E}X_i) \\
&= \frac{1}{2}\mathbb{E}(2\mathbb{P}[u + |X_{\partial u}](X_i - \mathbb{E}X_i)) \\
&= \frac{1}{2}\mathbb{E}\tanh Z(X_i - \mathbb{E}X_i) \\
&\quad \text{where } Z = \sum_{i \in \partial u} \theta_{ui} X_i + \theta_u = \partial_{(u,\partial u)} \cdot X_{\partial u} + \theta_u \\
\sum_{i \in \partial u} \theta_{ui} \lambda_i \nu_{u|i} &= \frac{1}{2}(\tanh Z)\theta_{(u,\partial u)} \cdot (X_{\partial u} - \mathbb{E}X_{\partial u}) \\
&= \frac{1}{2}\mathbb{E}(\tanh Z)(Z - \mu)
\end{aligned}$$

where $\mu = \mathbb{E}Z$ and $\theta_{(u,\partial u)}$ consists of θ_{ui} for $i \in \partial u$. We want this to be bounded > 0.

The first thing to note is that for any random variable with $\mathbb{E}Z = \mu$, this quantity is ≥ 0. This makes sense because $(\tanh x)(x - \mu)$ is convex near μ, and positive except in between 0 and μ.

Thus we'll just need to show some positive amount of mass (bounded away from 0) is bounded away from μ. To do this we'll need an anticoncentration inequality.

4 Anti-concentration inequality

First, let's pretend that Z is the uniform random variable U_d on $\{-1, 1\}^d$. Then $Z = \theta_{(u,\partial u)} \cdot U + \theta_u$. The entries of U are ± 1 so we might think of Z as like a binomial random variable, which we know has some tail. The two issues are (1) Z is not uniform, and (2) using the right anticoncentration result.

For (1), Lemma 2.1 gives us that the probability of $\mathbb{P}(x_{\partial u})$ is bounded from below by δ^d (not depending on n). Thus we can write

$$X_{\partial u} = cU_{\partial u} + (1 - c)W$$

where $c = (2\delta)^d$.

Theorem 4.1 (Littlewood-Offord, Erdős)**:** thm:gm-loe Let w_i be reals with $|w_i| \geq 1$. Let I be an open interval of length k. Let $\xi \in \{-1, 1\}^d$ be iid Bernoulli. Then

$$\mathbb{P}(w \cdot \xi \in I) \leq \text{sum of } k\text{th largest binomial coefficients } \binom{d}{\bullet}.$$

In particular, for $d \geq O(1)$,

$$\mathbb{P}(w \cdot \xi \in (t_0 - 1, t_0 + 1) \leq 2\frac{1}{2^r}\binom{r}{\lfloor \frac{r}{2} \rfloor} \leq \frac{1}{2}$$

(and actually is $\precsim \frac{1}{\sqrt{d}}$).

Lemma 4.2 (Lemma 7.2 rephrased)**:** Let $\vec{\theta}$ be a vector of $d \geq O(1)$ variables with each entry $|\theta_i| \geq \alpha$. Let $\vec{Y}$ be iid ± 1. Consider the variable

$$Z = c(\vec{\theta} \cdot Y) + (1-c)W$$

where W is any random variable. Let $\mu = \mathbb{E}Z$. Then

$$\mathbb{E}[(\mu - Z)\mathbb{1}_{Z \leq \mu - \alpha\frac{c}{2}}] \geq \frac{\alpha c^2}{8}.$$

Proof. For simplicity, WLOG $\mu = 0$. Use Theorem 4.1. Let m_1 be the probabilities that $Z \leq -\alpha\frac{c}{2}$, $Z \in (-\alpha\frac{c}{2}, (2-\frac{c}{2})\alpha)$, and $Z \geq (2-\frac{c}{2})\alpha$. We just minimmize the desired quantity given the constraints

$$\begin{aligned} m_1 + m_2 + m_3 &= 1 \\ m_2 &\leq 1 - \frac{c}{2} \\ m_1\mathbb{E}[-Z|Z \leq -\frac{\alpha c}{2}] + m_2(-\alpha\frac{c}{2}) + m_3(2-\frac{c}{2})\alpha &\geq \mu, \end{aligned}$$

and also noting the desired quantity is $\geq m_1(\frac{\alpha c}{2})$. □

Apply the lemma with $Y = X_{\partial u}$ and $c = \frac{1}{2}(2\delta)^d$ to get

$$\mathbb{E}((\mu - Z)\mathbb{1}_{Z \leq \mu - \alpha\frac{c}{2}}) \geq \frac{\alpha c^2}{8} = \Omega(\alpha\delta^{O(d)}).$$

Doing the calculation with bounding tanh (See lemma 7.3 in the paper), we get the bound

$$\mathbb{E}[(\tanh Z)(Z - \mu)] \geq \frac{\alpha^2\delta^{4d+1}}{8} \geq d\beta 2\tau^*.$$

which gives the key theorem 3.2.

5 $|S|$ is bounded

We bound the influence by a KL divergence, and then by mutual information. We can bound the sum of mutual informations by 1 (chain rule), so this gives an upper bound on how many times we have to grow our set.

