Learning Graph Structure in Discrete Markov Random Fields

Rui Wu, R. Srikant and Jian Ni

Abstract

We present a general algorithm for learning the structure of discrete Markov random fields from i.i.d. samples. The algorithm either achieves the same computational complexity or lowers the computational complexity of earlier algorithms for several cases, and provides a new low-computational complexity algorithm for the case of Ising models where the underlying graph is the Erdős-Rényi random graph $G \sim G(p, \frac{c}{p})$.

I. INTRODUCTION

In many models of networks, such as social networks and gene regulatory networks, each node in the network represents a random variable and the graph encodes the conditional independence relations among the random variables. A Markov random field is a particular such representation which has applications in a variety of areas (see [3] and the references therein). In a Markov field, the lack of an edge between two nodes implies that the two random variables are independent, conditioned on all the other random variables in the network.

As a concrete example of structure learning, consider a social network in which only the participants’ actions are observed. In particular, we do not observe or are unable to observe interactions between the participants. Our goal is to infer relationships among the nodes (participants) in such a network by understanding the correlations among the nodes. The canonical example used to illustrate such inference problems is the US Senate. Suppose one has access to the voting patterns of the senators over a number of bills (and not their party affiliations or any other information), the question we would like to answer is the following: can we say that a particular senator’s vote is independent of everyone else’s when conditioned on a few other senators’ votes? In other words, if we view the senators’ actions as forming a Markov Random Field (MRF), we want to infer the topology of the underlying graph.

Structure learning, i.e, learning the underlying graph structure of a Markov random field, refers to the problem of determining if there is an edge between each pair of nodes, given i.i.d. samples from the joint distribution of the random vector. There are many algorithms for structure learning and they can be broadly divided into two classes: ones based on conditional independence tests [5], [12], [2], [3] and ones based on $l_1$ regularized maximum likelihood estimation which results in a convex program [13].

In this paper, we propose a class of algorithms based on conditional independence tests for structure learning. For all the previously known algorithms for which analytical complexity bounds are available, the number of samples required to recover the graph correctly with high probability, i.e, the sample complexity, is $O(\log p)$, where $p$ is the size of the graph. Not surprisingly, the sample complexity for our algorithm is also $O(\log p)$ under reasonable assumptions. Therefore, we focus on computational complexity to compare our algorithm with other algorithms in the literature.
As can be seen in prior work, the complexity of an algorithm for structure learning depends on the assumptions made about the Markov random field. Common assumptions include an upper bound on the node degree of the underlying graph [5], [12], restrictions on the class of parameters of the joint probability distribution of the random variables to ensure correlation decay [5], [12], [2], lower bounds on the girth of the underlying graph [12], and a sparse, probabilistic structure on the underlying random graph [2]. Most prior work provide algorithms specialized to the type of assumptions made regarding the Markov random field. One of the main contributions of this paper is to show that there is a common framework from one which one can derive algorithms for the special cases studied in the above papers. The common framework does not necessarily yield the same algorithms as in prior work, but provides algorithms which have the same or lower computational complexity compared to prior work [5], [12], [2], or generalizes the class of models to which the algorithms can be applied [2]. It is not clear how to compare [13] with our work. The algorithm in [13] learns the structure by solving $l_1$ regularized convex optimizations under a set of incoherent conditions, therefore it has the same computational complexity for all Ising models. But the incoherent conditions do not have a clear interpretation as conditions for the graph parameters, and is NP-hard to verify for a give Ising model [10].

Our algorithms follow from a new necessary and sufficient condition for the existence of an edge in the underlying graph of a Markov random field (see Lemma 3.2). We show that one can check whether this condition is satisfied or not by using a low-complexity test for many classes of sparse graphs, including the ones mentioned above. Our analysis of the class of Ising models on Erdős-Rényi graphs was motivated by the results in [2] which studies the special of the so-called ferromagnetic Ising models defined over an Erdős-Rényi random graph. We also note that [3] extend the results in [2] to more general sparse graphs (beyond the Erdős-Rényi model). Our algorithm is more generally applicable than [2], [3], and in particular, applies to general Ising models. The results in [3] exclude certain parameter values for general Ising models by noting that these difficult cases have Lebesgue measure zero. However, the parameters of their algorithm appear to be difficult to choose when the Ising model parameters lie close to this zero Lebesgue measure set.

We present some preliminaries in the next section. In Section III, we present our algorithm under the assumption that the joint distribution of the random variables represented by the Markov random field is known, and focus on analyzing the computational complexity of the algorithm. We apply our algorithm to several special cases of Markov random fields and show that the algorithm recovers the graph structure correctly. These results are proved in Section IV, VI and VII. In Section VIII, we take sample complexity into account and present our algorithm assuming only i.i.d. samples from the distribution are given. We can easily extend the previous results to show that with sample complexity $O(\log p)$, the algorithm recovers the graph structure correctly with high probability in the above cases.

II. Preliminaries

A. Markov Random Fields (MRFs)

Let $X = (X_1, X_2, \ldots, X_p)$ be a random vector with distribution $P$ and $G = (V, E)$ be an undirected graph consisting $|V| = p$ nodes with each node $i$ associated with the $i^{th}$ element $X_i$ of $X$. Before we define an MRF, we introduce the notation $X_S$ to denote any subset $S$ of the random variables in $X$. A random vector and graph pair $(X, G)$ is called an MRF if it satisfies one of the following three Markov properties:
1) Pairwise Markov: \(X_i \perp X_j | X_{V \setminus \{i,j\}}, \forall (i,j) \notin E\), where \(\perp\) denotes independence.

2) Local Markov: \(X_i \perp X_{V \setminus \{i\}} | X_i, \forall i \in V\), where \(N_i\) is the set of neighbors of node \(i\).

3) Global Markov: \(X_A \perp X_B | X_S\), if \(S\) separates \(A, B\) on \(G\). In this case, we say \(G\) is an I-map of \(X\).

Further if \(G\) is an I-map of \(X\) and the global Markov property does not hold if any edge of \(G\) is removed, then \(G\) is called a minimal I-map of \(X\).

In all three cases, \(G\) encodes a subset of the conditional independence relations of \(X\) and we say that \(X\) is Markov with respect to \(G\). We note that the global Markov property implies the local Markov property, which in turn implies the pairwise Markov property.

When \(P(x) > 0, \forall x\), the three Markov properties are equivalent, i.e., if there exists a \(G\) under which one of the Markov properties is satisfied, then the other two are also satisfied. Further, in the case when \(P(x) > 0, \forall x\), there exists a unique minimal I-map of \(X\). The unique minimal I-map \(G = (V, E)\) is constructed as follows:

1) Each random variable \(X_i\) is associated with a node \(i \in V\).

2) \((i, j) \notin E\) if and only if \(X_i \perp X_j | X_{V \setminus \{i,j\}}\).

In this case, we consider the case \(P(x) > 0, \forall x\) and are interested in learning the structure of the associated unique minimal I-map. We will also assume that, for each \(i, X_i\) takes on values in a discrete, finite set \(\mathcal{X}\). We will also be interested in the special case where the MRF is an Ising model, which we describe next.

### B. Ising Model

Ising models are a type of well-studied pairwise Markov random fields. In an Ising model, each random variable \(X_i\) takes values in the set \(\mathcal{X} = \{-1, +1\}\) and the joint distribution is parameterized by constants called edge coefficients \(J\) and external fields \(h\):

\[
P(x) = \frac{1}{Z} \exp \left( \sum_{(i,j) \in E} J_{ij} x_i x_j + \sum_{i \in V} h_i x_i \right).
\]

If \(h = 0\), we say the Ising model is zero-field. If \(J_{ij} \geq 0\), we say the Ising model is ferromagnetic.

Ising models have the following useful property. Given an Ising model, the conditional probability \(P(X_{V \setminus S} | x_S)\) corresponds to an Ising model on \(V \setminus S\) with edge coefficients \(J_{ij}, i, j \in V \setminus S\) unchanged and modified external fields \(h_i + h_i', i \in V \setminus S\), where \(h_i' = \sum_{(i,j) \in E, j \in S} J_{ij} x_j\) is the additional external field on node \(i\) induced by fixing \(X_S = x_S\).

### C. Random Graphs

A random graph is a graph generated from a prior distribution over the set of all possible graphs with a given number of nodes. Let \(\chi_p\) be a function on graphs with \(p\) nodes and let \(C\) be a constant. We say \(\chi_p \geq C\) almost always for a family of random graphs indexed by \(p\) if \(P(\chi_p \geq C) \to 1\) as \(p \to \infty\). Similarly, we say \(\chi_p \to C\) almost always for a family of random graphs if \(\forall \epsilon > 0, P(|\chi_p - C| > \epsilon) \to 1\) as \(p \to \infty\). This is a slight variation of the definition of almost always in \([1]\).

The Erdős-Rényi random graph \(G(p, \frac{c}{p})\) is a graph on \(p\) nodes in which the probability of an edge being in the graph is \(\frac{c}{p}\) and the edges are generated independently. We note that, in this random graph, the average degree of a node is \(c\). In this paper, when we consider random graphs, we only consider the Erdős-Rényi random graph \(G(p, \frac{c}{p})\).
**D. High-Dimensional Structure Learning**

In this paper, we are interested in inferring the structure of the graph $G$ associated with an MRF $(X, G)$. We will assume that $P(x) > 0, \forall x$, and $G$ will refer to the corresponding unique minimal I-map. The goal of structure learning is to design an algorithm that, given $n$ i.i.d. samples $\{X^{(k)}\}_{k=1}^n$ from the distribution $P$, outputs an estimate $\hat{G}$ which equals $G$ with high probability when $n$ is large. We say that two graphs are equal when their node and edge sets are identical.

In the classical setting, the accuracy of estimating $G$ is considered only when the sample size $n$ goes to infinity while the random vector dimension $p$ is held fixed. This setting is restrictive for many contemporary applications, where the problem size $p$ is much larger than the number of samples. A more suitable assumption allows both $n$ and $p$ to become large, with $n$ growing at a slower rate than $p$. In such a case, the structure learning problem is said to be high-dimensional.

An algorithm for structure learning is evaluated both by its computational complexity and sample complexity. The computational complexity refers to the number of computations required to execute the algorithm, as a function of $n$ and $p$. When $G$ is a deterministic graph, we say the algorithm has sample complexity $f(p)$ if, for $n = O(f(p))$, there exist constants $c$ and $\alpha > 0$, independent of $p$, such that $\Pr(\hat{G} = G) \geq 1 - \frac{c}{p^\alpha}$ for all $P$ which are Markov with respect to $G$. When $G$ is a random graph drawn from some prior distribution, we say the algorithm has sample complexity $f(p)$ if the above is true almost always. In the high-dimensional setting $n$ is much smaller than $p$. In fact, we will show that, for the algorithms described in this paper, $f(p) = \log p$.

For ease of exposition, we will first consider the computational complexity of an algorithm assuming that the distribution $P(x)$ is known. For a deterministic graph, when $P(x)$ is known, we say that an algorithm is correct if $\hat{G} = G$. For a graph drawn from a distribution, such as the Erdős-Rényi graph, when $P(x)$ is known, we say that an algorithm is correct if $\hat{G} = G$ almost always. Instead of the distribution $P(x)$, if we only have access to $n$ i.i.d. samples drawn from $P(x)$, then certain conditional independence tests have to be performed using empirical estimates of certain conditional probabilities. As can be easily seen later, the complexity of computing each such empirical estimate is $n$, therefore, the computational complexity in this case is simply $n$ times the computational complexity when $P(x)$ is known.

**III. The Basic Algorithm and Its Computational Complexity**

For ease of exposition, we consider the computational and sample complexities separately. Further, as we will see, the the evaluation of sample complexity follows from standard concentration bounds used in estimation theory. However, the computational complexity is a function of the structure of the MRF and is the main focus of the paper. In this section, we assume that the distribution $P(x)$ is known and therefore, the sample complexity is irrelevant. Given the distribution, our goal is to infer the structure of $P$ using an algorithm with low computational complexity.

Learning the structure of a graph is equivalent to learning if there exists an edge between every pair of nodes in the graph. Therefore, we would like to develop a test to determine if there exists an edge between two nodes or not. First, consider a pair of nodes $i$ and $j$ which do not have an edge between them. This means that there exists a set of nodes $S$ such that the removal of $S$ from the graph separates $i$ and $j$. From the global Markov property, this implies that $X_i \perp X_j \mid X_S$. However, as the following example shows, the converse is not true in general: in other words, if there exists an $S$ such that $X_i \perp X_j \mid X_S$, it does not imply that there is no edge between $i$ and $j$. 
Example 3.1: Consider the Ising model

\[ P(x) = \frac{1}{Z} \exp(J_{12}x_1x_2 + J_{23}x_2x_3 + J_{13}x_1x_3), \quad x_i \in \{\pm 1\}. \]

Let \( J_{12} = 1, J_{13} = -J_{23} = J > 0 \). The graph for this distribution is a triangle connecting all the three nodes. When \( J \) satisfies the equation \( e^J + e^{-J} = 2e^2 \), it is easy to verify that \( X_1, X_2 \) are marginally independent, i.e., \( X_1 \perp X_2 | X_S \) for \( S = \emptyset \). If we interpret \( J_{ij} > 0 \) as a positive dependence between \( X_i \) and \( X_j \), then the fact that \( X_1 \perp X_2 \) is the result of the positive dependence between \( X_1 \) and \( X_2 \) over edge \((1, 2)\) being canceled by the negative dependence over path \((1, 3, 2)\). In order to detect the existence of edge \((1, 2)\), we have to “block” all paths between nodes 1 and 2, other than the direct edge. In particular, we define \( T = \{3\} \), and note that \( \forall S, X_1 \not\perp X_2 | X_S, X_T \). In the next lemma, we will see that this basic idea generalizes to a test for the detection of an edge in the graph: whenever \( i \) and \( j \) have an edge between them, we can always find a \( T \) such that \( \forall S, X_1 \not\perp X_2 | X_S, X_T \).

