Learning Ising and Potts Models with Latent Variables

Surbhi Goel
The University of Texas at Austin

Abstract

We study the problem of learning graphical models with latent variables. We give the *first* efficient algorithms for learning: 1) ferromagnetic Ising models with latent variables under *arbitrary* external fields, and 2) ferromagnetic Potts model with latent variables under unidirectional non-negative external field. Our algorithms have optimal dependence on the dimension but suffer from a sub-optimal dependence on the underlying sparsity of the graph.

Our results rely on two structural properties of the underlying graphical models. These in turn allow us to design an influence function which can be maximized greedily to recover the structure of the underlying graphical model. These structural results may be of independent interest.

1 Introduction

Graphical models are a popular framework for expressing high dimensional distributions by using an underlying graph to represent conditional dependencies among the variables. Learning the underlying dependency structure of a graphical model using samples drawn from the distribution is a core problem in understanding graphical models. Much progress has been made in the recent years towards developing efficient algorithms for learning fundamental models such as Ising model, Potts model and Markov random fields (MRFs) with near optimal sample and time complexity under the assumptions of sparsity and/or correlation decay.

The *structure learning* problem becomes even more challenging when the underlying model is allowed to have latent (or latent) variables. Compared to fully observed models, latent variable models can induce more complex dependencies among the observed variables once the latent variables are marginalized over. In this work we restrict ourselves to a special class of latent variable models where the interactions are restricted to be pairwise only between observed and latent variables . In this work, we consider both Ising and Potts models with such latent variables.

In the Ising case, these are known as Restricted Boltzmann machines (RBMs). RBMs have been used for various unsupervised learning tasks (Hinton and Salakhutdinov, 2006; Larochelle and Bengio, 2008; Salakhutdinov et al., 2007; Hinton and Salakhutdinov, 2009) since their inception in the early 2000s. A RBM with latent variables induces a probability distribution over n observed variables $X \in \{\pm 1\}^n$ and m latent variables $Y \in \{\pm 1\}^m$ as follows,

$$\Pr[X = x, Y = y] \propto \exp(x^T J y + h^T x + g^T y) \quad (1)$$

Here $J \in \mathbb{R}^{n \times m}$ is the interaction matrix, $h \in \mathbb{R}^n, g \in \mathbb{R}^m$ are the external fields. Alternatively, a RBM can be viewed as a bipartite graph between the set of observed and latent variables with edge weights given by J. Here we use *restricted* to refer to the bipartite nature of the interaction graph.

Generalizing the above to non-binary state, we say a q-state Restricted Potts model (RPM) with latent variables induces a probability distribution over n observed variables $X \in [q]^n$ and m latent variables $Y \in [q]^m$ with

$$\Pr[X = x, Y = y] \propto \exp\left(\sum_{i \in [n], j \in [m]} J_{ij}\delta(x_i, y_j) + \sum_{i \in [n]} h_i \delta(x_i, 0) + \sum_{j \in [m]} g_j \delta(y_j, 0)\right) \quad (2)$$

where $\delta(a, b) = 0$ if $a \neq b$ else 0. We restrict the RPM to have non-negative external field only on state 0.

Recently Bresler et al. (2019) gave the first algorithm to learn *ferromagnetic* ($J \geq 0$) RBMs with non-negative external fields ($h, g \geq 0$). They applied the

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

famous Griffiths-Hurst-Sherman correlation inequality (Griffiths et al., 1970; Kelly and Sherman, 1968) to prove that a certain influence function is submodular and subsequently used a simple greedy algorithm to maximize the influence function. Their work relied heavily on the GHS inequality which requires the external fields to be *consistent*, that is, have the same sign. Also, their result does not extent to RPMs, it is not clear whether the concavity of magnetization holds in the Potts models.

In this paper we focus on 1) learning *ferromagnetic* RBMs with *arbitrary* external fields, and 2) learning ferromagnetic RPMs (*non-binary*) with non-negative external field supported on one state. We give the *first* efficient algorithms for these two settings.

Our Results. The first contribution of our paper is a key structural property of ferromagnetic RBMs with arbitrary external fields.

Theorem 1 (Informal version of Theorem 1)**.** *In a ferromagnetic RBM with arbitrary fields, for any pair of observed nodes u and v that share a common latent neighbor in the underlying graph, the covariance between u and v is at least some positive constant independent of n.*

The above key property gives us the following structure learning result for RBMs.

Theorem 2 (Informal version of Theorem 6)**.** *Consider a ferromagnetic RBM with arbitrary external fields such that all non-zero interactions are bounded below by α and the sum of absolute weights of outgoing edges of every node (plus absolute value of external field) is bounded above by λ. Then there is an algorithm that recovers the markov blanket of each observed variable in time $\tilde{O}_{\alpha,\lambda}(n^2)$ and sample complexity $O_{\alpha,\lambda}(\log n)$[1].*

The second contribution of our paper is a key structural property of ferromagnetic RPMs with non-negative external fields on one state.

Theorem 3 (Informal version of Theorem 7)**.** *Let u be an observed node in a ferromagnetic q-state RPM with non-negative external field on state 0. For any pair of observed nodes u and v that share a common latent neighbor in the underlying graph, the probability of u and v being simultaneously 0 is greater than the product of the probabilities of each of them being 0 by at least some positive constant independent of n.*

The above key property gives us the following structure learning result for RPMs.

[1]The sub-script indicates that the dependency on α, λ is suppressed. Also $\tilde{O}$ hides logarithmic dependencies.