Using Pinskler's inequality relating the total variation distance and KL divergence,

$$d_{TV}(P, Q) \leq \sqrt{\frac{1}{2}D_{KL}(P||Q)}$$

we have

$$\begin{aligned}\frac{1}{2}v_{u|i}^{\text{avg}} &= \mathbb{P}(i+)\mathbb{P}(i-)|\mathbb{P}(u+|i+) - \mathbb{P}[u+|i-]| \\ &= |\mathbb{P}(u+,i+)(1-\mathbb{P}(i+)) - \mathbb{P}(u+,i-)\mathbb{P}(i+)| \\ &= |\mathbb{P}(u+,i+) - \mathbb{P}(u+)\mathbb{P}(i+)| \\ &= d_{TV}(\mathbb{P}(u,i), \mathbb{P}(u)\mathbb{P}(i)) \\ &\leq \sqrt{\frac{1}{2}D_K L(\mathbb{P}(u,i), \mathbb{P}(u)\mathbb{P}(i))} \\ &= \sqrt{\frac{1}{2}I(u:i)}.\end{aligned}$$

Exactly the same analysis goes through when conditioning on x_{S_l} (the first l nodes added to S). Now suppose that the estimated $\widehat{\nu}_{u|i;S_l}^{\text{avg}}$ is at most ε away from the true average. For each i that is added, we then have

$$\begin{aligned}\sqrt{\frac{1}{2}I(X_u : X_i|S_l)} &\geq \frac{1}{2}(\widehat{\nu}_{u|i;S_l}^{\text{avg}} - \varepsilon) \\ I(X_u : X_i|S_l) &\geq \frac{1}{2}(\tau - \varepsilon)^2.\end{aligned}$$

Now use the chain rule for mutual information to obtain

$$1 \geq H(X_u) \geq I(X_u|X_S) = \sum_{k=1}^{|S|} I(X_u : X_{j_k}|X_{j_1}, \ldots, X_{j_{k-1}}) \geq |S|\frac{1}{2}(\tau-\varepsilon)^2$$

so the set grows to size

$$|S| \leq \frac{2}{(\tau-\varepsilon)^2}.$$

6 Estimating ν

This is straightforward using Chernoff's inequality and getting your δs and εs in order. Let's just perform a sanity check—estimating how many samples we need.

Lemma 6.1 (Lemma 3.2 in [Bre])**:** Let $l \leq \frac{n}{4} - 2$. If the number of samples is $\geq \frac{144(l+3)}{\varepsilon^2\delta^{2l}} \ln \frac{n}{\varepsilon'}$, then

$$\mathbb{P}(\forall u, i \in V, S \subseteq V\backslash\{u,i\}, |S| \leq l, |\nu_{u|i;S}^{\text{avg}} - \widehat{\nu}_{u|i;S}^{\text{avg}}| \leq \varepsilon) \geq 1 - \varepsilon'.$$

We have to union bound over all subsets of size at most l, but this is negligible since it introduces a log factor. The key part is the error parameter we want for $\widehat{\mathbb{P}}(x_S)$ as S ranges over all subsets with $|S| \leq l+2$. If we allow $\widehat{\mathbb{P}}(x_S)$ to have ε'' error, then because terms of the form $\widehat{\mathbb{P}}(x_S)$ appears in the denominator of the sum of ν's, we get that the ν's have error $O\left(\frac{\varepsilon''}{\delta^l}\right)$ as $\delta^l \leq \delta^{|S|} \leq \mathbb{P}(x_S)$. So we need

$$O\left(\frac{\varepsilon''}{\delta^l}\right) \leq \varepsilon,$$

and on the order of $O\left(\frac{1}{(\varepsilon'')^2}\right) = O(\delta^l) = O(\delta^{|S|})$ samples, where $|S| = O\left(\frac{1}{(\tau^*)^2}\right)$. This gives the doubly exponential dependence on d.

7 Open questions

1. Can we find an algorithm with running time exponential rather than doubly exponential?

2. Extension to Markovian models: What if we have alphabet size > 2? Given $x \in \Sigma^V$, and given $f_{ij} : \Sigma^2 \to \mathbb{R}$ with f_{ij} constant for $ij \notin E$ and with range in between α and β otherwise, and $f_i : \Sigma \to \mathbb{R}$ with range $\leq \gamma$, find G.

3. Can we extend to hypergraphs? For example, in the Ising model, what if we had $x_i x_j x_k$ terms? What about the Markov model, where we allow functions f_{ijk}?

4. What is the optimal algorithm if we are only required to learn the distribution up to some statistical distance, rather than give the exact graph?

5. What if we assume a random graph? (I.e., we have a generative model for the graph.) Is the problem easy "on average"?

References

[Bre] Guy Bresler. "Efficiently learning Ising models on arbitrary graphs". In: (). arXiv: 1411.6156v2.

[Gam13] David Gamarnik. "Correlation decay method for decision, optimization, and inference in large-scale networks". In: *Tutorials in Operations Research, Vol. 10.* April 2015. 2013, pp. 108–121. ISBN: 9780984337842. DOI: http://dx.doi.org/10.1287/educ.2013.0119.

[SW12] Narayana P. Santhanam and Martin J. Wainwright. "Information-theoretic limits of selecting binary graphical models in high dimensions". In: *IEEE Transactions on Information Theory* 58.7 (2012), pp. 4117–4134. ISSN: 00189448. DOI: 10.1109/TIT.2012.2191659. arXiv: 0905.2639.