The following lemma provides a test for checking if there exists an edge between two nodes.

Lemma 3.2: Consider two nodes \( i \) and \( j \) in \( G \). Then, \((i, j) \not\in E \) if and only if \( \exists S, \forall T, X_i \perp X_j | X_S, X_T \).

Proof: Recall from the definition of the minimal I-map that \((i, j) \not\in E \) if and only if \( X_i \perp X_j | X_{V \setminus \{i,j\}} \). Therefore, the statement of the lemma is equivalent to

\[ I(X_i; X_j | X_{V \setminus \{i,j\}}) = 0 \iff \min_S \max_T I(X_i; X_j | X_S, X_T) = 0, \]

where \( I(X_i; X_j | X_S) \) denotes the mutual information between \( X_i \) and \( X_j \) conditioned on \( X_S \), and we have used the fact that \( X_i \perp X_j | X_S \) is equivalent to \( I(X_i; X_j | X_S) = 0 \). Notice that

\[ \min_S \max_T I(X_i; X_j | X_S, X_T) = \min_{S,T} \max_{T \supset S} I(X_i; X_j | X_T) \]

and \( \max_{T \supset S} I(X_i; X_j | X_T) \) is an increasing function in \( S \). The minimization over \( S \) is achieved at \( S = V \setminus \{i, j\} \), i.e.,

\[ I(X_i; X_j | X_{V \setminus \{i,j\}}) = \min_S \max_T I(X_i; X_j | X_S, X_T). \]

According to the lemma, if we allow \( S \) and \( T \) to be any subsets of the node set \( V \), we can recover each edge correctly for any graph. But the computational complexity for this test is very high. If we use mutual information to test for conditional independence, then to compute mutual information \( I(X_i; X_j | X_S, X_T) \), we need to sum over every possible choice of \((x_i, x_j, x_S, x_T)\) and therefore, the computational complexity for a single conditional independence test is \( O(|\mathcal{X}|^2 + |S \cup T|) \). If \( S \) and \( T \) are arbitrary subsets of the nodes, then \(|S \cup T|\) can be comparable to \( p \), and the complexity of a single conditional independence test will be exponential in \( p \). Further, this test has to be performed over possible choices of \( S \) and \( T \) for every pair of nodes. The number of possible choices for \( S \) and \( T \) is also exponential in \( p \). The main focus of the paper is to show that we can restrict the sizes of \( S \) and \( T \). In particular, if we require the size of \( S \) to be at most \( D_1 \) and the size of \( T \) to be at most \( D_2 \), the corresponding computational complexity (of performing the test for the existence of an edge for every pair of nodes) is given by \( O \left( (|\mathcal{X}|^2)^{D_1} + D_2 \right) \).

Thus, the goal is to show that \( D_1, D_2 \) can be small in many cases of interest.

Define the separator between a pair of nodes to be the smallest set that separates all the non-direct paths between them. It is easy to see that it suffices to choose \( D_1 \) to be the maximum separator size between non-neighbor nodes and \( D_2 \) to be the maximum separator size between neighbor nodes to recover the
graph correctly: for each pair of non-neighbor nodes \(i, j\), there exists some separator \(S\) of size at most \(D_1\) separating them, so \(X_i, X_j\) are independent conditioned on \(X_S\); for each pair of neighbor nodes \(i, j\) and any given \(S\), there exist some separator \(T\) of size at most \(D_2\) blocking all paths other than the direct edge between them, so \(X_i, X_j\) are not independent conditioned on \(X_S, X_T\).

If we relax the requirement of \(X_i, X_j\) being exactly independent conditioned on \(X_S\), we might be able to choose \(D_1, D_2\) to be much smaller than the sizes of the exact separators. In some cases we will see, though the exact separator size can be large, non-neighbor nodes might be approximately independent conditioned on only a small subset of the separator. This observation motivates us to consider the following test which includes a small constant \(\epsilon_2\) to quantify the level of approximate independence.

**Algorithm 1** \(\text{Exact}_\text{Alg}(D_1, D_2, \epsilon_2)\)

```python
for i, j ∈ V do
    if ∃S with |S| ≤ D_1, ∀T with |T| ≤ D_2, ∆(X_i; X_j|X_S, X_T) ≤ \frac{\epsilon_2}{2}
        then
            (i, j) ∉ E
        else
            (i, j) ∈ E
    end if
end for
```

Here we call the algorithm \(\text{Exact}_\text{Alg}\) as we assume \(P(x)\) is known. We use \(\Delta\) to represent a conditional independence test, i.e., \(\Delta(X_i; X_j|X_S, X_T) = 0\) if and only if \(X_i \perp X_j|X_S, X_T\). When \(\Delta(X_i; X_j|X_S, X_T)\) is small, we say that \(X_i, X_j\) are approximately independent conditioned on \(X_S, X_T\). In this paper, we use two types of conditional independence tests:

- **Mutual Information Test (I-test):** \(\Delta(X_i; X_j|X_S, X_T) = I(X_i; X_j|X_S, X_T)\). The
- **Probability Test (P-test):** \(\Delta(X_i; X_j|X_S, X_T) = \max_{x_i, x_j, x_S, x_T} [P(x_i|x_j, x_S, x_T) − P(x_i)|x_j', x_S, x_T]\).

Later on, we will see that, in some cases, it is better to use one of the above tests than the other for lowering the sample complexity. For clarity, when we specifically use the I-test (or the P-test), we index the corresponding algorithm or assumption by I (or P). For example, \(\text{Exact}_\text{Alg}_P\) would mean that we are implementing the algorithm with the P-test.

In the rest of the section, we demonstrate the power of this algorithm by identifying some important cases where \(D_1, D_2\) can be chosen to be small while guaranteeing the correctness of the algorithm. Therefore, in these cases the structure can be learned with low complexity.

### A. Bounded Degree Graphs

When the graph has maximum degree \(d\), for any non-neighbor nodes \(i, l\), set \(S = N_i\) separates \(i\) and \(l\) and the size of \(S\) is at most \(d\), while for any neighbor nodes \(i, j\), set \(T = N_i \setminus j\) blocks all the paths other than the direct edge between \(i, j\) and the size of \(T\) is at most \(d − 1\). So we can choose \(D_1 = d, D_2 = d − 1\) for our algorithm, thus trivially leading to the following theorem.

**Theorem 3.3:** For MRFs on a graph with maximum degree \(d\), \(\text{Exact}_\text{Alg}(d, d − 1, 0)\) recovers the graph correctly.
B. Ferromagnetic Ising Models

For Ising models, each term $J_{ij}x_ix_j$ in the distribution corresponds to an edge $(i, j)$ in the graph, i.e., $(i, j)$ is an edge if and only if $J_{ij} \neq 0$. As in the example, we intuitively interpret $J_{ij} > 0$ as a positive dependence between $X_i$ and $X_j$ along the edge $(i, j)$, and $J_{ij} < 0$ as a negative dependence between $X_i$ and $X_j$ along the edge $(i, j)$. We can similarly consider the dependence along a path between nodes $i, j$. If there are even number of negative edges on the path, we say $X_i$ and $X_j$ are positively dependent along the path; otherwise, we say $X_i$ and $X_j$ are negatively dependent along the path. Under the ferromagnetic assumption, all $J_{ij}$'s are nonnegative, therefore for any pair of nodes $i, j$, $X_i$ and $X_j$ are positively dependent along any path between them. In particular, when $i, j$ are neighbors, $X_i$ and $X_j$ are positively dependent along edge $(i, j)$ and the non-direct paths between $i, j$ can only make $X_i, X_j$ more positively dependent. The situation in the example where the positive dependence is canceled by another negatively dependent path cannot happen here. So we do not need to use $T$ to block the other paths and can choose $D_2 = 0$ in our algorithm.

**Theorem 3.4:** For ferromagnetic Ising models, let $D_1$ be maximum separator size between non-neighbor nodes. $\text{Exact}_{\text{Alg}}(D_1, 0, 0)$ recovers the graph correctly.

**Proof:** See Section V.

C. Ising Models on Bounded Degree Graphs with Correlation Decay and Large Girth

In this subsection, we will assume that the maximum degree of the graph is $d$ and the edge coefficients are bounded, i.e., $J_{\text{min}} \leq |J_{ij}| \leq J_{\text{max}}$. As for the general bounded degree graph case, we can use algorithm $\text{Exact}_{\text{Alg}}(d, d − 1, 0)$ to correctly recover the graph. However, under the additional assumptions that the Ising model has the correlation decay property and the graph has large girth, we want to show that the low complexity algorithm $\text{Exact}_{\text{Alg}}(1, 0, \epsilon_2)$ for some constant $\epsilon_2$ is enough to recover the graph. Before we state the main result, we present the intuition behind it.

Correlation decay is a property of MRFs which says that, for any pair of nodes $i, j$, the correlation of $X_i$ and $X_j$ decays with the distance between $i, j$. When a MRF has correlation decay, the correlation of $X_i$ and $X_j$ is mainly determined by the short paths between nodes $i, j$, and the contribution due to the long paths is negligible. It is known that when $J_{\text{max}}$ is small compared with $d$, the Ising model has correlation decay. Moreover, the girth of a graph is defined as the length of the shortest cycle in the graph, and large girth implies that there is no short cycle in the graph.

When the Ising model is in the correlation decay regime and the girth of the graph is large in terms of the correlation decay parameters, there is at most one short path between any pair of non-neighbor nodes, and no short paths other than the direct edge between any pair of neighboring nodes. Naturally, we can use $S$ of size 1 to approximately separate any pair of non-neighbor nodes and do not need $T$ to block the other paths for neighbor nodes as the correlations are mostly due to the direct edges. Therefore we choose $D_1 = 1, D_2 = 0$ in the algorithm for this case. For simplicity, let $\beta = \frac{4J_{\text{max}}d}{(d-1)\tanh J_{\text{max}}}$ and $\alpha = (d-1)\tanh J_{\text{max}}$, which are parameters related to the correlation decay property we will see later.

**Theorem 3.5:** Assume $(d-1)\tanh J_{\text{max}} < 1$. Let

$$g_1 \triangleq 2\ln \left[\beta \left(\frac{1}{\alpha} \lor \ln 2\right)\right] \ln \frac{1}{\alpha},$$
where $A = \frac{1}{1800}(1 - e^{-4J_{\min}})e^{-8dJ_{\max}}$, and let $\epsilon_2 = 48Ae^{4dJ_{\max}}$. Then if the girth $g > g_1$, $Exact\_Alg(1, 0, \epsilon_2)$ recovers the graph correctly.

Proof: See Section VI.

The complexity of the above algorithm is $O(p^3)$, which in general is much better than that of algorithm $Exact\_Alg(d, d - 1, 0)$. The complexity can be further reduced. We note that in testing whether there is an edge between nodes $i$ and $j$, we can restrict ourselves to $S$ and $T$ which are subsets of $N_i$ (the neighbors of $i$) to separate or block the short paths between $i, j$. Therefore, if somehow we can identify a set $L_i \subset V$ which contains $N_i$, then we can modify our algorithm to search $S$ and $T$ only in $L_i$ rather than over all the subsets of $V \setminus \{i, j\}$, which reduces the complexity of searching. For this purpose, consider the following metric to test for the correlation of $X_j$ with $X_i$:

$$\max_{x_i,x_j,x_j'} |P(x_i|x_j) - P(x_i|x_j')|.$$ 

With some abuse of terminology, we will call this metric, the correlation of $X_j$ with $X_i$. Under correlation decay and large girth assumptions, we can in fact show that for any node $j \in N_i$, the correlation of $X_j$ with $X_i$ is lower bounded by some constant $\epsilon_1$. This motivates us to generalize our algorithm by including a correlation test as a preprocessing step, i.e., we only include in $L_i$ nodes whose correlation with $X_i$ is above $\epsilon_1$. Clearly such an $L_i$ contains $N_i$.

Algorithm 2 $Exact\_Alg(D_1, D_2, \epsilon_1, \epsilon_2)$

for $i \in V$ do
  $L_i = \{j \in V \setminus i, \max_{x_i,x_j,x_j'} |P(x_i|x_j) - P(x_i|x_j')| > \frac{\epsilon_1}{2}\}$.
  for $j \in L_i$ do
    if $\exists S \subseteq L_i$ with $|S| \leq D_1, \forall T \subseteq L_i$ with $|T| \leq D_2, \Delta(X_i; X_j|X_S, X_T) \leq \frac{\epsilon_2}{2}$ then
      $j \notin N_i$
    else
      $j \in N_i$
    end if
  end for
end for

We have the following result when using the new algorithm.

Theorem 3.6: Assume $(d - 1)\tanh J_{\max} < 1$, $g_1, \epsilon_2$ are defined as above. Let

$$g_2 = \ln \left[ \beta \left( \frac{1}{B} \lor \ln 2 \right) \right] + 1,$$

where $B = \frac{1}{1800}(1 - e^{-4J_{\min}})$. Let $\epsilon_1 = 48A$. Then if the girth $g > \max\{g_1, g_2\}$, $Exact\_Alg(1, 0, \epsilon_1, \epsilon_2)$ recovers the graph correctly.

Proof: See Section VI.