Theorem 4 (Informal version of Theorem 9)**.** *Consider a ferromagnetic q-state RPM with non-negative external field on state 0 such that all non-zero interactions are bounded below by α and the sum of absolute weights of outgoing edges of every node (plus external field) is bounded above by λ. Then there is an algorithm that recovers the markov blanket of each observed variable in time $\tilde{O}_{\alpha,\lambda}(n^2)$ and sample complexity $O_{\alpha,\lambda}(\log n)$.*

Note that our bounds are similar to those in Bresler (2015). In both our results, the dependence on $1/\alpha$ is exponential and that on λ is doubly exponential. However, we note that for ferromagnetic RBMs with consistent fields, Bresler et al. (2019) have a singly exponential dependence on λ which is optimal. Note that once we have learned the underlying structure of the graphical model, we can use standard techniques to learn a representation in the form of a Markov Random Field of the RBM/RPM (see Bresler et al. (2019)).

Our Techniques. For our first key structural result, we use Percus's transformation (Percus, 1975) on the variables that enables us to use symmetry arguments to prove that the covariance is bounded away from 0. Similarly, for the second key structural result, we use the Random Cluster Expansion (Fortuin and Kasteleyn, 1972) to lower bound the quantity of interest. Both results give a stronger version of the FKG based correlation inequality through a more involved analysis with special care to avoid a dimension dependence.

For learning RBMs/RPMs, in the spirit of the influence maximization algorithm due to Bresler (2015), we design a corresponding influence function in both setups and greedily maximize the same in order to iteratively build the neighborhood of each observed vertex. Using an information theoretic argument, we can show that our iterative algorithm followed by pruning returns us the exact neighborhood of each vertex.

Related Work. Structure learning for graphical models is a well studied problem, with major focus on the fully-observed model. The first algorithms were proposed by Chow and Liu (1968) for learning undirected graphical models on trees. Subsequently, various algorithms were proposed for structure learning under varying assumptions on the underlying model (Lee et al., 2007; Ravikumar et al., 2010; Yang et al., 2012; Bresler, 2015; Vuffray et al., 2016; Klivans and Meka, 2017; Hamilton et al., 2017; Wu et al., 2019). Bresler (2015) proposed a simple greedy algorithm based on influence maximization for assumption-free structure learning of Ising models. His algorithm achieved optimal sample/time complexity in terms of the dimension however depended doubly exponentially

on the degree of the underlying graph. Subsequently Vuffray et al. (2016) and Klivans and Meka (2017) proposed alternative techniques to remove the doubly exponential dependence. Klivans and Meka (2017) were the first to give efficient algorithms for non-binary models. The dependence on alphabet size was further improved by Wu et al. (2019).

The problem of structure recovery in the presence of latent variables is not as well understood as the fully-observed setting. For locally tree-like models, Anandkumar et al. (2013) gave efficient algorithms for recovery under correlation decay assumption. Assuming that the latent variables are distributed according to a Gaussian distribution, Nussbaum and Giesen (2019) proposed a likelihood model for sparse + low rank model for structure learning. The most relevant to our work is that of Bresler et al. (2019) which proposed the first algorithm for structure recovery of ferromagnetic RBMs with non-negative external fields using concavity of magnetization. Unlike their setup, we extend to larger state space and allow the external fields to be arbitrary in the case of binary state space at the cost of a worse dependence on α, λ. The presence of inconsistent external fields allows for different biases on different latent nodes. For example, in a social network, if we want to model the distribution of votes for party A vs party B then it is likely that individuals have different inherent inclinations (arbitrary external field) and are subsequently influenced to vote similarly as their friends (ferromagnetic interactions). These biases often create conflicts and often make the problem much more challenging. It is well-known that arbitrary external fields can greatly change the complexity of closely related problems such as approximating the partition function (Goldberg and Jerrum, 2007).

2 Preliminaries

We consider a binary RBM on underlying bipartite graph $G = ([n], [m], E)$ over observed variables X and latent variables Y. Each configuration of observed/latent variables $\in \pm 1$ is assigned probability according to (1) where J indicates the weighted edges. We let Z^{IS} be the partition function (normalizing constant). Similarly, we consider a q-state RPM where each configuration of observed/latent variables $\in [q]$ is assigned probability according to (2). We make the following assumptions,

Minimum ferromagnetic interaction: For all $i \in [n], j \in [m]$, if $J_{ij} \neq 0$ then $J_{ij} \geq \alpha$.

Weight sparsity: For all $i \in [n]$, $\sum_j |J_{ij}| + |h_i| \leq \lambda$ and for all $j \in [m]$, $\sum_i |J_{ij}| + |g_j| \leq \lambda$.

Additionally, for the q-state RPM we assume that $h, g \geq 0$. Note that we do not need this assumption on external fields for the Ising case.

Remark: For the binary RBM, our model covers a more general class of *locally consistent* RBMs where for each $j \in [m]$, $J_{ij} \geq 0$ for all $i \in [n]$ (ferromagnetic) or $J_{ij} \leq 0$ for all $i \in [n]$ (anti-ferromagnetic). This can be reduced to the ferromagnetic case straightforwardly. If there exists j such that $J_{ij} \leq 0$ for all i (locally consistent) then we can map $Y_j \to -Y_j$ without affecting the marginal on X and the model is ferromagnetic at j. The change of variable will reverse the external field at j however since we do not make any assumption on the sign of the external field, our model assumptions still hold. We can repeat this for all such j and the model can therefore be made ferromagnetic.

Learning task. Define $N(u) := \{j : J_{uj} \neq 0\}$ to be the graph-theoretic neighborhood of observed node u and define $N_2(u) = \{i : \exists\ j, J_{ij}, J_{uj} \neq 0\}$ to be the two-hop graph-theoretic neighborhood. We also define $N_2^{mkv}(u)$ to be the two-hop Markov neighborhood, that is, the smallest set $S \subseteq [n] \backslash \{u\}$ such that X_u is conditionally independent of X_v for all $v \in [n] \backslash (S \cup \{u\})$. Our objective is to recover the two-hop Markov neighborhood of each observed variable. In both our setting, this will correspond exactly to the two-hop graph-theoretic neighborhood of each observed variables.

3 Binary RBMs

In this section we will first present our structural result for binary RBMs and subsequently show how to use the result to obtain a learning algorithm.

3.1 Key Property: Conditional Covariance

We show that for two observed nodes sharing a common latent neighbor, the covariance is positive and bounded away from 0. The main motivation to believe that such a structural result holds is the famous FKG inequality Percus (1975); Sylvester (1976) which states that for ferromagnetic Ising models with arbitrary external field the covariance of any two nodes is non-negative. We extend this result to show that if the interactions are large, then this covariance is indeed bounded away from 0.

Let us define the conditional covariance for observed nodes $u, v \in [n]$ and a subset of observed nodes $S \subseteq [n] \backslash \{u, v\}$ with configuration x_S as follows,

$$\mathsf{Cov}(u, v | X_S = x_S) := \mathbb{E}[X_u X_v | X_S = x_S] - \mathbb{E}[X_u | X_S = x_S]\ \mathbb{E}[X_v | X_S = x_S].$$

Here X_S denotes the subset of coordinates of X selected by S.

We also define the notion of average conditional covariance as follows, $\mathsf{Cov}_{\mathsf{avg}}(u, v|S) = \mathbb{E}_{x_S}[\mathsf{Cov}(u, v|X_S = x_S)]$. We will prove the following property of the conditional covariance:

Theorem 5. *Under our assumptions, for fixed node u and any fixed subset of observed nodes $S \subseteq [n]\backslash\{u\}$ with configuration x_S, then for all $v \in N_2(u)\backslash S$,*

$$\mathsf{Cov}(u, v|X_S = x_S) \geq \alpha^2 \cdot \exp(-12\lambda).$$

Proof. Observe that we can restrict to proving the above result for $S = \emptyset$ since conditioning over a set of observed variables ($X_S = x_S$)) will give us a new binary RBM which satisfies our assumptions. Moreover, the edges between the the remaining nodes remain the same with the same edge weights.

Percus's Transformation. We will utilize the transformation proposed by Percus Percus (1975). It uses a simple idea of making two copies of the underlying graph and using the symmetry of this transformation to prove useful properties. The probability of a configuration under this new distribution $\mathcal{D}$ is

$$\begin{aligned}&\Pr[X = x, Y = y, X' = x', Y' = y'] \\ &\quad \propto \exp(x^T Jy + h^T x + g^T y + {x'}^T Jy' + h^T x' + g^T y')\end{aligned}$$

Observe that $\Pr[X = x, Y = y, X' = x', Y' = y'] = \Pr[X = x', Y = y', X' = x, Y' = y] = \Pr[X = x, Y = y] \Pr[X' = x', Y' = y']$.

Define $\underline{X}_i = \frac{X_i - X'_i}{\sqrt{2}}$, $\underline{Y}_i = \frac{Y_i - Y'_i}{\sqrt{2}}$ and $\overline{X}_i = \frac{X_i + X'_i}{\sqrt{2}}$, $\overline{Y}_i = \frac{Y_i + Y'_i}{\sqrt{2}}$. Then we have

$$\begin{aligned}&\Pr[X = x, Y = y, X' = x', Y' = y'] \\ &\quad \propto \exp(x^T Jy + h^T x + g^T y + {x'}^T Jy' + h^T x' + g^T y') \\ &\quad = \exp\left(\overline{x}^T J\overline{y} + \underline{x}^T J\underline{y} + \sqrt{2}h^T\overline{x} + \sqrt{2}g^T\overline{y}\right).\end{aligned}$$

Thus under this transformation, we can rewrite the covariance in the following way,

$$\begin{aligned}\mathsf{Cov}(u, v) &= \mathbb{E}[X_u X_v] - \mathbb{E}[X_u]\mathbb{E}[X_v] \\ &= \frac{1}{2}(\mathbb{E}_{\mathcal{D}}[X_u X_v] - \mathbb{E}_{\mathcal{D}}[X_u X'_v]) \\ &\quad + \frac{1}{2}(\mathbb{E}_{\mathcal{D}}[X'_u X'_v] - \mathbb{E}_{\mathcal{D}}[X'_u X_v]) \\ &= \frac{1}{2}\mathbb{E}_{\mathcal{D}}[(X_u - X'_u)(X_v - V'_v)] \\ &= \mathbb{E}_{\mathcal{D}}[\underline{X}_u \underline{X}_v]\end{aligned}$$

The above follows from the independence of (X, Y) and (X', Y') and the symmetry of the Percus transformation.

Positivity of Covariance. First we will show that the covariance is positive. Let k be the common neighbor of u and v, that is, $J_{uk}, J_{vk} \neq 0$. There must be such a k since $v \in N_2(u)$. We will define the following useful terms, $\gamma(\underline{x}, \underline{y}) = \exp(\underline{x}^T J\underline{y} - \underline{x}_u J_{uk}\underline{y}_k - \underline{x}_v J_{vk}\underline{y}_k)$ and $\Delta(\overline{x}, \overline{y}) = \exp(\overline{x}^T J\overline{y} + \sqrt{2}h^T\overline{x} + \sqrt{2}g^T\overline{y})$. Using these, we have,