To compute $L_i$ for all the nodes, the complexity is $O(p^2)$. As the Ising model has correlation decay, we will show that the size of each $L_i$ is upper bounded by some constant, independent of $p$, so the complexity of the conditional independence tests for nodes in $L_i$ is only $O(1)$ for each node. Therefore, the total complexity is $O(p^2 + p) = O(p^2)$.
D. Ising Models on Random Graph $G(p, \frac{c}{p})$ with Correlation Decay

In this subsection, we will assume that $J_{\text{min}} \leq |J_{ij}| \leq J_{\text{max}}$. As before, we will first present the intuition behind the result before stating the main result of this subsection. For the family of random graphs $G(p, \frac{c}{p})$, the parameter $c$ is the average degree of each node in the graph. When $J_{\text{max}}$ is small compared to $c$, it is known from prior work that an Ising model defined over a graph drawn at random from $G(p, \frac{c}{p})$ satisfies the following two properties with high probability:

- The correlation between any pair of nodes $X_i$ and $X_j$ is insignificant only if the distance between them in the graph is larger than some threshold. We say that a path in the graph is long if the length of the path is above this threshold. Otherwise, we say that the path is short.
- The number of short paths between non-neighboring nodes is at most 2 and the number of short paths, other than the direct edge, between neighboring nodes is at most 1.

From these observations, it is natural to conclude that we should choose $D_1 = 2, D_2 = 1$ for our algorithm to recover the graph correctly with high probability. The computational complexity for this algorithm is $O(p^5)$.

**Theorem 3.7:** Assume $c \tanh J_{\text{max}} < 1$. Then for some $\epsilon_2 = \Omega(1)$, Exact_Alg($2, 1, \epsilon_2$) recovers the graph correctly almost always.

**Proof:** See Section VII.

If we further assume that the Ising model is ferromagnetic, by the result of section III-B, we can again choose $D_2 = 0$ and immediately extend the above result to show that Exact_Alg($2, 0, \epsilon_2$) for some constant $\epsilon_2$ recovers the graph correctly with high probability. The computational complexity is $O(p^4)$. But we show that we can do better in this case by first showing that the correlation between neighbor nodes is lower bounded by some $\epsilon_1 = \Omega(1)$. Then, we can use a preprocessing step to estimate the neighborhood of each node before we apply the main algorithm. The result is stated in the next theorem.

**Theorem 3.8:** Assume $c \tanh J_{\text{max}} < 1$ and the Ising model is ferromagnetic. Then for some $\epsilon_1 = \Omega(1)$, $\epsilon_2 = \Omega(1)$, Exact_Alg($2, 0, \epsilon_1, \epsilon_2$) recovers the graph correctly almost always.

**Proof:** See Section VII.

We will see that the complexity of the algorithm is only $O(p^2)$.

IV. CORRECT RECOVERY FOR GENERAL DISCRETE MRFs

In this section, we provide conditions under which the algorithms presented in the previous section recover the graph structure correctly for general discrete MRFs. Then to prove the theorems for Ising models presented in the previous section, it is enough to show that the assumptions are satisfied by the corresponding Ising models. This will be done in subsequent sections.

The following assumptions are stated with respect to some given $D_1, D_2$. The theorems presented after the assumptions simply state that, if our algorithms are applied to a given $P(x)$ using these $D_1$ and $D_2$, then the algorithm will find the graph structure of the MRF correctly. As seen in the previous section, for many MRFs of interest, it is possible to identify appropriate values of $D_1$ and $D_2$ such that the following assumptions are satisfied. The assumptions are stated next.
Assumption A1: $\forall (i, j) \in E$

$$\min_{S \subseteq V \setminus \{i, j\}, |S| \leq D_1} \max_{T \subseteq V \setminus \{i, j\}, |T| \leq D_2} \Delta(X_i; X_j | X_S, X_T) > \epsilon_2,$$

and $\forall (i, j) \notin E$

$$\min_{S \subseteq V \setminus \{i, j\}, |S| \leq D_1} \max_{T \subseteq V \setminus \{i, j\}, |T| \leq D_2} \Delta(X_i; X_j | X_S, X_T) \leq \frac{\epsilon_2}{4}.$$

Assumption A1 says that, when performing the conditional independence test, the dependencies between neighbor nodes (edges) are clearly non-zero, and the dependencies between non-neighbor nodes are sufficiently small. So we can perform a threshold test to separate the edges and non-edges. The above assumption immediately lead to the following theorem.

**Theorem 4.1:** If for some $D_1, D_2$ assumption A1 holds, then $\text{Exact}_\text{Alg}(D_1, D_2, \epsilon_2)$ recovers the graph correctly.

Next, we state another assumption which helps in establishing the correctness of the algorithm with the preprocessing step mentioned in the previous section.

**Assumption A2:** $\forall (i, j) \in E$, $\max_{x_i, x_j, x'_j} |P(x_i | x_j) - P(x_i | x'_j)| > \epsilon_1$.

Assumption A2 says that the correlations between neighbor nodes are lower bounded so that we can incorporate a preprocessing step in the algorithm which may reduce the total computational complexity if the correlation between the nodes decays fast. Thus, we have the following theorem.

**Theorem 4.2:** If for some $D_1, D_2$ assumption A1, A2 hold, then $\text{Exact}_\text{Alg}(D_1, D_2, \epsilon_1, \epsilon_2)$ recovers the graph correctly.

In the rest of this section, we will present some discussion in preparation for the proofs in the next section.

### A. Sparse Graphs with Correlation Decay

One important class of MRFs to which we want to apply our algorithms is the class of MRFs on sparse graphs with correlation decay. By a sparse graph, we mean that the number of short paths between any pair of nodes is small. As explained in the previous section, due to the correlation decay property, the long paths contribute little in computing $\Delta$, so we only require $S$ and $T$ to separate the short paths between nodes in the conditional independence test. Therefore, applying our algorithms with small $D_1, D_2$ is sufficient to recover the graph correctly. The notion of a short path depends on the type of correlation decay we have and we will consider correlation decay properties for both bounded degree graphs and random graph $G(p, \frac{\epsilon}{p})$ in this paper. For later use, we give the following definition of a set that disconnects all short paths between a pair of nodes.

**Definition 4.3:** We say that $S$ is a $\gamma$-separator of nodes $i, j$ if all non-direct paths of length at most $\gamma$ between $i, j$ are disconnected when $S$ is removed from the graph. In this definition, the length of a path is the number of edges on the path.
B. Conditional Independence Test \( \Delta \)

In the previous section, we have seen that the conditional independence test \( \Delta \) can take two forms: the \( P \)-test and the \( I \)-test. If we are given \( P(x) \), then to establish the correctness of our algorithms, we can use either test. However, later we will see that the sample complexity of the algorithm depends on both the model we work with and the form of \( \Delta \) we use. In particular, for Ising models on bounded degree graphs, we will should use \( P \)-test and for Ising models on random graphs, we use the \( I \)-test. Keeping this mind, in the next two sections, for simplicity, we only prove that Assumption \( A1 \) is satisfied for the \( P \)-test in bounded degree graphs and is satisfied for the \( I \)-test in random graphs.

V. Ferromagnetic Ising Models

Ferromagnetic Ising models are Ising models in which all the edge coefficients \( J_{ij} \) are nonnegative. One important property of ferromagnetic Ising models is association, which characterizes the positive dependence among the nodes.

Definition 5.1: [7] We say a collection of random variables \( X = (X_1, X_2, \ldots, X_n) \) is associated, or the random vector \( X \) is associated, if

\[
\text{Cov}(f(X), g(X)) \geq 0
\]

for all nondecreasing functions \( f \) and \( g \) for which \( E f(X), E g(X), E f(X)g(X) \) exist.

Proposition 5.2: [9] The random vector \( X \) of a ferromagnetic Ising model (possibly with external fields) is associated.

Association is a stronger property than positive correlation between random variables. It allows us to justify the intuition that, for a pair of neighbor nodes, the paths other than the direct edge between them only make them more positively dependent. In particular, we have the following lemma.

Lemma 5.3: \( \forall (i, j) \in E, S \subset V \setminus \{i, j\} \) and \( \forall x_S \),

\[
\max_{x_i, x_j, x'_j} |P(x_i|x_j, x_S) - P(x_i|x'_j, x_S)| \geq \frac{1}{16}(1 - e^{-4J_{\text{min}}})e^{-4|N_S|J_{\text{max}}}.
\]

Proof: See appendix.

From this lemma, we see that when \( i, j \) are neighbor nodes, for any set of nodes \( S \), the \( P \)-test is always larger than 0 for any \( x_S \), i.e., \( X_i \) and \( X_j \) are not independent conditioned \( X_S \). Therefore, if we fix \( T = \emptyset \) and allow the size of \( S \) to be as large as the maximum separator size between non-neighbor nodes, the algorithm recovers the graph correctly, which leads to Theorem 3.4.

We also note that assumption \( A2 \) is automatically satisfied by the ferromagnetic Ising model for some constant \( \epsilon_1 \). This result immediately follows from Lemma 5.3 by setting \( S = \emptyset \).

Corollary 5.4: \( \forall (i, j) \in E \),

\[
\max_{x_i, x_j, x'_j} |P(x_i|x_j) - P(x_i|x'_j)| \geq \frac{1}{16}(1 - e^{-4J_{\text{min}}}) \triangleq \epsilon_1.
\]

Hence we can add a preprocessing step and apply algorithm \( \text{Alg}(d, 0, \epsilon_1, \epsilon_2) \) to recover the graph.
VI. ISING MODELS ON BOUNDED DEGREE GRAPHS

In this section, we will assume that the maximum degree of the graph is \(d\) and the edge coefficients are bounded, i.e., \(J_{\min} \leq |J_{ij}| \leq J_{\max}\). We have already seen that in Theorem 3.3 that for such Ising models, our algorithm with \(D_1 = d, D_2 = d - 1\) recovers the graph correctly. The computational complexity of this algorithm is \(O(p^{2d+1})\). We notice that when the graph is sparse, this complexity can be greatly reduced. For example, consider Ising models on a tree with maximum degree \(d\). As there is exactly one path between each pair of nodes, the algorithm \(\text{Exact\_Alg}(1, 0, 0)\) is sufficient to learn the structure, and the complexity is only \(O(p^3)\). This suggests that the maximum degree of a graph is not always the right parameter to quantify the complexity of learning the graph. In particular, for Ising models on sparse graphs with correlation decay, in which the number of short paths between any pair of nodes is small, the complexity of the algorithm can be made much smaller than \(O(p^{2d+1})\).

For Ising models on bounded degree graphs (not necessarily zero-field), we have the following strong spatial mixing result [15].

**Proposition 6.1:** Consider any \(i \in V, \Lambda \subset V \setminus i\) and let \(\sigma\), \(\eta\) be two configurations of \(\Lambda\). (A configuration of \(\Lambda\) is a set of values taken by the random variables in \(\Lambda\).) Let \(\Delta = \{v \in \Lambda : \sigma_v \neq \eta_v\}\) and \(l = d(i, \Delta)\) be the length of the shortest path from \(i\) to \(\Delta\). For all \(x_i\),

\[
e^{-4J_{\max}d[(d - 1) \tanh J_{\max}]^{d-1}} \leq \frac{P(x_i|\sigma)}{P(x_i|\eta)} \leq e^{4J_{\max}d[(d - 1) \tanh J_{\max}]^{d-1}}.
\]

This proposition implies that, in the correlation decay regime \((d - 1) \tanh J_{\max} < 1\), the Ising model has exponential correlation decay, i.e., the correlation between a pair of nodes decays exponentially with their distance.

**Corollary 6.2:** Assume \((d - 1) \tanh J_{\max} < 1\). \(\forall i, j \in V, d(i, j) = l\), then for any set \(S\) and \(\forall x_i, x_j, x_j', x_S\),

\[
|P(x_i|x_j, x_S) - P(x_i|x_j', x_S)| \leq 4J_{\max}d[(d - 1) \tanh J_{\max}]^{d-1} \triangleq \beta \alpha^l,
\]

where \(\beta = \frac{4J_{\max}d}{(d - 1) \tanh J_{\max}}\) and \(\alpha = (d - 1) \tanh J_{\max}\).