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[\underline{X}_u \underline{X}_v] &= \frac{1}{(Z^{IS})^2} \sum_{x,x',y,y'} \underline{x}_u \underline{x}_v \exp(\overline{x}^T J\overline{y} + \underline{x}^T J\underline{y} \\ &\qquad + \sqrt{2}h^T\overline{x} + \sqrt{2}g^T\overline{y}) \\ &= \frac{1}{(Z^{IS})^2} \sum_{x,x',y,y'} \underline{x}_u \underline{x}_v \exp((\underline{x}_u J_{uk} \\ &\qquad + \underline{x}_v J_{vk})\underline{y}_k)\gamma(\underline{x}, \underline{y})\Delta(\overline{x}, \overline{y}) \\ &= \frac{1}{(Z^{IS})^2} \sum_{x,x',y,y'} \sum_{i=0}^{\infty} \underline{x}_u \underline{x}_v \left(\frac{(\underline{x}_u J_{uk} + \underline{x}_v J_{vk})^i \underline{y}_k^i}{i!}\right) \\ &\quad \times \gamma(\underline{x}, \underline{y})\Delta(\overline{x}, \overline{y}) \\ &= \frac{1}{(Z^{IS})^2} \sum_{x,x',y,y'} \sum_{i=0}^{\infty} \sum_{j=0}^{i} \frac{1}{i!}\binom{i}{j} J_{uk}^j J_{vk}^{i-j} \underline{x}_u^{j+1} \underline{x}_v^{i+1-j} \underline{y}_k^i \\ &\quad \times \gamma(\underline{x}, \underline{y})\Delta(\overline{x}, \overline{y})\end{aligned} \tag{3}$$

The following lemma shows that each term in the summation is non-negative.

Lemma 1. *For all $A \in \mathbb{Z}_+^n, B \in \mathbb{Z}_+^n$ and function f over $\overline{x}, \overline{y}$ such that $f \geq 0$,*

$$\sum_{x,x',y,y'} \prod_{a \in [n]} \underline{x}_a^{A_a} \prod_{b \in [m]} \underline{y}_b^{B_b} f(\overline{x}, \overline{y}) \geq 0.$$

Proof. Observe that for any $i \in [n]$, exchanging $x_i \leftrightarrow x'_i$ does not change the summation, however it changes $\underline{x}_i \to -\underline{x}_i$ while leaving $\overline{x}_i \to \overline{x}_i$ unchanged. Thus, if A_i is odd, then the summation will be 0. Therefore, for the term to be non-zero, for all $i \in [n]$, A_i must be even. Similarly, for all $j \in [m]$, B_j must be even. Now since $f \geq 0$ and there are only even powers, the summation must be positive. □

It is easy to see that $\gamma(\underline{x}, \underline{y})$ can be expanded as a multivariate polynomial over $\underline{x}, \underline{y}$ with non-negative coefficients (since $J \geq 0$)[2]. Therefore, applying Lemma 1, we have for all $i \geq j$,

$$\sum_{x,x',y,y'} \underline{x}_u^{j+1} \underline{x}_v^{i+1-j} \underline{y}_k^i \gamma(\underline{x}, \underline{y})\Delta(\overline{x}, \overline{y}) \geq 0.$$

[2]Since γ is an exponential function of a polynomial with non-negative coefficients, using taylor expansion of e^a, we will overall get a polynomial with all non-negative coefficients.

This implies that the covariance is indeed *non-negative*.

Lower Bound on Covariance. We will show that in fact the covariance is at least a constant independent of n. Since all terms are non-negative, we can lower bound the expression by the term corresponding to $i = 2$ and $j = 1$. This yields only squares of $\underline{x}_u, \underline{x}_v, \underline{y}_k$ as follows,

$$(3) \geq \frac{1}{(Z^{IS})^2} \sum_{x,x',y,y'} J_{uk} J_{vk} \underline{x}_u^2 \underline{x}_v^2 \underline{y}_k^2 \gamma(\underline{x}, \underline{y}) \Delta(\overline{x}, \overline{y})$$
$$\geq \frac{\alpha^2}{(Z^{IS})^2} \sum_{x,x',y,y'} \underline{x}_u^2 \underline{x}_v^2 \underline{y}_k^2 \gamma(\underline{x}, \underline{y}) \Delta(\overline{x}, \overline{y}).$$

Here the second inequality follows from noting that by our assumption $J_{uk}, J_{vk} \neq 0$ and hence must be at least α. The last obstacle is to bound the remaining expression independent of n. We use the following lemma to bound the same.

Lemma 2. *We have,*

$$\frac{1}{(Z^{IS})^2} \sum_{x,x',y,y'} \underline{x}_u^2 \underline{x}_v^2 \underline{y}_k^2 \gamma(\underline{x}, \underline{y}) \Delta(\overline{x}, \overline{y}) \geq \exp(-12\lambda).$$

The proof requires a very careful analysis to avoid the naive dimension dependence. Now, using Lemma 2 gives us the desired result. □

Corollary 1. *For $u \neq v \in [n]$ such that there exists $w \in [m]$ with $(u,k), (v,k) \in E$ and a subset of observed nodes $S \subseteq [n] \backslash \{u, v\}$, $\mathsf{Cov}_{\mathsf{avg}}(u, v|X_S) \geq \alpha^2 \cdot \exp(-12\lambda)$.*

Proof. Since for any $X_S = x_S$, by Lemma 1, the covariance is bounded below by $\alpha^2 \exp(-12\lambda)$, hence the expectation is also bounded by the same quantity. □

Remark 1. *Observe that the above lemma also shows that $N_2(u) \subseteq N_2^{mkv}(u)$. It is not hard to see that $N_2^{mkv}(u) \subseteq N_2(u)$ by the structure of the RBM therefore $N_2(u) = N_2^{mkv}(u)$.*

Remark 2. *The key structural result can be extended to the setting in which there are edges between latent and observed variables using the same techniques, however now the bound will be exponentially worse in terms of the length of the shortest path connecting two observed nodes similar to Bresler et al. (2019).*

3.2 Algorithm: Greedy Maximization

In this section we present the main algorithm (Algorithm 1) and a proof of its correctness. Our algorithm and analysis is similar to the influence maximization algorithms for learning Ising models as in Bresler (2015). Our algorithm exploits the key property to maximize conditional covariance to greedily build the two-hop neighborhood.

Let us define the empirical conditional covariance computed using a sample $\mathcal{S} = \{X^{(1)}, \ldots, X^{(M)}\}$ of size M by $\widehat{\mathsf{Cov}}_{\mathsf{avg}}(\cdot, \cdot|\cdot)$. For a given threshold $\tau > 0$, our algorithm is as follows:

Algorithm 1 LearnRBMNbhd(u)

1: Set $S := \phi$
2: Let $i^* = \arg\max_{v \in [n] \backslash S \cup \{u\}} \widehat{\mathsf{Cov}}_{\mathsf{avg}}(u, v|S)$, and $\eta^* = \max_v \widehat{\mathsf{Cov}}_{\mathsf{avg}}(u, v|S)$
3: **if** $\eta^* \geq \tau$ **then**
4: $\quad S = S \cup \{i^*\}$
5: **else**
6: $\quad$ Go to Step 8
7: Go to Step 2
8: (*Pruning:*) For each $v \in S$, if $\widehat{\mathsf{Cov}}_{\mathsf{avg}}(u, v|S \backslash \{v\}) < \tau$, remove v
9: Return S

Theorem 6. *For $\tau = \alpha^2 \exp(-12\lambda)/2$ and $\delta = \exp(-\lambda)/2$, with probability $1 - \zeta$,* LearnRBMNbhd(u) *outputs* exactly *the two-hop neighborhood of observed variable u for*

$$M \geq \Omega\left((\log(1/\zeta) + T^* \log(n)) \frac{2^{2T^*}}{\tau^2 \delta^{2T^*}} \right) \text{ for } T^* = \frac{8}{\tau^2}.$$

The algorithm runs in time $O(T^* M n)$ for node u.

Proof. The proof follows along the same lines as Bresler (2015). We will first show that our estimates of conditional covariance are close to the true values given M samples. We will then show that after T iterations, set S contains a superset of the two-hop neighbors. Lastly we will show that our refining step removes all nodes except the two-hop neighbors.

Closeness of Estimates Denote by $\mathcal{A}(l, \epsilon)$ the event such that for all u, v and S with $|S| \leq l$, simultaneously, $\left| \widehat{\mathsf{Cov}}_{\mathsf{avg}}(u, v|S) - \mathsf{Cov}_{\mathsf{avg}}(u, v|S) \right| \leq \epsilon$.

Lemma 3. *For fixed $l, \epsilon, \zeta \geq 0$, if the number of samples is $\Omega\left((\log(1/\zeta) + l \log(n)) \frac{2^{2l}}{\epsilon^2 \delta^{2l}} \right)$. then $\Pr[A(l, \epsilon)] \geq 1 - \zeta$.*

We defer the proof of the above lemma to the appendix. Choosing $M = \Omega\left((\log(1/\zeta) + T^* \log(n)) \frac{2^{2T}}{\tau^2 \delta^{2l}} \right)$, we have $A := A(T^*, \tau/2)$ holds for $T^* = 8/\tau^2$ with probability $1 - \zeta$. From now on we assume A holds.

Entropy Gain. We will show that the conditional mutual information is bounded below by a function of the average conditional covariance thus at each iteration of the algorithm we are increasing the overall entropy of X_u.

Lemma 4. *For $u \neq v \in [n]$ and a subset of observed nodes $S \subseteq [n]\backslash\{u, v\}$ with configuration x_S,*

$$\sqrt{2I(X_u; X_v|X_S)} \geq \mathsf{Cov}_{\mathsf{avg}}(u, v|S)$$

Upper Bound on Size of S. We will show that $|S| \leq T^*$ before pruning. Let the sequence of added nodes be $i_1, \ldots, i_T$ for some T and $S_l = \{i_1, \ldots, i_l\}$ for $1 \leq l \leq T$. For each $j \in T$, we have $\widehat{\mathsf{Cov}}_{\mathsf{avg}}(u; i_j|X_{S_j}) \geq \tau$ (by Step 3). If $T \geq T^*$, then we have $\mathsf{Cov}_{\mathsf{avg}}(u; i_j|X_{S_j}) \geq \tau/2$ for all $j \leq T^* + 1$ (since A holds). Thus we have,

$$\begin{aligned} 1 \geq H(X_u) &\geq I(X_u|X_S) \\ &= \sum_{j=1}^{T} I(X_u; X_{i_j}|S_{j-1}) \geq \frac{T^*+1}{8}\tau^2. \end{aligned}$$

Here the inequalities follow from standard properties of entropy and mutual information. This leads to a contradiction since $T^* = \frac{8}{\tau^2}$. Thus, we have $T \leq T^*$. Observe that each iteration requires $O(Mn)$ time and at most T^* iterations take place prior to pruning. Also pruning takes $O(Mn)$ time, giving us a total runtime of $O(T^*Mn)$.

Recovery of Two-hop Neighborhood. We will show that $N_2(u) \subseteq S$. Suppose $N_2(u) \not\subseteq S$, then there exists $v \in N_2(u)$. By Lemma 5, we know that $\mathsf{Cov}_{\mathsf{avg}}(u, v|X_S) \geq \alpha^2 \exp(-12\lambda) = 2\tau$. Since A holds and $|S| \leq 8/\tau^2$, we have $\widehat{\mathsf{Cov}}_{\mathsf{avg}}(u, v|X_S) \geq 3\tau/2$, thus the algorithm would not have terminated. This is a contradiction, thus $N_2(u) \subseteq S$ before pruning.