**Proof:** For some given \(x_i, x_j, x_j', x_S\), w.l.o.g. assume \(P(x_i|x_j, x_S) \geq P(x_i|x_j', x_S)\). Applying the above proposition with \(\Lambda = \{j\} \cup S, \Delta = \{j\}\),

\[
|P(x_i|x_j, x_S) - P(x_i|x_j', x_S)| \leq 1 - \frac{P(x_i|x_j', x_S)}{P(x_i|x_j, x_S)} \leq 1 - e^{-4J_{\max}d[(d - 1) \tanh J_{\max}]^{d(i, j) - 1}} \\
\leq 4J_{\max}d[(d - 1) \tanh J_{\max}]^{d(i, j) - 1}.
\]

The above corollary suggests that, when \(d(i, j)\) is sufficiently large, then \(X_j\) does not significantly influence \(X_i\). This suggests that we can set a threshold \(h\) such that, for any given node, the influence of other nodes that are farther than \(h\) is sufficiently small. In other words, if for every pair of non-neighbors \(i\) and \(j\), there exists a set \(S\) of size \(D_1\) which is a \(h\)-separator, then conditioned on \(X_S\), \(X_i\) and \(X_j\) are approximately independent. Similarly, if for every pair of neighbors \(i\) and \(j\), there exists a set \(T\) of size \(D_2\) which is a \(h\)-separator, then conditioned on \(X_T\), the dependence between \(X_i\) and \(X_j\) is dominated by the edge between them. The following theorem shows that assumption \(A_{1,p}\) is satisfied for graphs in which we can find such a \(D_1\) and \(D_2\).
Theorem 6.3: Assume \((d - 1) \tanh J_{\text{max}} < 1\). Fix \(D_1, D_2\). Let
\[
h = \frac{\ln \left[\frac{\beta (\frac{1}{4} \vee \ln 2)}{\ln \frac{1}{\alpha}}\right]}{1 + \frac{\ln \left(\frac{1}{\alpha}\right)}{\ln \frac{1}{\alpha}}},
\]
where \(A = \frac{1}{1890} (1 - e^{-4J_{\text{min}}}) e^{-8(D_1 + D_2)dJ_{\text{max}}}\), and let \(\epsilon_2 = 48AE(4(D_1 + D_2)dJ_{\text{max}})\). Assume the maximum sizes of \(h\)-separator between non-neighbor and neighbor nodes are \(D_1, D_2\). Then \(\forall (i, j) \in E\),
\[
\begin{align*}
&\min_{S \subseteq V \setminus \{i, j\}, |S| \leq D_1} \max_{T \subseteq V \setminus \{i, j\}, |T| \leq D_2} \max_{x_i, x_j, x_s, x_T} |P(x_i | x_j, x_s, x_T) - P(x_i | x'_j, x_s, x_T)| > \epsilon_2, \\
&\forall (i, j) \notin E,
\end{align*}
\]
and \(\forall (i, j) \notin E\),
\[
\begin{align*}
&\min_{S \subseteq V \setminus \{i, j\}, |S| \leq D_1} \max_{T \subseteq V \setminus \{i, j\}, |T| \leq D_2} \max_{x_i, x_j, x'_s, x_T} |P(x_i | x_j, x_s, x_T) - P(x_i | x'_j, x_s, x_T)| \leq \frac{\epsilon_2}{4}.
\end{align*}
\]

Proof: See appendix.

From this theorem, we know that the algorithm \(\text{Exact}_{-}\text{Alg}(D_1, D_2, \epsilon_2)\) recovers the graph correctly. Now, Theorem 3.5 follows easily. In particular, if the girth of the graph is \(g\), then the second shortest path between any pair of nodes is no less than \(\frac{g}{2}\). Applying the above theorem, we see that assumption A1 holds for \(D_1 = 1, D_2 = 0\) under the assumption in Theorem 3.5. Therefore, by Theorem 4.1, we have proved Theorem 3.5. The computational complexity for this algorithm is \(O(p^3)\), which is the same as the complexity for learning a tree structure. This is not a coincidence as a large girth graph indeed looks like a tree locally.

To prove Theorem 3.6, we need to show that the correlations between neighbor nodes are lower bounded by some \(\epsilon_1\) as required by assumption A2. In fact, this is true as there is no short path other than the direct path between neighbor nodes, so the correlation over the direct path dominates in computing the total correlation and itself is lower bounded by some constant. We note that the length of the second shortest path between neighbor nodes is no less than \(g - 1\).

Lemma 6.4: Assume \((d - 1) \tanh J_{\text{max}} < 1\). If
\[
g > \frac{\ln \left[\frac{\beta (\frac{1}{4} \vee \ln 2)}{\ln \frac{1}{\alpha}}\right]}{1 + \frac{\ln \left(\frac{1}{\alpha}\right)}{\ln \frac{1}{\alpha}}} + 1 \triangleq g_2,
\]
where \(A = \frac{1}{1890} (1 - e^{-4J_{\text{min}}}) e^{-8(D_1 + D_2)dJ_{\text{max}}}\). Let \(\epsilon_1 = 48A\). \(\forall (i, j) \in E\), we have
\[
\max_{x_i, x_j, x'_j} |P(x_i | x_j) - P(x_i | x'_j)| > \epsilon_1.
\]

Proof: The proof is similar as Theorem 6.3 with \(D_1 = D_2 = 0\) and is omitted. \(\square\)

Theorem 3.6 with \(\Delta\) being the \(P\)-test is then a simple corollary of Theorem 3.5 with the above lemma. For algorithm \(\text{Exact}_{-}\text{Alg}(1, 0, \epsilon_1, \epsilon_2)\), the correlation test runs over each pair of nodes and the complexity of computing \(L_i\) for all the nodes is \(O(p^3)\). By the correlation decay property in Corollary 6.2, for any node \(i \in V\), if node \(j\) is of distance \(l_{\epsilon_1} = \frac{\ln \frac{\epsilon_1}{2}}{\ln \frac{1}{\alpha}}\) away from \(i\), we have
\[
\max_{x_i, x_j, x'_j} |P(x_i | x_j) - P(x_i | x'_j)| < \beta a^{l_{\epsilon_1}} \leq \frac{\epsilon_1}{4}.
\]
Therefore, \(L_i\) only includes nodes within distance \(l_{\epsilon_1}\) from \(i\) and the size \(|L_i| \leq d^{l_{\epsilon_1}}\) as the maximum degree is \(d\); i.e., \(L = \max_i |L_i| \leq d^{l_{\epsilon_1}}\). As this upper bound on \(L\) is a constant independent of \(p\), the number of choices over \(j, S, T\) among \(L_i\) is only a constant, so the complexity of the conditional independent tests after the preprocessing step is only \(O(1)\) for each node. Therefore, the total complexity is \(O(p^3 + p) = O(p^3)\).
VII. ISING MODELS ON RANDOM GRAPHS

In this section, we assume the graph $G$ is generated from the prior $G(p, c)$ in which each edge is in $G$ with probability $c$ and the average degree for each node is $c$. For each edge, assume $J_{\min} \leq |J_{ij}| \leq J_{\max}$. For this random graph, the maximum degree scales as $O\left(\frac{\ln p}{\ln \ln p}\right)$ with high probability [1]. Thus, we cannot use the results for bounded degree graphs even though the average degree remains bounded as $p \to \infty$.

It is known that when $J_{\max}$ is small compared with the average degree $c$, the random graph is in the correlation decay region. More specifically, fix any $\epsilon \in (0, 1)$ and define $\gamma_p := \log p \left(3 + \frac{\epsilon}{\log c}\right)$. The following theorem shows that nodes of distance $\gamma_p$ away from node $i$ have diminishing impact on $i$. The proofs of all the theorems in this section can be found in the appendix.

**Theorem 7.1:** Assume $\alpha = c \tanh J_{\max} < 1$. Then, the following property is true almost always. Let $G$ be a graph generated from the prior $G(p, c)$. If $i, j$ are not neighbors in $G$ and $S$ is a $\gamma_p$-separator of $i, j$, then $\forall x_i, x_j, x_j', x_S,
\begin{align*}
|P(x_i|x_j, x_S) - P(x_i|x_j', x_S)| \leq B_{\gamma_p}(\tanh J_{\max})^{\gamma_p} = o(p^{-\kappa})
\end{align*}$
for all Ising models $P$ on $G$, where $\kappa = \frac{\log \alpha}{4 \log e}$.

We call a path short if it is shorter than $\gamma_p$ and the following theorem formally shows that the number of short paths between any pair of nodes is small.

**Theorem 7.2:** The size of the $\gamma_p$-separator between neighbor nodes is at most 1 and between non-neighbor nodes is at most 2 almost always.

The theorem suggests that we should pick $D_1 = 2, D_2 = 1$ in the algorithm. Together with the correlation decay property, we can show the following results about the I-test for neighbor and non-neighbor nodes in the random graph.

**Theorem 7.3:** $\forall i \in V, \forall j \notin N_i$, let $S$ contain a $\gamma_p$-separator between $i, j$ and $|S| \leq 3$, then almost always

$$I(X_i; X_j|X_S) = o(p^{-2\kappa}).$$

**Theorem 7.4:** $\forall i \in V, \forall j \in N_i$, let $T$ contain a $\gamma_p$-separator between $i, j$ and $|T| \leq 3$, then almost always

$$I(X_i; X_j|X_T) = \Omega(1).$$

Note that, in the above two theorems, $S$ or $T$ in fact plays the role of $S \cup T$ in the structure learning algorithm, but to avoid having to write $S \cup T$ many times, we have used $S$ in Theorem 7.3 since it deals with non-neighbor nodes and $T$ in Theorem 7.4 since it deals with neighbor nodes. Now Theorem 3.7 is a simple corollary of the above results with $\Delta$ being the I-test.

When the Ising model is ferromagnetic, the result for the random graph is similar to that of a deterministic graph. For each graph sampled from the prior distribution, the dependence over the edges are positive. If $i, j$ are neighbors in the graph, having additional paths between them makes them more positively dependent, so we do not need to block those paths with a set $T$ to detect the edge and set $D_2 = 0$. In fact, we can prove a stronger result for neighbor nodes than the general case.
**Theorem 7.5:** \( \forall i \in V, \forall j \in N_i, \) let \( S \) be any set with \( |S| \leq 2 \), then almost always
\[
I(X_i; X_j | X_S) = \Omega(1).
\]

Moreover, the same intuition allows us to show a lower bound on the correlation between neighbors so assumption \( A2 \) is also satisfied by some constant \( \epsilon_1 \). This result is formally stated in Corollary 5.4.

Putting all the above together, we complete the proof of Theorem 3.8 with \( \Delta \) being the \( I \)-test.

From Theorem 7.1 we know that if \( j \) is more than \( \gamma_p \) away from \( i \), the correlation between them decays as \( o(p^{-\alpha}) \). For any constant threshold \( \frac{p}{\epsilon} \), these far-away nodes are excluded from the candidate neighbor set \( L_i \) when \( p \) is large. It is shown in the proof of [11, Lemma 2.1] that for \( G(p, \frac{\epsilon}{p}) \), the number of nodes in the \( \gamma_p \)-ball around \( i \) is not big with high probability. We state this result in the following proposition.

**Proposition 7.6:** \( \forall i \in V, |B(i, \gamma_p)| = O(c^{\gamma_p} \log p) \) almost always, where \( B(i, \gamma_p) \) is the set of all nodes which are at most \( \gamma_p \) hops away from \( i \).

From this proposition, we get
\[
L = \max_i |L_i| \leq |B(i, \gamma_p)| = O(c^{\gamma_p} \log p) = O(p^{\frac{1}{\alpha^2}} \log p) = O(p^{\frac{1}{4}}).
\]

So the total complexity of algorithm \( \text{Exact\_Alg}(2, 0, \epsilon_1, \epsilon_2) \) is \( O(p^2 + pL^2) = O(p^2) \).

### VIII. Sample Complexity

In this section, we turn to the more realistic situation where we do not know the distribution but have \( n \) i.i.d. samples \( \{X^{(k)}\}_{k=1}^n \) from the distribution. Let \( \hat{P} \) be the empirical probability distribution and \( \hat{\Delta} \) be the conditional independence test, where the exact probabilities \( P \) are replaced by the corresponding empirical estimates \( \hat{P} \). Note that \( \hat{P} \) is a notation for any generic conditional probability estimate such as \( \hat{P}(x_i | x_j) \) and does not specifically refer to the entire joint distribution \( \hat{P}(x) \). The only change to our algorithms is to use the empirical tests instead of the exact tests.

**Algorithm 3** \( \text{Alg}(D_1, D_2, \epsilon_2) \)

\begin{algorithmic}
  \For {\( i, j \in V \)}
  \State \( (i, j) \notin E \iff \exists S \text{ with } |S| \leq D_1, \forall T \text{ with } |T| \leq D_2, \hat{\Delta}(X_i; X_j | X_S, X_T) \leq \frac{\epsilon_2}{2} \).
  \EndFor
\end{algorithmic}

**Algorithm 4** \( \text{Alg}(D_1, D_2, \epsilon_1, \epsilon_2) \)

\begin{algorithmic}
  \For {\( i \in V \)}
  \State \( L_i = \{ j \in V \setminus i, \max_{x_i, x_j, x'_j} |\hat{P}(x_i | x_j) - \hat{P}(x_i | x'_j)| > \frac{\epsilon_2}{2} \} \).
  \For {\( j \in L_i \)}
  \State \( j \in N_i \iff \exists S \text{ with } |S| \leq D_1, \forall T \text{ with } |T| \leq D_2, \hat{\Delta}(X_i; X_j | X_S, X_T) \leq \frac{\epsilon_2}{2} \).
  \EndFor
  \EndFor
\end{algorithmic}

Next we show a set of concentration results for the empirical quantities for any general MRF.

**Lemma 8.1:** Fix \( \gamma > 0 \). Let \( L = \max_i |L_i| \). For \( \forall \alpha > 0 \),
1) Assume $\gamma \leq \frac{1}{4}$. If

$$n > \frac{2[(2 + \alpha) \log p + 2 \log |X|]}{\gamma^2},$$

then for some constant $c'$, with probability $1 - \frac{c'}{p^5}$, $\forall i, j \in V, \forall x_i, x_j$,

$$|\hat{P}(x_i|x_j) - P(x_i|x_j)| < 4\gamma;$$

2) Assume $\gamma \leq \frac{\delta}{2}$ and $P(x_S) > \delta, \forall S \subset V, |S| \leq D_1 + D_2 + 1$. If

$$n > \frac{2[(1 + \alpha) \log p + (D_1 + D_2 + 1) \log L + (D_1 + D_2 + 2) \log |X|]}{\gamma^2},$$

then for some constant $c''$, with probability $1 - \frac{c''}{p^5}$, $\forall i \in V, \forall j \in L_i, \forall S \subset L_i, |S| \leq D_1 + D_2, \forall x_i, x_j, x_S$,

$$|\hat{P}(x_i|x_j, x_S) - P(x_i|x_j, x_S)| < \frac{2\gamma}{\delta};$$

3) Assume $\gamma \leq \frac{1}{2|X|^1+D_2+2} < 1$,

$$n > \frac{2[(1 + \alpha) \log p + (D_1 + D_2 + 1) \log L + (D_1 + D_2 + 2) \log |X|]}{\gamma^2},$$

then for some constant $c''$, with probability $1 - \frac{c''}{p^5}$, $\forall i, j \in V, |S| \leq D_1 + D_2, \forall x_i, x_j, x_S$,

$$|\hat{I}(X_i; X_j|X_S) - I(X_i; X_j|X_S)| < 8|X|^{D_1+D_2+2}\sqrt{\gamma}.$$

**Proof:** See appendix. 

When $n$ is large, the above lemma shows that the empirical estimates are very close to the exact conditional probabilities and mutual information. The computational complexities of the algorithms in the previous sections have to be multiplied by $n$ to get the corresponding computational complexities for the case where only samples from $P(x)$ are given.