Now if $v \notin N_u(S)$ then $\mathsf{Cov}(u, v|X_{S\backslash\{v\}}) = 0$ since conditional on the 2-hop neighborhood, X_u and X_v are independent, therefore they will be removed. Whereas, by Lemma 5, if $v \in N_u(S)$ then $\mathsf{Cov}(u, v|X_{S\backslash\{v\}}) \geq 2\tau$ and our test will not remove it (estimates of covariance are correct withing $\alpha/2$). Thus we will exactly obtain the neighborhood at the end of the algorithm. □

Remark 3. *Bresler et al. (2019) showed a hardness result for structure learning of RBMs by reduction from learning sparse parities with noise over the uniform distribution. Under our assumption of $J \geq 0$, this construction is not achievable. We refer the reader to the supplementary for more details.*

4 Non-binary RPMs

In this section we will first present our structural result for q-state RPMs and subsequently show how to use the result to obtain a learning algorithm similar to the binary case.

4.1 Key Property: Conditional Influence

We define the following conditional influence function for observed nodes $u, v \in [n]$ and a subset of observed nodes $S \subseteq [n]\backslash\{u, v\}$ as follows,

$$\begin{aligned} \mathsf{inf}(u, v|S) = \Pr[X_u = 0|X_S = 0^S, X_v = 0] \\ - \Pr[X_u = 0|X_S = 0^S]. \end{aligned}$$

We will prove the following useful property of inf,

Theorem 7. *Under our assumptions, for fixed node u and any fixed subset of observed nodes $S \subseteq [n]\backslash\{u\}$, for all $v \in N_2(u)\backslash S$,*

$$\mathsf{inf}(u, v|S) \geq \frac{q-1}{q} \cdot \alpha^2 \cdot \frac{\exp(-3\lambda)}{(\exp(\lambda) + q - 1)^3}.$$

Proof. To prove the theorem, we can assume that $S = \emptyset$ since our assumptions are still satisfied for this model. Our proof will use the Fortuin-Kasteleyn Random Cluster (RC) Fortuin and Kasteleyn (1972) expansion of Potts models.

Random Cluster Model. We give a brief overview of the random cluster model and useful properties for our analysis. We refer the reader to Fortuin and Kasteleyn (1972); Cioletti and Vila (2016) for a more detailed exposition.

For the graph G corresponding to the Potts model in (2), let $\lambda(x, y)$ denote the probability measure. The Random Cluster (RC) model is introduced by randomly choosing a set of edges from E which are said to be occupied. Let $\omega \in \{0, 1\}^E$ be the indicator for the occupied edges, then the random cluster probability measure ϕ is defined as follows,

$$\phi(\omega) = \frac{1}{Z^{RC}} \cdot B(\omega) \cdot \prod_{c \in C(\omega)} (\exp(H(c)) + q - 1)$$

where $C(\omega)$ is the set of clusters (connected components containing observed and latent variables) into which $V = [n] \cup [m]$ is partitioned by ω and $H(c) = \sum_{i \in c_{obs}} h_i + \sum_{j \in c_{lat}} g_j$ with c_{obs} (c_{lat})) being the observed (latent) nodes in c. B is defined as follows,

$$B(\omega) := \prod_{\omega_{ij}=1} p_{ij} \cdot \prod_{\omega_{ij}=0} (1 - p_{ij})$$

where $p_{ij} = 1 - \exp(-J_{ij})$. Here Z^{RC} is the partition function (normalizing factor). The RC model has been shown to satisfy the FKG property under our setting. More formally,

Theorem 8 (Fortuin and Kasteleyn (1972); Cioletti and Vila (2016)[3])**.** *For non-decreasing functions f, g over the standard partial order on ω, $\mathbb{E}_\phi[f \cdot g] \geq \mathbb{E}_\phi[f] \cdot \mathbb{E}_\phi[g]$.*

There is a nice coupling between the Potts model and the RC expansion defined by the Edwards-Sokal (ES) model Edwards and Sokal (1988). On the graph G, it is a coupled distribution over $[q]^V \times \{0,1\}^E$. A pair of configurations (x, y) and ω are compatible if $\omega_{ij} = 1 \implies x_i = y_j$ for all $(i, j) \in E$. Denote $\Delta(x, y, \omega)$ to be the indicator for this condition. Then the joint distribution between (x, y) and ω is as follows,

$$\nu(x, y, \omega) = \frac{1}{Z^{ES}} \cdot B(\omega) \cdot \Delta(x, y, \omega) \cdot \exp\left(\sum_{i \in [n]} h_i \delta(x_i, 0) + \sum_{j \in [m]} g_j \delta(y_j, 0)\right)$$

where Z^{ES} is the partition function for the above. The ES coupling relates to the Potts and RC model as follows,

Lemma 5 (Edwards and Sokal (1988); Cioletti and Vila (2016))**.** *Marginals of ES match exactly the Potts and RC model, that is,*

$$\sum_{\omega \in \{0,1\}^E} \nu(x, y, \omega) = \lambda(x, y) \text{ and } \sum_{\substack{x \in [q]^n \\ y \in [q]^m}} \nu(x, y, \omega) = \phi(\omega).$$

Relating Influence to Connectivity. Using the ES coupling and FKG property of RC, we will show that we can lower bound the influence function to a weighted connectivity in the RC model.

Lemma 6. *For any $u, v \in [n]$, we have,*

$$\lambda(X_u = X_v = 0) - \lambda(X_u = 0) \cdot \lambda(X_v = 0) \geq \mathbb{E}_\phi\left[\mathbb{1}[u \leftrightarrow v] \cdot \frac{\exp\left(H(C_u(\omega))\right)}{\left(\exp\left(H(C_u(\omega))\right) + q - 1\right)^2}\right].$$

where $u \leftrightarrow v$ denoted that u is in the same connected component for given ω and $C_u(\omega)$ denotes the cluster containing u in $C(\omega)$.