From this lemma, we can also see that the sample complexity of an algorithm using the $P$-test depends on both $\varepsilon_2$ and $\delta$, while that of an algorithm using the $I$-test only depends on $\varepsilon_2$. Thus, if we can find a lower bound $\delta$ on $P(x_S)$ which is independent of the size of the graph, then we prefer the $P$-test; otherwise, we prefer the $I$-test. For Ising models on bounded degree graphs, we have the following constant lower bound on $P(x_S)$ for any bounded size set $S$, therefore, we prefer the $P$-test over the $I$-test.

**Lemma 8.2:** $\forall S \subset V, \forall x_S, P(x_S) \geq 2^{-|S|} \exp((-2(d + 1)|S|^2 J_{\text{max}})).$

**Proof:** See appendix.

For Ising models on the random graph $G(p, \xi)$, we cannot get a constant lower bound using the same proof technique above as the degree of any node can be potentially unbounded. In this case, it turns out that an algorithm using $I$-test gives lower sample complexity.

In the next few subsections, we state consistency theorems for the algorithms based on samples. These theorems are essentially simple modifications of the corresponding theorems for the case where $P(x)$ is known.
A. Ising models on Bounded Degree Graphs

Assume the maximum degree of the graph is \( d \). In Section III-A, we have seen \( \text{Exact\_Alg}(d, d - 1, 0) \) can be used to recover the graph when the distribution \( P(x) \) is known. The quantity \( \Delta(X_i; X_j|X_S, X_T) \) used in the conditional independence test is 0 when \( i, j \) are not neighbors and \( S \) is a separator and is positive when \( i, j \) are neighbors and \( T \) is a separator. So \( \epsilon_2 = 0 \) is enough to determine if there is an edge between \( i, j \). But when only samples are given, \( \hat{\Delta} \) is positive with high probability due to sample noise and therefore, setting \( \epsilon_2 = 0 \) does not work in this case. We need to bound \( \Delta \) away from 0 so that, with high probability, the noise can be tolerated. To do this, we further assume that \( J_{\min} \leq |J_{ij}| \leq J_{\max} \). Then we have the following lower bound from [5, Proposition 2].

**Proposition 8.3:** When \( i, j \) are neighbors and \( T = N_i \setminus j \), there is a choice of \( x_i, x_j, x'_j, x_S, x_T \) such that

\[
|P(x_i|x_j, x_S, x_T) - P(x_i|x'_j, x_S, x_T)| \geq \frac{\tanh(2J_{\min})}{2e^{2J_{\max}} + 2e^{-2J_{\max}}}.
\]

Together with Lemma 8.2, we let

\[
\epsilon_2 = \frac{\tanh(2J_{\min})}{2e^{2J_{\max}} + 2e^{-2J_{\max}}}, \quad \delta = 2^{-2d+1} \exp(-2(d+1)(2d-1)J_{\max}).
\]

For the structure learning algorithm to work, we require that with high probability \( \tilde{P}(x_i|x_j, x_S, x_T) \) is at most \( \epsilon_2 \) away from \( P(x_i|x_j, x_S, x_T), \forall i, j, S, T, x_i, x_j, x_S, x_T \). By the concentration results, we get the following theorem for Ising models on bounded degree graphs.

**Theorem 8.4:** \( \epsilon_2, \delta \) are defined as above. Let \( \gamma = \frac{\epsilon_2 \delta}{8} \wedge \frac{\delta}{2} < 1 \). If \( n > \frac{2[(2d+1+\alpha)\log p + (2d+1)\log 2]}{\gamma^2} \), the algorithm \( \text{Alg}_{P}(d, d - 1, \epsilon_2) \) recovers \( G \) with probability \( 1 - \frac{c}{p^n} \) for some constant \( c \).

B. Ferromagnetic Ising Models on Bounded Degree Graphs

Assume the maximum degree of the graph is \( d \) and \( J_{\min} \leq |J_{ij}| \leq J_{\max} \). Combining the theorems in Section III for bounded degree graphs and ferromagnetic Ising models, we can use \( \text{Exact\_Alg}(d, 0, 0) \) to recover the graph when the distribution \( P(x) \) is known. When we only have samples, again we need to bound \( \Delta \) away from 0 for neighbors to tolerate the sample noise. We note that, when \( |S| \leq d \), \( |N_S| \leq d|S| \leq d^2 \) as the graph has maximum degree \( d \). Therefore, by Lemma 5.3, we let

\[
\epsilon_2 = \frac{1}{16} \left(1 - e^{-4J_{\min}}\right) e^{-4d^2J_{\max}}.
\]

Let \( \delta \) be the one defined in the previous subsection. By the concentration results, we get the following theorem for ferromagnetic Ising models on bounded degree graphs.

**Theorem 8.5:** \( \epsilon_2, \delta \) are defined as above. Let \( \gamma = \frac{\epsilon_2 \delta}{8} \wedge \frac{\delta}{2} < 1 \). If \( n > \frac{2[(d+2+\alpha)\log p + (d+2)\log 2]}{\gamma^2} \), the algorithm \( \text{Alg}_{P}(d, 0, \epsilon_2) \) recovers \( G \) with probability \( 1 - \frac{c}{p^n} \) for some constant \( c \).

As mentioned in Section V, the correlation between neighbor nodes is lower bounded by a constant \( \epsilon_1 \). For the preprocessing step to work, We require that with high probability \( \tilde{P}(x_i|x_j) \) is at least \( \epsilon_1 \) away from \( P(x_i|x_j), \forall i, j, x_i, x_j \). When the Ising model is further in the correlation decay region, with high probability, we have \( L = \max_i L_i \leq d^{\frac{d}{4}} \) as in Section VI, so the computational complexity is \( O(np^2) \).

**Theorem 8.6:** Assume \( d \tan h J_{\max} < 1, \epsilon_1, \epsilon_2, \delta \) are defined as above. Let \( \gamma = \frac{\epsilon_1}{32} \wedge \frac{\epsilon_2 \delta}{16} \wedge \frac{\delta}{2} < 1 \). If \( n > \frac{2[(2+\alpha)\log p + (d+1)\log d + (d+2)\log 2]}{\gamma^2} \), the algorithm \( \text{Alg}_{P}(d, 0, \epsilon_1, \epsilon_2) \) recovers \( G \) with probability \( 1 - \frac{c}{p^n} \) for some constant \( c \).
C. Ising Models on Bounded Degree Graphs with Correlation Decay and Large Girth

The extension from structure learning algorithms using exact distribution to algorithms using samples is easy in this case. In the proof of Theorem 6.3, the exact $P$-tests for neighbor nodes and non-neighbor nodes are bounded away from $\frac{\epsilon_2}{8}$ by at least $\frac{\epsilon_2}{4}$. To show the empirical $P$-test works, we only require that with high probability $\hat{P}(x_i|x_j, x_S, x_T)$ is at most $\frac{\epsilon_2}{8}$ away from $P(x_i|x_j, x_S, x_T)$, $\forall i, j, S, T, x_i, x_j, x_S, x_T$; i.e., the sample noise level is comparable with the impact of the long paths under correlation decay. Similarly, to show the empirical correlation test in the preprocessing step works, we only require that with high probability $\hat{P}(x_i|x_j)$ is at least $\frac{\epsilon_4}{8}$ away from $P(x_i|x_j)$, $\forall i, j, x_i, x_j$. By the concentration results, we get the following theorem for the algorithm with a preprocessing step. The computational complexity is $O(np^2)$.

**Theorem 8.7:** Assume $(d - 1) \tanh J_{\text{max}} < 1$. $g_1, g_2, \epsilon_1, \epsilon_2, \delta$ are defined as above. Let $\gamma = \frac{\epsilon_1^2}{32} \wedge \frac{\epsilon_2^2}{16} \wedge \frac{\delta}{2}$. If
\[
n > \frac{2 \left[ (2 + \alpha) \log p + 2 \gamma_1 \log d + 3 \log 2 \right]}{\gamma^2},
\]
the algorithm $Alg_{g}(1, 0, \epsilon_1, \epsilon_2)$ recovers $G$ with probability $1 - \frac{c}{p^n}$ for some constant $c$.

D. Ising Models on Random Graph $\mathcal{G}(p, \frac{c}{p})$ with Correlation Decay

The extension to random graphs is also straightforward. As in the large girth graph case, we only require that with high probability $\hat{I}(X_i; X_j | X_S, X_T)$ is at most $\frac{\epsilon_2}{8}$ away from $I(X_i; X_j | X_S, X_T)$, $\forall i, j, S, T, X_i, X_j, X_S, X_T$, for the empirical $I$-test, and $\hat{P}(x_i|x_j)$ is at least $\frac{\epsilon_4}{8}$ away from $P(x_i|x_j)$, $\forall i, j, x_i, x_j$, for the empirical correlation test. The computational complexity for $Alg_1(2, 1, \epsilon_2)$ is $O(np^2)$ and for $Alg_1(2, 0, \epsilon_1, \epsilon_2)$ is $O(np^2)$.

**Theorem 8.8:** Assume $c \tanh J_{\text{max}} < 1$. $\epsilon_2$ is defined as above. Let $\gamma = \left( \frac{\epsilon_2}{32} \right)^2 \wedge \frac{1}{64} < 1$. If $n > \frac{2 \left[ (2 + \alpha) \log p + 3 \log L + 5 \log 2 \right]}{\gamma^2}$, the algorithm $Alg_1(2, 1, \epsilon_2)$ recovers the graph $G$ almost always.

**Theorem 8.9:** Assume $c \tanh J_{\text{max}} < 1$. $\epsilon_1, \epsilon_2$ are defined as above. Let $\gamma = \frac{\epsilon_1}{32} \wedge \left( \frac{\epsilon_2}{32} \right)^2 \wedge \frac{1}{32} < 1$. If $n > \frac{2 \left[ (2 + \alpha) \log p + 3 \log L + 5 \log 2 \right]}{\gamma^2}$, the algorithm $Alg_1(2, 0, \epsilon_1, \epsilon_2)$ recovers the graph $G$ almost always.

IX. Conclusions

In this paper, we have presented a unified framework for structure learning in discrete Markov random fields. The key idea behind our algorithms is a new necessary and sufficient condition for the existence of an edge between two nodes in an MRF (see Lemma 3.2). For appropriate choices of $S$ and $T$ in the lemma, the complexity of using this condition is quite low for many classes of sparse graphs, including bounded degree graphs with correlation decay and Ising models defined over a graph drawn from the class of Erdős-Rényi graphs.

In particular, the unified framework yields algorithms which have the same computational complexity as prior algorithms for bounded degree graphs with or without correlation decay as in [5], [12], and for ferromagnetic Ising models over Erdős-Rényi graphs in [2]. In addition, our framework allows us to derive an algorithm for general Ising models over Erdős-Rényi graphs, which appears to be a new result. More importantly, Lemma 3.2 seems to provide a unified basis for deriving low-complexity algorithms for many types of MRFs. The primary application of the paper is for structure learning in the high-dimensional regime, i.e., one where the number of samples is small compared to the number of nodes in the graph.
Thus, in addition to computational complexity results, we have provided results on sample complexity, which follow from standard Hoeffding bound arguments.

ACKNOWLEDGMENTS

We thank Anima Anandkumar and Sreekanth Annapureddy for useful discussions. In particular, we would like to thank Anandkumar for suggesting the use of the SAW tree in the proof of Lemma C.7 and Annapureddy for suggesting the proof of Lemma 3.2.

APPENDIX A

BOUNDED DEGREE GRAPH

A. Proof of Lemma 8.2

Let \( N_s \) be the neighbor nodes of \( S \).

\[
P(x_S) = \sum_{x_N} P(x_N)P(x_S|x_N)
\]

\[
\geq \min_{x_S} P(x_S|x_N)
\]

\[
= \min_{x_S} \frac{\exp(x^T_SJ_Sx_S + x^T_SJ_Nx_N)}{\sum_{x'_S} \exp(x'^T_SJ_Sx'_S + x'^T_SJ_Nx_N)}
\]

\[
\geq 2^{|S|} \max_{x_S,x_N} \exp(x^T_SJ_Sx_S + x^T_SJ_Nx_N)
\]

\[
\geq 2^{-|S|} \exp(-2(|S|^2J_{max} + |S||N_S|J_{max}))
\]

B. Proof of Theorem 6.3

First consider \((i, j) \in E \). Without loss of generality, assume \( J_{ij} > 0 \). By the assumption on the size of the \( h \)-separator between neighbor nodes, there exists \( T' \subset N_i, |T'| \leq D_2 \) such that, when the set \( T' \) is removed from the graph, the length of any path from \( i \) to \( j \) is no less than \( h \). For any \( S \), let \( T = T' \setminus S \). To simplify the notation, let \( R = S \cup T \) and \( W = V \setminus R \). For any value \( x_R \), let \( Q \) be the joint probability of \( X_W \) conditioned on \( X_R = x_R \), i.e., \( Q(X_W) = P(X_W|x_R) \). \( Q \) has the same edge coefficients for the unconditioned nodes, but is not zero-field as conditioning induces external fields. Let \( \tilde{Q} \) denote the joint probability when edge \((i, j)\) is removed from \( Q \). We note that \( Q \) and \( \tilde{Q} \) satisfy the same correlation decay property as \( P \), so

\[
\tilde{Q}(1, 1) = \tilde{Q}(X_i = 1)\tilde{Q}(X_j = 1|X_i = 1)
\]

\[
\geq \tilde{Q}(X_i = 1)[\tilde{Q}(X_j = 1|X_i = -1) - \beta h^i]
\]

\[
\geq \tilde{Q}(X_i = 1)[\tilde{Q}(X_j = 1|X_i = -1) - \beta h]
\]
Similarly, \( \bar{Q}(-1, -1) \geq \bar{Q}(X_i = -1)[\bar{Q}(X_j = -1|X_i = 1) - \beta \alpha^h] \). Then,
\[
\bar{Q}(1, 1)\bar{Q}(-1, -1) \\
\geq \bar{Q}(X_i = 1)\bar{Q}(X_i = -1)[\bar{Q}(X_j = 1|X_i = -1) - \beta \alpha^h][\bar{Q}(X_j = -1|X_i = 1) - \beta \alpha^h] \\
\geq \bar{Q}(1, -1)\bar{Q}(-1, 1) - 2\beta \alpha^h
\]

Using the above inequality, we have the following lower bound on the \( P \)-test quantity.
\[
\max_{x_i, x_j, x_j'} |P(x_i|x_j, x_S, x_T) - P(x_i|x_j', x_S, x_T)| \\
\geq |Q(x_i = 1|x_j = 1) - Q(x_i = 1|x_j = -1)| \\
= \left| \frac{Q(x_i = 1, x_j = 1)}{Q(x_j = 1)} - \frac{Q(x_i = 1, x_j = -1)}{Q(x_j = -1)} \right| \\
= \left| \frac{Q(x_i = 1, x_j = 1)Q(x_i = -1, x_j = -1) - Q(x_i = 1, x_j = -1)Q(x_i = -1, x_j = 1)}{Q(x_j = 1)Q(x_j = -1)} \right| \\
\geq \frac{\left| e^{2J_{ij}}\bar{Q}(1, 1)\bar{Q}(-1, -1) - e^{-2J_{ij}}\bar{Q}(1, -1)\bar{Q}(-1, 1) \right|}{\left( e^{J_{ij}}\bar{Q}(1, 1) + e^{-J_{ij}}\bar{Q}(-1, 1) \right) \left( e^{-J_{ij}}\bar{Q}(1, -1) + e^{J_{ij}}\bar{Q}(-1, 1) \right)} \\
\geq e^{-2J_{ij}} \left[ (e^{2J_{ij}} - e^{-2J_{ij}})\bar{Q}(1, -1)\bar{Q}(-1, 1) - 2e^{2J_{ij}}\beta \alpha^h \right] \\
=(1 - e^{-4J_{ij}})\bar{Q}(1, -1)\bar{Q}(-1, 1) - 2\beta \alpha^h \\
\geq (1 - e^{-4J_{\min}})\bar{Q}(1, -1)\bar{Q}(-1, 1) - 2\beta \alpha^h.
\]

Let \( \bar{Q} \) denote the joint probability when all the external field terms are removed from \( \bar{Q} \); i.e.,
\[
\bar{Q}(X_W) \propto \bar{Q}(X_W)e^{h_{W^c}X_W}
\]
As there are at most \((D_1 + D_2)d\) edges between \( R \) and \( W \), we have \( ||h_W||_1 \leq (D_1 + D_2)dJ_{\max} \). Hence, for any subset \( U \subset W \) and value \( x_U \),
\[
\bar{Q}(x_U) = \frac{\bar{Q}(x_U)}{\sum x'_{U} \bar{Q}(x'_{U})} \\
= \frac{\sum_{x_{W \setminus U}} \bar{Q}(x_U, x_{W \setminus U})e^{h_{W^c}x_W}}{\sum_{x'_{U}} \sum_{x_{W \setminus U}} \bar{Q}(x'_U, x'_{W \setminus U})e^{h_{W^c}x'_{W^c}}} \\
\geq \bar{Q}(x_U)e^{-(D_1 + D_2)dJ_{\max}} \frac{e^{(D_1 + D_2)dJ_{\max}}}{\sum_{x_{W \setminus U}} \bar{Q}(x_U, x_{W \setminus U})e^{h_{W^c}x_W}} \\
= e^{-2(D_1 + D_2)dJ_{\max}} \bar{Q}(x_U).
\]

Moreover, \( \bar{Q} \) is zero-field by definition and again has the same correlation decay condition as \( P \), hence
\[
\bar{Q}(1, -1) + \bar{Q}(1, 1) = \bar{Q}(X_i = 1) = \frac{1}{2} \\
\frac{\bar{Q}(1, -1)}{\bar{Q}(1, 1)} \geq e^{-\beta \alpha^h},
\]
which gives the lower bound $\tilde{Q}(1, -1) \geq \frac{1}{2(1 + e^{\beta h})}$. Therefore, we have

$$\tilde{Q}(1, -1) \geq \frac{e^{-2(D_1 + D_2) \beta \alpha h}}{2(1 + e^{\beta h})}.$$ 

The same lower bound applies for $\tilde{Q}(-1, 1)$. Hence,

$$\max_{x_i, x_j, x'_j} |P(x_i | x_j, x_S, x_T) - P(x_i | x'_j, x_S, x_T)|$$

$$\geq \frac{(1 - e^{-4J_{\min}}) e^{-4(D_1 + D_2) \beta \alpha h}}{4(1 + e^{\beta h})^2} - 2 \beta \alpha h$$

$$\geq \frac{(1 - e^{-4J_{\min}}) e^{-4(D_1 + D_2) \beta \alpha h}}{36} - 2 \beta \alpha h$$

$$> \epsilon_2.$$ 

The second inequality uses the fact that $e^{\beta h} < 2$. The last inequality is by the choice of $h$.

Next consider $(i, j) \notin E$. By the choice of $h$, there exists $S \subset N_i, |S| \leq D_1$ such that, when the set $S$ is removed from the graph, the distance from $i$ to $j$ is no less than $h$. Let $T$ set with $|T| \leq D_2$. As there is no edge between $i, j$, the joint probability $Q$ and $\tilde{Q}$ are the same. Then $\forall x_S, x_T, x_i, x_j,$

$$|P(x_i | x_j, x_S, x_T) - P(x_i | x_j, x_S, x_T)|$$

$$= \tilde{Q}(x_i | x_j) - \tilde{Q}(x_i | -x_j)$$

$$= \tilde{Q}(x_i, x_j) \tilde{Q}(-x_i, -x_j) - \tilde{Q}(x_i, -x_j) \tilde{Q}(-x_i, x_j).$$

Similar as above, we have

$$\tilde{Q}(x_j) \geq e^{-2(D_1 + D_2) \beta \alpha h} \tilde{Q}(x_j) = \frac{1}{2} e^{-2(D_1 + D_2) \beta \alpha h}.$$ 

The same bound applies for $\tilde{Q}(-x_j)$. Therefore,

$$|P(x_i | x_j, x_S, x_T) - P(x_i | x_j, x_S, x_T)|$$

$$\leq 4 e^{4(D_1 + D_2) \beta \alpha h} |\tilde{Q}(x_i, x_j) \tilde{Q}(-x_i, -x_j) - \tilde{Q}(x_i, -x_j) \tilde{Q}(-x_i, x_j)|.$$ 

By correlation decay and the fact $\beta \alpha h < \ln 2 < 1$,

$$Q(x_i, x_j) Q(-x_i, -x_j) = Q(x_i | x_j) Q(x_j) Q(-x_i | -x_j) Q(-x_j)$$

$$\leq (Q(x_i | -x_j) + \beta \alpha h) Q(x_j) (Q(-x_i | -x_j) + \beta \alpha h) Q(-x_j)$$

$$\leq Q(x_i, -x_j) Q(-x_i, x_j) + 3 \beta \alpha h.$$ 

Similarly, we can show $Q(x_i, x_j) Q(-x_i, -x_j) \geq Q(x_i, -x_j) Q(-x_i, x_j) - 2 \beta \alpha h$. Hence, by the choice of $h$,

$$|P(x_i | x_j, x_S, x_T) - P(x_i | x_j, x_S, x_T)| \leq 12 e^{4(D_1 + D_2) \beta \alpha h} \beta \alpha h \leq \epsilon_2.$$


APPENDIX B
FERROMAGNETIC ISING MODELS

A. Proof of Lemma 5.3

By Proposition 5.2, \( X \) is associated, which gives
\[
P(X_i = 1, X_j = 1)P(X_i = -1, X_j = -1) \geq P(X_i = 1, X_j = -1)P(X_i = -1, X_j = 1), \forall i, j.
\]

If \( P \) is symmetric, i.e., \( P(x) = P(-x) \), we have
\[
P(X_i = 1, X_j = 1) \geq P(X_i = 1, X_j = -1), \forall i, j,
\]
which shows that \( P(X_i = 1, X_j = 1) \geq \frac{1}{4}, \forall i, j \in N_i, S \subset V. \)

A. Self-Avoiding-Walk Tree and Some Basic Results

This subsection introduces the notion of a self-avoiding-walk (SAW) tree, first introduced in [14], and presents some properties of a SAW tree. For an Ising model on a graph \( G \), fix an ordering of all the nodes. We say \( \text{dge}(i, j) \) is larger (smaller resp.) than \( (i, l) \) with respect to node \( i \) if \( j \) comes after (before resp.) \( l \) in the ordering. The SAW tree rooted at node \( i \) is denoted as \( T_{\text{saw}}(i; G) \). It is essentially the tree of self-avoiding walks originated from node \( i \) except that the terminal nodes closing a cycle are also included in the tree with a fixed value \( +1 \) or \( -1 \). In particular, a terminal node is fixed to \( +1 \) (resp. \( -1 \)) if the closing edge of the cycle is larger (resp. smaller) than the starting edge with respect to the terminal node. Let \( A \) denote the set of all terminal nodes in \( T_{\text{saw}}(i; G) \) and \( x_A \) denote the fixed configuration on \( A. \) For

APPENDIX C
RANDOM GRAPHS

The proofs in this section are related to the techniques developed in [2], [3]. The key differences are in adapting the proofs for general Ising models, as opposed to ferromagnetic models. We point out similarities and differences as we proceed with the section.

A. Self-Avoiding-Walk Tree and Some Basic Results

This subsection introduces the notion of a self-avoiding-walk (SAW) tree, first introduced in [14], and presents some properties of a SAW tree. For an Ising model on a graph \( G \), fix an ordering of all the nodes. We say \( \text{dge}(i, j) \) is larger (smaller resp.) than \( (i, l) \) with respect to node \( i \) if \( j \) comes after (before resp.) \( l \) in the ordering. The SAW tree rooted at node \( i \) is denoted as \( T_{\text{saw}}(i; G) \). It is essentially the tree of self-avoiding walks originated from node \( i \) except that the terminal nodes closing a cycle are also included in the tree with a fixed value \( +1 \) or \( -1 \). In particular, a terminal node is fixed to \( +1 \) (resp. \( -1 \)) if the closing edge of the cycle is larger (resp. smaller) than the starting edge with respect to the terminal node. Let \( A \) denote the set of all terminal nodes in \( T_{\text{saw}}(i; G) \) and \( x_A \) denote the fixed configuration on \( A. \) For
set $S \subset V$, let $U(S)$ denote the set of all non-terminal copies of nodes in $S$ in $T_{\text{saw}}(i; G)$. Notice that there is a natural way to define conditioning on $T_{\text{saw}}(i; G)$ according to the conditioning on $G$; specifically, if node $j$ in graph $G$ is fixed to a certain value, the non-terminal copies of $j$ in tree $T_{\text{saw}}(i; G)$ are fixed to the same value.

One important result is [8, Theorem 7], motivated by [14], says that the conditional probability of node $i$ on graph $G$ is the same as the corresponding conditional probability of node $i$ on tree $T_{\text{saw}}(i; G)$, which is easier to deal with.

**Proposition C.1:** Let $S$ be a subset of $V$. $\forall x_i, x_S, P(x_i|x_S; G) = P(x_i|x_{U(S)}, x_A; T_{\text{saw}}(i; G))$.

Next we list some basic results which will be used in later proofs. First we have the following lemma about $\gamma_p$-separators from [2]. Theorem 7.2 is an immediate result of this lemma.

**Lemma C.2:** [2] For all $i, j \in V$, the size of the smallest $\gamma_p$-separator of $i, j$ is at most 2 almost always.

Let $B(i, l; T_{\text{saw}}(i; G))$ be the set of nodes of distance $l$ from $i$ on the tree $T_{\text{saw}}(i; G)$. Recall that $A$ is the set of terminal nodes in the tree. Let $\tilde{A}$ be the subset of $A$ that are of distance at most $\gamma_p$ from $i$. The size of $B(i, l; T_{\text{saw}}(i; G))$ and $\tilde{A}$ are upper bounded as follows.