[3]The FKG property for the RC was first studied by Fortuin and Kasteleyn (1972) under zero external field. For the setting of one-directional non-negative external field, this property was proven by Cioletti and Vila (2016). We refer the reader to Cioletti and Vila (2016) for the most general setup where this holds.

The intuition behind the relation to connectivity is that in the ES coupling, the vertices in the same cluster have the same state, hence if the probability of the nodes being connected is high, they are likely to have the same state. Following a careful analysis, the weighted expectation can be lower bounded independent of dimension n using the following lemma.

Lemma 7. *For any $u, v \in [n]$, we have*

$$\mathbb{E}_\phi\left[\mathbb{1}[u \leftrightarrow v] \cdot \frac{\exp\left(H(C_u(\omega))\right)}{\left(\exp\left(H(C_u(\omega))\right) + q - 1\right)^2}\right] \geq \frac{q-1}{q} \cdot \alpha^2 \cdot \frac{\exp(-3\lambda)}{(\exp(\lambda) + q - 1)^3}$$

Note that $\mathsf{inf}(u, v|\emptyset) = \frac{\lambda(X_u = X_v = 0)}{\lambda(X_v = 0)} - \lambda(X_u = 0) \geq \lambda(X_u = X_v = 0) - \lambda(X_u = 0) \cdot \lambda(X_v = 0)$ since $\lambda(X_v = 0) \leq 1$. Now using Lemma 7 gives us the desired result.

□

4.2 Algorithm: Greedy Maximization

Similar to the binary RBM setup, we iteratively build the neighborhood by maximizing the conditional influence. The algorithm essentially remains the same as Algorithm 1 replacing $\mathsf{Cov}_{\mathsf{avg}}$ by inf. Our algorithm gives us the following guarantee,

Theorem 9. *For $\tau = \frac{q-1}{2q} \cdot \alpha^2 \cdot \frac{\exp(-3\lambda)}{(\exp(\lambda)+q-1)^3}$ and $\delta = \exp(-\lambda)/q$, with probability $1 - \zeta$, our algorithm outputs* exactly *the two-hop neighborhood of observed variable u for*

$$M \geq \Omega\left((\log(1/\zeta) + T^* \log(n)) \frac{2^{2T^*}}{\tau^2 \delta^{2T^*}}\right) \text{ for } T^* = \frac{2}{\tau}.$$

The algorithm runs in time $O(T^* M n)$ for each node u.

The proof follows from essentially the same arguments as the RBM case with a simpler argument on the bound of set S. We refer the reader to the supplementary for explicit details. Note that the dependence on q is exponential. To improve this is an interesting open question.

5 Experimental Evaluations

Synthetic experiments. For our synthetic experiments, we sample exactly from a binary RBM. Due to high cost of sampling data from an RBM (state space is exponential in the dimension of observed variables), we restrict to low dimensions of the observed variables <= 15. We consider a bipartite graphs with the same number of hidden and observed nodes ($n = m$) and

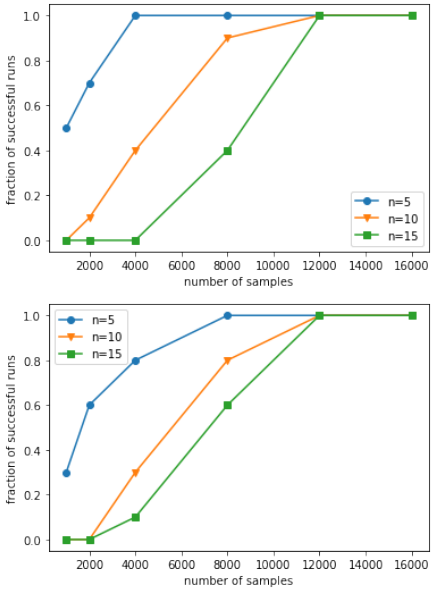

Figure 1: Fraction of runs for which the two-hop neighborhood was exactly recovered with varying sample size. The underlying graph has edge weights 1. The two plots are have different magnitudes of external fields on the hidden nodes.

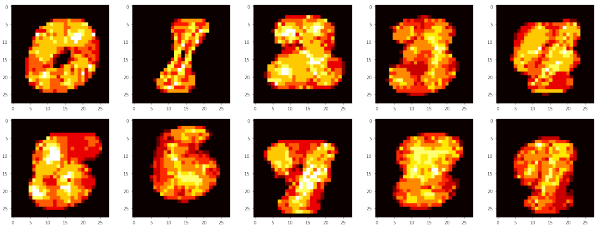

Figure 2: Heat-maps representing frequency of node to be in a two-hop neighborhood of another node for binarized MNIST.

constant bounded degree ($d = 2$). The edges are selected so as observed node i is connected to latent node i and $(i+1)\%n$ (cyclically to the corresponding hidden node and the next node with edge weight 1. The external field is randomly selected with value ± 0.2 (Figure 1 - top) and $\pm$ 0.4 (Figure 1 - bottom) on each latent variable. We set $\tau = 0.02$ for the experiment. Under this setup, we vary the input dimension as well as the number of samples and plot the fraction of runs (out of 10) in which we were able to recover the graph exactly. This is consistent with our analysis. The choice of parameters is arbitrary and we observed similar performance up to tuning τ.

MNIST experiments. The bottleneck for running on synthetic data was the data generation step and not the algorithm. Therefore, we also run our algorithm on binarized MNIST (pixels set to 0 and 1 with threshold at 0.5). Here $n = 784$ and number of samples per digit is $M \approx 5000$. For each digit class, we run our algorithm to recover a bounded size neighborhood (set to 8). We plot a heat-map corresponding to the frequency of a node appearing in the two-hop neighborhood of another node (see Figure 2). Note that background pixels do not appear in any neighborhoods since they are always 0 in our data and hence have no positive correlation with other pixels. This heat-map is able to identify the relevant structure of the digit and importance of nodes. We observe that the algorithm recovers a subset of the geometric neighborhood along with some other positively correlated pixels (see Figure 3). This learned structure can be incorporated in to downstream tasks such as classification.

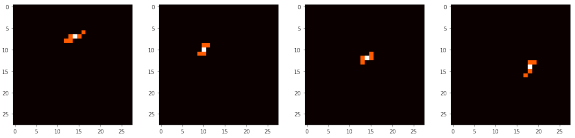

Figure 3: Visualizing learned neighborhoods for digit 0 data for selected nodes. Here the pixel in white represents the selected node and the ones in red denote the neighborhood. Observe that the neighborhood learned does not exactly match the geometric neighborhood.

6 Conclusions and Open Problems

In this work we present key structural properties of ferromagnetic binary RBMs with arbitrary external fields and ferromagnetic q-state RPMs with non-negative field on a fixed state. Subsequently we show how to use these properties to iteratively build the two-hop neighborhood of each node. Our algorithms run in optimal time and sample complexity in terms of the dimension however pay doubly exponentially in the upper bound on the weights and exponentially in the number of states. This seems to be an artifact of the approach of maximizing influence in general whereas algorithms using convex optimization are able to avoid this dependence for fully-observed graphical models. A natural open question is to improve this dependency. Alternatively, proving a stronger structural result such as weak-submodularity could lead to better dependence. More broadly, understanding the most expressive class of RBMs that allow efficient structure learning is a worthwhile future direction to pursue. Further understanding non-binary RPMs with external field on more than one states is an interesting direction however even the FKG condition is not known to be satisfied for the corresponding RC model.

Acknowledgments. The author would like to thank Sumegha Garg and Jessica Hoffmann for comments on the initial draft, and Adam Klivans, Frederic Koehler and Josh Vekhter for useful discussions. The author was supported by the JP Morgan AI PhD Fellowship. Part of this work was done while the author was visiting the Simons Institute for Theoretical Computer Science, Berkeley.

References

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

Bresler, G. (2015). Efficiently learning ising models on arbitrary graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 771–782. ACM.

Bresler, G., Koehler, F., and Moitra, A. (2019). Learning restricted boltzmann machines via influence maximization. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 828–839.

Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467.

Cioletti, L. and Vila, R. (2016). Graphical representations for ising and potts models in general external fields. *Journal of Statistical Physics*, 162(1):81–122.

Edwards, R. G. and Sokal, A. D. (1988). Generalization of the fortuin-kasteleyn-swendsen-wang representation and monte carlo algorithm. *Physical review D*, 38(6):2009.

Fortuin, C. M. and Kasteleyn, P. W. (1972). On the random-cluster model: I. introduction and relation to other models. *Physica*, 57(4):536–564.

Goldberg, L. A. and Jerrum, M. (2007). The complexity of ferromagnetic ising with local fields. *Combinatorics, Probability and Computing*, 16(1):43–61.

Griffiths, R. B., Hurst, C. A., and Sherman, S. (1970). Concavity of magnetization of an ising ferromagnet in a positive external field. *Journal of Mathematical Physics*, 11(3):790–795.

Hamilton, L., Koehler, F., and Moitra, A. (2017). Information theoretic properties of markov random fields, and their algorithmic applications. In *Advances in Neural Information Processing Systems*, pages 2463–2472.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

Hinton, G. E. and Salakhutdinov, R. R. (2009). Replicated softmax: an undirected topic model. In *Advances in neural information processing systems*, pages 1607–1614.

Kelly, D. G. and Sherman, S. (1968). General griffiths' inequalities on correlations in ising ferromagnets. *Journal of Mathematical Physics*, 9(3):466–484.

Klivans, A. and Meka, R. (2017). Learning graphical models using multiplicative weights. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 343–354. IEEE.

Larochelle, H. and Bengio, Y. (2008). Classification using discriminative restricted boltzmann machines. In *Proceedings of the 25th international conference on Machine learning*, pages 536–543. ACM.

Lee, S.-I., Ganapathi, V., and Koller, D. (2007). Efficient structure learning of markov networks using l_1-regularization. In *Advances in neural Information processing systems*, pages 817–824.

Nussbaum, F. and Giesen, J. (2019). Ising models with latent conditional gaussian variables. In *Algorithmic Learning Theory*, pages 669–681.

Percus, J. (1975). Correlation inequalities for ising spin lattices. *Communications in Mathematical Physics*, 40(3):283–308.

Ravikumar, P., Wainwright, M. J., Lafferty, J. D., et al. (2010). High-dimensional ising model selection using l_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319.

Salakhutdinov, R., Mnih, A., and Hinton, G. (2007). Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798. ACM.

Sylvester, G. S. (1976). Inequalities for continuous-spin ising ferromagnets. *Journal of Statistical Physics*, 15(4):327–341.

Vuffray, M., Misra, S., Lokhov, A., and Chertkov, M. (2016). Interaction screening: Efficient and sample-optimal learning of ising models. In *Advances in Neural Information Processing Systems*, pages 2595–2603.

Wu, S., Sanghavi, S., and Dimakis, A. G. (2019). Sparse logistic regression learns all discrete pairwise graphical models. In *Advances in Neural Information Processing Systems*, pages 8069–8079.

Yang, E., Allen, G., Liu, Z., and Ravikumar, P. K. (2012). Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, pages 1358–1366.