**Lemma C.3:** [11, Lemma 2.2] For $1 \leq l \leq a \log p$, where $0 < a < \frac{1}{2 \log c}$, we have

$$\max_i |B(i, l; T_{\text{saw}}(i; G))| = O(c^l \log p), \text{ almost always.}$$

**Lemma C.4:** $\forall i \in V, |\tilde{A}| \leq 1$ in $T_{\text{saw}}(i; G)$ almost always.

**Proof:** Each terminal node in $\tilde{A}$ corresponds to a cycle connected to $i$ with the total length of the cycle and the path to $i$ at most $\gamma_p$. Let $OLO_l$ denote the subgraph consists of two connected circles with total length $l$. This structure has $l - 1$ nodes and $l$ edges. Let $H = \{OLO_l, l \leq 2\gamma_p\}$ and $N_H$ denote the number of subgraphs containing an instance from $H$. Then it is equivalent to show that there is at most 1 such small cycle close to each node or $N_H = 0$ almost always.

$$E[N_H] \leq \sum_{l=1}^{2\gamma_p} \binom{p}{l-1} (l-1)!(l-1)^2 \frac{c^l}{p} \leq O(\sum_{l=1}^{2\gamma_p} p^{-l}l^2 c^l) = O(p^{-\frac{1}{2}}c^{2\gamma_p}) \leq O(p^{-\frac{1}{2}}) = o(1).$$

So, $P(N_H \geq 1) = o(1)$. £££

**B. Correlation Decay in Random Graphs**

This subsection is to prove Theorem 7.1 which characterizes the correlation decay property of a random graph.

First we state a correlation decay property for tree graphs. This result shows that having external fields only makes the correlation decay faster.

**Lemma C.5:** Let $P$ be a general Ising model with external fields on a tree $T$. Assume $|J_{ij}| \leq J_{\text{max}}$. $\forall i, j \in T$,

$$|P(x_i|x_j) - P(x_i|x'_j)| \leq (\tanh J_{\text{max}})^{d(i,j)}. $$

**Proof:** The basic idea in the proof is get an upper bound that does not depend on the external field. To do this, we proceed as in the proof of Lemma 4.1 in [4]. First, as noted in [4], w.l.o.g. assume the tree is a line from $i$ to $j$. Then, we prove the result by induction on the number of hops in the line.
1) \(d(i, j) = 1\) or \(j \in N_i\). The graph has only two nodes. We have

\[
P(x_i | x_j) = \frac{e^{J_{ij}x_i x_j + h_i x_i}}{e^{J_{ij}x_j + h_i} + e^{-J_{ij}x_j - h_i}}.
\]

Hence,

\[
|P(x_i | x_j) - P(x_i | x'_j)| = \frac{|e^{2J_{ij}} - e^{-2J_{ij}}|}{(e^{J_{ij} + h_i} + e^{-J_{ij} - h_i})(e^{-J_{ij} + h_i} + e^{J_{ij} - h_i})}
\]

\[
= \frac{|e^{2J_{ij}} - e^{-2J_{ij}}|}{e^{2J_{ij}} + e^{-2J_{ij}} + e^{2h_i} + e^{-2h_i}}
\]

This function is even in both \(J_{ij}\) and \(h_i\). Without loss of generality, assume \(J_{ij} \geq 0, h_i \geq 0\). It is not hard to see that the RHS is maximized when \(h_i = 0\). So

\[
|P(x_i | x_j) - P(x_i | x'_j)| \leq \tanh |J_{ij}| \leq \tanh J_{\max}.
\]

The inequality suggests that, when there is external field, the impact of one node on the other is reduced.

2) Assume the claim is true for \(d(i, j) \leq k\). For \(d(i, j) = k + 1\), pick any \(l\) on the path from \(i\) to \(j\), and note that \(X_i \rightarrow X_l \rightarrow X_j\) forms a Markov chain. Moreover, \(d(i, l) \leq k\) and \(d(l, j) \leq k\).

\[
|P(x_i | x_j) - P(x_i | x'_j)| = \sum_{x_l} P(x_i | x_l)P(x_l | x_j) - \sum_{x_l} P(x_i | x_l)P(x_l | x'_j)
\]

\[
= |P(x_i | x_l)(P(x_l | x_j) - P(x_l | x'_j)) + P(x_i | x'_l)(P(x'_l | x_j) - P(x'_l | x'_j))|
\]

\[
= (P(x_i | x_l) - P(x_i | x'_l))(P(x_l | x_j) - P(x_l | x'_j))
\]

\[
\leq (\tanh J_{\max})^{d(i, l)}(\tanh J_{\max})^{d(l, j)} = (\tanh J_{\max})^{d(i, j)}
\]

The third equality follows by observing that \(P(x_l | x_j) - P(x_l | x'_j) = -(P(x'_l | x_j) - P(x'_l | x'_j))\). The last inequality is by induction.

Writing the conditional probability on a graph as a conditional probability on the corresponding SAW tree, we can apply the above lemma and show the correlation decay property for random graphs.

**Lemma C.6:** Let \(P\) be a general Ising model on a graph \(G\). Fix \(i \in V, \forall j \notin N_i\), let \(S\) be the \(\gamma\)-separator between \(i, j\) and \(B = B(i, \gamma; T_{saw}(i; G))\), then \(\forall x_i, x_j, x'_j, x_S\),

\[
|P(x_i | x_j, x_S) - P(x_i | x'_j, x_S)| \leq |B|(\tanh J_{\max})^\gamma.
\]

**Proof:** Let \(Z\) be the subset of \(U(j)\) on \(T_{saw}(i; G)\) that is not separated by \(U(S)\) from \(i\). By the
In the above, \( \gamma \) for ferromagnetic Ising models on the random graph \( G \) comes in handy when proving the other two theorems. This result was conjectured to hold in \([2]\) holds for general Ising models, whereas the proof in \([2]\) is specific to ferromagnetic Ising models. Lemma 3 in \([2]\) is by the choice of \( S \) as the \( \gamma \)-separator and the definition of \( Z \). Step (c) uses the fact that \( Z \) is separated from \( i \) by \( B \). In (d), \( x_B, x_B'' \) represent the maximizer and minimizer respectively. Step (e) is by telescoping the sign of \( x_B \).

Notice that the Hamming distance between \( x_B, x_B'' \) is at most \( |B| \), and we can apply the above lemma to each pair as the conditioning terms differ only on one node. The above proof is similar to the proof of Lemma 3 in \([2]\). However, in going from step (c) to step (d) above, it is important to note that our proof holds for general Ising models, whereas the proof in \([2]\) is specific to ferromagnetic Ising models. ■

**Proof of Theorem 7.1:** As in \([2]\), setting \( \gamma = \gamma_p \) in the above lemma and noticing that

\[
|B(i, \gamma_p; T_{\text{saw}}(i; G))| = O(c^{\gamma_p} \log p),
\]

we get

\[
|P(x_i | x_j, x_S) - P(x_i | x_j', x_S)| \leq O((c \tanh J_{\text{max}})^{\gamma_p} \log p) = O(p^{-\frac{\log c}{(1+c)(\log p)}} \log p) = o(p^{-\kappa}).
\]

C. Asymptotic Lower Bound on \( P(x_i | x_R) \) When \( |R| \leq 3 \)

This subsection is to prove that \( P(x_i | x_R) \) is lower bounded by some constant when \( |R| \leq 3 \). This result comes in handy when proving the other two theorems. This result was conjectured to hold in \([2]\) for ferromagnetic Ising models on the random graph \( G(p, \frac{\varepsilon}{p}) \) without a proof. Here we prove that it is also true for general Ising models on the random graph.

**Lemma C.7:** \( \forall i \in V, \forall R \subset V, |R| \leq 3 \), there exists a constant \( C \) such that \( \forall x_i, x_R, P(x_i | x_R) \geq C \) almost always.

This basic idea is that the conditional probability \( P(x_i | x_R) \) is equal to some conditional probability on a SAW tree, which in turn is viewed as some unconditional probability on the same tree with induced external fields. Then we apply a tree reduction to the SAW tree till only the root is left, and show that the induced external field on the root is bounded, which implies that the probability of the root taking +1 or −1 is bounded.
On a tree graph, when calculating a probability which involves no nodes in a subtree, we can reduce the subtree by simply summing (marginalizing) over all the nodes in it. This reduction produces an Ising model on the rest part of the tree with the same $J_{ij}$ and $h_i$ except for the root of the subtree, which would have an induced external field due to the reduction of the subtree. The probability we want to calculate remains unchanged on this new tree. Such induced external fields are bounded according to the following lemma.

**Lemma C.8:** Consider a leaf node 2 and its parent node 1. The induced external field $h_1'$ on node 1 due to summation over node 2 satisfies

$$|h_1'| \leq |h_2| \tanh |J_{12}|.$$

We first prove an inequality which is used in the proof of the above lemma.

**Lemma C.9:** \(\forall x \geq 0, y \geq 0,\)

$$e^{2x \tanh y} \geq \frac{e^{x+y} + e^{-x-y}}{e^{x-y} + e^{-x+y}}.$$

**Proof:** Let \(u = \tanh y \in [0, 1],\) then \(y = \frac{1}{2} \ln \frac{1+u}{1-u}.\) The required result is equivalent to showing that

$$e^{2yu}[(1+u)e^{-x} + (1-u)e^x] > (1+u)e^x + (1-u)e^{-x}.$$

Define

$$f_u(z) = (1+u)e^{uz} + (1-u)e^{(1+u)z} - (1+u)e^z - (1-u).$$

Clearly, \(f_u(0) = 0,\) and

$$f_u'(z) = (1+u)[ue^{uz} + (1-u)e^{(1+u)z} - e^z].$$

By the convexity of \(e^z, ue^{uz} + (1-u)e^{(1+u)z} \geq e^z.\) Hence, \(f_u'(z) \geq 0,\) which implies \(f_u(z) \geq 0.\) We finish the proof by noticing that the original inequality is equivalent to \(f_u(2x) \geq 0.\)

**Proof of Lemma C.8:**

$$\sum_{x_2} e^{J_{12}x_1x_2 + h_2x_2} = e^{J_{12}x_1 + h_2} + e^{-J_{12}x_1 - h_2} \propto e^{h_1'x_1}.$$  

Comparing the ratio of \(x_1 = \pm 1,\) we get

$$\frac{e^{J_{12}+h_2} + e^{-J_{12}-h_2}}{e^{-J_{12}+h_2} + e^{J_{12}-h_2}} = e^{h_1'} = e^{2h_1'}.$$

So

$$h_1' = \frac{1}{2} \log \frac{e^{J_{12}+h_2} + e^{-J_{12}-h_2}}{e^{-J_{12}+h_2} + e^{J_{12}-h_2}} \leq |h_2| \tanh |J_{12}|.$$

The last inequality follows from Lemma C.9.

It is easy to see that \(|h_1'| \leq |h_2| \tanh |J_{\max}| < |h_2|.\) By induction, we can bound the external field induced by the whole subtree.
Proof of Lemma C.7: First we have

\[ P(x_i|x_R) = P(x_i|x_U(R), x_A; T_{saw}(i; G)) \]

\[ = \sum_{x_B} P(x_i|x_B, x_{\tilde{U}(R)}, x_{\tilde{A}}; T_{saw}(i; G)) P(x_B|x_U(R), x_A; T_{saw}(i; G)) \]

\[ \geq \min_{x_B} P(x_i|x_B, x_{\tilde{U}(R)}, x_{\tilde{A}}; T_{saw}(i; G)) \]

\[ = P(x_i|x_B, x_{\tilde{U}(R)}, x_{\tilde{A}}; T_{saw}(i; G)) \triangleq Q(x_i), \]

where \( Q \) is the probability on the tree with external fields induced by \( x_B, x_{\tilde{U}(R)}, x_{\tilde{A}} \). We only need to consider the external fields on the parent nodes of \( B, \tilde{U}(R), \tilde{A} \) as the conditional probability is on a tree. The nodes affected by \( B \) are all \( \gamma_p \) away from \( i \) and the total number of them is no larger than \( |B| \), which is bounded by Lemma C.3. The number of nodes affected by \( \tilde{U}(R), \tilde{A} \) is no larger than \( |\tilde{U}(R)| + |\tilde{A}| \).

By Lemma C.2 and Lemma C.4, \( |\tilde{U}(R)| \leq 2|R| \) and \( |\tilde{A}| \leq 1 \) almost always. Applying the reduction technique to the tree till a single root node \( i \), by Lemma C.8, we bound the induced external field on \( i \) as

\[ |h_i| \leq [(\tanh J_{max})^\gamma |B| + (|\tilde{U}(R)| + |\tilde{A}|)] J_{max} \]

\[ \leq O((c \tanh J_{max})^\gamma \log n + 2|R| + 1) \]

\[ \leq O(n^{-\kappa} + 7) = O(1). \]

So,

\[ Q(x_i) = \frac{e^{h_i x_i}}{e^{h_i x_i} + e^{-h_i x_i}} \geq \Omega(e^{-2|h_i|}) = \Omega(1). \]

When \( p \) is large enough, there exists some constant \( C \) such that \( P(x_i|x_R) \geq C \). 

D. Proof of Theorem 7.3

Let \( S \) be a \( \gamma_p \)-separator of node \( i, j \) with size \( |S| \leq 3 \). It is straightforward to show that \( I(X_i; X_j| X_S) = o(p^{-2\kappa}) \) in a manner similar to [2, Lemma 5]. The only difference is that the correlation decay property in Theorem 7.1 takes a different form, which is easier to apply, therefore the proof there needs to be modified accordingly. We also note that the constant \( C \) in Lemma C.7 is referred to as \( f_{\min}(S) \) in [2]. The details are omitted here.

E. Proof of Theorem 7.4

When \( j \) is a neighbor of \( i \), conditioned on the approximate separator \( T \), there is one copy of \( j \) which is a child of the root \( i \) in the SAW tree and is the only copy that within \( \gamma_p \) from \( i \). In Theorem 7.4, we show that the effect of conditioning on \( T \) is bounded and this copy of \( j \) has a nontrivial impact on \( i \). With a little abuse of notation, we use \( j \) to denote this copy of \( j \) in \( T_{saw}(i; G) \).
**Proof of Theorem 7.4:** W.l.o.g assume \( J_{ij} > 0 \). As in Lemma C.6,

\[
\max_{x, x'} |P(x | x', x_T) - P(x | x_T)|
\]

\[
= \max_{x, x'} |P(x | x_{U(j)}, x_{U(T)}, x_A; T_{saw}(i; G)) - P(x | x'_{U(j)}, x_{U(T)}, x_A; T_{saw}(i; G))|
\]

\[
= \max_{x} \sum_{x_B} P(x | x_B, x_{U(T)}, x_A; T_{saw}(i; G)) P(x_B | x_{U(T)}, x_A; T_{saw}(i; G))
\]

\[
\geq \min_{x_B} P(x_i = + | x_j = +, x_B, x_{U(T)}, x_A; T_{saw}(i; G)) - \max_{x_B} P(x_i = + | x_j = -, x_B, x_{U(T)}, x_A; T_{saw}(i; G))
\]

\[
= P(x_i = +1 | x_j = +1, x_B^M, x_{U(T)}, x_A; T_{saw}(i; G)) - P(x_i = +1 | x_j = -1, x_B^M, x_{U(T)}, x_A; T_{saw}(i; G))
\]

\[
\geq Q(x_i = +1 | x_j = +1) - Q(x_i = +1 | x_j = -1) - |B|(\tanh J_{max})^\gamma,
\]

where \( Q \) is the probability measure on the reduced graph with only nodes \( i, j \). We have

\[
Q(x_i = +1 | x_j = +1) - Q(x_i = +1 | x_j = -1) = e^{2J_{ij}} - e^{-2J_{ij}}
\]

\[
\geq \frac{e^{2J_{ij}} + e^{-2J_{ij}} + e^{2h_i} + e^{-2h_i}}{e^{2J_{min}} + e^{-2J_{min}} + e^{2h_i} + e^{-2h_i}} = \Omega(e^{-2|h_i|}).
\]

The external fields in \( Q \) are induced by the conditioning on \( B, \tilde{U}(T), \tilde{A} \). As in the proof of Lemma C.7, we have \( |h_i| \leq O(1) \), so \( Q(x_i = +1 | x_j = +) - Q(x_i = +1 | x_j = -) = \Omega(1) \). Hence,

\[
\max_{x_i, x_j} |P(x_i | x_j, x_S) - P(x_i | x'_j, x_S)| \geq \Omega(1) - O(p^{-\kappa}) = \Omega(1).
\]

Using this result, the lower bound \( I(X_i; X_j | X_T) = \Omega(1) \) simply follows from the proof of [2, Lemma 7]. Again we note that the constant \( C \) in Lemma C.7 is referred to as \( f_{min}(T) \) in [2]. The details are omitted here.

---

**F. Proof of Theorem 7.5**

The proof of the theorem needs the following lemma.

**Lemma C.10:** \( X \) is a ferromagnetic Ising model (possibly with external fields). \( \forall i \in V, \forall S \subset V \setminus i, \)

\[
P(x_i = +1 | x_S = +1) \geq P(x_i = +1 | x_S = -1).
\]

**Proof:** For any node \( j \in S \), consider the probability \( \tilde{P}(x_i, x_j) = P(x_i, x_j | x_{S \setminus j}) \). \( \tilde{P} \) is still ferromagnetic and hence is associated. Then we have

\[
\tilde{P}(x_i = +1, x_j = +1) \tilde{P}(x_i = -1, x_j = -1) \geq \tilde{P}(x_i = +1, x_j = -1) \tilde{P}(x_i = -1, x_j = +1).
\]
After some algebraic manipulation, we get
\[
\bar{P}(x_i = +1|x_j = +1) \geq \bar{P}(x_i = +1|x_j = -1).
\]

This is equivalent saying that
\[
P(x_i = +1|x_j = +1, x_{S \setminus j} = +1) \geq P(x_i = +1|x_j = -1, x_{S \setminus j} = +1).
\]
So flipping one node from +1 to −1 reduces the conditional probability regardless the what value the rest of the nodes take. Continuing this process till we flip all the nodes in \(S\), we get the result
\[
P(x_i = +1|x_S = +1) \geq P(x_i = +1|x_S = -1).
\]

\[\blacksquare\]

**Proof of Theorem 7.5:** For \((i, j) \in E\), assume \(J_{ij} > 0\). Following the proof of Theorem 7.4,
\[
\max_{x_i, x_j} |P(x_i|x_j, x_S) - P(x_i|x_j', x_S)|
\]
\[= \max_{x_i, x_j} |P(x_i|x_U(j), x_{U(S)}, x_{A}; T_{saw}(i; G)) - P(x_i|x_U(j), x_{U(S)}, x_{A}; T_{saw}(i; G))|
\]
\[\geq P(x_i = +1|x_{U(j)} = +1, x^m_{B}, x_{U(S)}, x_{A}; T_{saw}(i; G)) - P(x_i = +1|x_{U(j)} = -1, x^M_{B}, x_{U(S)}, x_{A}; T_{saw}(i; G)).
\]

The only difference here is that we might have more than one copy of \(j\) in \(\tilde{U}(j)\). Let \(Z = \tilde{U}(j) \setminus j\). By the above lemma, we have
\[
\max_{x_i, x_j} |P(x_i|x_j, x_S) - P(x_i|x_j', x_S)|
\]
\[\geq P(x_i = +1|x_j = +1, x_Z = +1, x^m_{B}, x_{U(S)}, x_{A}; T_{saw}(i; G))
\]
\[- P(x_i = +1|x_j = -1, x_Z = +1, x^m_{B}, x_{U(S)}, x_{A}; T_{saw}(i; G))
\]
\[+ P(x_i = +1|x_j = -1, x_Z = -1, x^m_{B}, x_{U(S)}, x_{A}; T_{saw}(i; G))
\]
\[- P(x_i = +1|x_j = -1, x_Z = -1, x^M_{B}, x_{U(S)}, x_{A}; T_{saw}(i; G))
\]
\[\geq Q(x_i = +1|x_j = +1) - Q(x_i = +1|x_j = -1) - |B| \log H_{\max},
\]
As the size of \(Z\) is only a constant, by the same reasoning, we finish the theorem. \[\blacksquare\]

**APPENDIX D**

**PROOF OF LEMMA 8.1**

Before proving the concentration lemma, we first present the following two lemmas which helps handle the mutual information terms in the proof.

**Lemma D.1:** If \(0 < \epsilon < 1\), then \(0 < -\epsilon \log \epsilon < 1\).

**Proof:** We note that the \(-\epsilon \log \epsilon\) is concave in \(\epsilon\). By elementary calculus, one can show the above result. \[\blacksquare\]

**Lemma D.2:** [6, Theorem 17.3.3] If \(||P - Q||_1 \leq \frac{1}{2}\), then \(|H(P) - H(Q)| \leq -||P - Q||_1 \log \frac{||P - Q||_1}{|X|}\).

When \(||P - Q||_1 \leq \frac{1}{2}\), the RHS is increasing in \(||P - Q||_1\).

**Proof of Lemma 8.1:** By definition, \(\forall S \subset V\) and \(\forall x_S\), \(|1_{\{X^y_S = x_S\}} - P(x_S)| \leq 1\) and
\[
\hat{P}(x_S) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{X^y_S = x_S\}}.
\]
By the Hoeffding inequality,
\[
P \left( |\hat{P}(x_S) - P(x_S)| \geq \gamma \right) = P \left( \left| \sum_{i=1}^{n} 1_{\{x_S^i = x_S\}} - nP(x_S) \right| \geq n\gamma \right) \leq 2e^{-\frac{n\gamma^2}{2n}} \leq 2e^{-\frac{\gamma^2}{2}}.
\]

1) By the union bound, we have
\[
P \left( \exists S \subset V, |S| \leq 2, \exists x_S, |\hat{P}(x_S) - P(x_S)| \geq \gamma \right) < p^2|\mathcal{X}|2e^{-\frac{\gamma^2}{2}} = 2e^{-\frac{\gamma^2}{2} + 2\log p|\mathcal{X}|}
\]

For the choice of \(n\), with probability \(1 - \frac{\epsilon''}{p^2}\), \(\forall i, j \in V, \forall x_i, x_j\),

\[
|\hat{P}(x_i, x_j) - P(x_i, x_j)| < \gamma, |\hat{P}(x_i) - P(x_i)| < \gamma,
\]

which gives \(\hat{P}(x_j) > P(x_j) - \gamma \geq \frac{1}{2} - \gamma \geq \frac{1}{4} \) as \(\gamma < \frac{1}{4}\). Hence,

\[
\frac{|\hat{P}(x_i|x_j) - P(x_i|x_j)|}{\hat{P}(x_j)} \leq \frac{\hat{P}(x_i|x_j)|P(x_j) - P(x_j)|}{P(x_j)} + \frac{\hat{P}(x_i|x_j) - P(x_i|x_j)|}{P(x_j)} \leq 2\gamma = 4\gamma.
\]

2) By the union bound, we have
\[
P \left( \exists i \in V, \exists S \subset L_i, |S| \leq D_1 + D_2 + 1, \exists x_S, |\hat{P}(x_S) - P(x_S)| \geq \gamma, |\hat{P}(x_i, x_S) - P(x_i, x_S)| \geq \gamma \right)
\]

\[
< 2pL^{D_1 + D_2 + 1}|\mathcal{X}|^{D_1 + D_2 + 2}e^{-\frac{\gamma^2}{2}}
\]

\[
< 4e^{-\frac{\gamma^2}{2} + \log p + (D_1 + D_2 + 1) \log L + (D_1 + D_2 + 2) \log |\mathcal{X}|}.
\]

For the choice of \(n\), with probability \(1 - \frac{\epsilon''}{p^4}\), \(\forall i \in V, \forall j \in L_i, \forall S \subset L_i, |S| \leq D_1 + D_2, \forall x_i, x_j, x_S\),

\[
|\hat{P}(x_i, x_j, x_S) - P(x_i, x_j, x_S)| \leq \gamma, |\hat{P}(x_j, x_S) - P(x_j, x_S)| \leq \gamma,
\]

which gives \(\hat{P}(x_j, x_S) > P(x_j, x_S) - \gamma \geq \frac{\delta}{2} \) as \(\gamma < \frac{\delta}{2}\). Hence,

\[
\frac{|\hat{P}(x_i|x_j, x_S) - P(x_i|x_j, x_S)|}{\hat{P}(x_j, x_S)} \leq \frac{|\hat{P}(x_i, x_j, x_S)P(x_j, x_S) - P(x_i, x_j, x_S)\hat{P}(x_j, x_S)|}{\hat{P}(x_j, x_S)} \leq \frac{\hat{P}(x_i, x_j, x_S)|P(x_j, x_S) - P(x_j, x_S)|}{\hat{P}(x_j, x_S)} + \frac{\hat{P}(x_j, x_S)|\hat{P}(x_i, x_j, x_S) - P(x_i, x_j, x_S)|}{\hat{P}(x_j, x_S)} \leq \frac{2\gamma}{\delta}.
\]
3) As in the previous case, for the choice of \( n \), with probability \( 1 - \frac{\epsilon}{p^n} \), \( \forall i, j \in V, \forall S \subset L_i, |S| \leq D_1 + D_2, \forall x_i, x_j, x_S, \)

\[
|\hat{P}(x_i, x_j, x_S) - P(x_i, x_j, x_S)| \leq \gamma, \\
|\hat{P}(x_j, x_S) - P(x_j, x_S)| \leq \gamma, \\
|\hat{P}(x_S) - P(x_S)| \leq \gamma.
\]

So we get

\[
||\hat{P}(X_i, X_j, X_S) - P(X_i, X_j, X_S)||_1 \leq |\mathcal{X}|^{D_1 + D_2 + 2} \gamma \leq \frac{1}{2}.
\]

By Lemma D.2,

\[
|\hat{H}(X_i, X_j, X_S) - H(X_i, X_j, X_S)| \\
\leq -||\hat{P}(X_i, X_j, X_S) - P(X_i, X_j, X_S)||_1 \log \frac{||\hat{P}(X_i, X_j, X_S) - P(X_i, X_j, X_S)||_1}{|\mathcal{X}|^{D_1 + D_2 + 2}} \\
\leq -|\mathcal{X}|^{D_1 + D_2 + 2} \gamma \log \gamma = -2|\mathcal{X}|^{D_1 + D_2 + 2} \gamma \log \sqrt{\gamma} \\
\leq 2|\mathcal{X}|^{D_1 + D_2 + 2} \sqrt{\gamma}.
\]

The last inequality follows from Lemma D.1 with \( \epsilon = \sqrt{\gamma} \). Similarly, we have the same upper bound for \( |\hat{H}(X_i, X_S) - H(X_i, X_S)|, |\hat{H}(X_j, X_S) - H(X_j, X_S)| \) and \( |\hat{H}(X_S) - H(X_S)| \). We finish the proof by noticing that

\[
I(X_i; X_j|X_S) = H(X_i, X_S) + H(X_j, X_S) - H(X_i, X_j, X_S) - H(X_S).
\]

\[\blacksquare\]

REFERENCES


