Testing Properties of Ising Models

by

Sai Nishanth Dikkala

B.Tech. in Computer Science and Engineering
Indian Institute of Technology, 2014

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2017

© Massachusetts Institute of Technology 2017. All rights reserved.

Author . . . .
Department of Electrical Engineering and Computer Science
January 31, 2017

Certified by . . . .
Constantinos Daskalakis
Associate Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by . . . .
Leslie A. Kolodziejski
Chairman, Department Committee on Graduate Students

Testing Properties of Ising Models

by

Sai Nishanth Dikkala

Submitted to the Department of Electrical Engineering and Computer Science
on January 31, 2017, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

Given samples from an unknown multivariate distribution p, is it possible to distinguish whether p is the product of its marginals versus p being ε-far from every product distribution? Similarly, is it possible to distinguish whether p equals a given distribution q versus p and q being ε-far from each other? These problems of testing independence and goodness-of-fit have received enormous attention in statistics, information theory, and theoretical computer science, with sample-optimal algorithms known in several interesting regimes of parameters [14, 15, 17, 18, 20]. Unfortunately, it has also been understood that these problems become intractable in large dimensions, necessitating exponential sample complexity.

Motivated by the exponential lower bounds for general distributions as well as the ubiquity of Markov Random Fields (MRFs) in the modeling of high-dimensional distributions, we study distribution testing on *structured* multivariate distributions, and in particular the prototypical example of MRFs: *the Ising Model*. We demonstrate that, in this structured setting, we can avoid the curse of dimensionality, obtaining sample and time efficient testers for independence and goodness-of-fit which yield a sample complexity of $\mathrm{poly}(n)/\varepsilon^2$ on n-node Ising models. Along the way, we develop new tools for establishing concentration of functions of the Ising model, using the exchangeable pairs framework developed by Chatterjee [27], and improving upon this framework. In particular, we prove tighter concentration results for multi-linear functions of the Ising model in the high-temperature regime. We also prove a lower bound of n/ε on the sample complexity required for testing uniformity and independence of n-node Ising models.

Thesis Supervisor: Constantinos Daskalakis
Title: Associate Professor of Electrical Engineering and Computer Science

Acknowledgments

First and foremost I would like to thank my advisor Prof. Konstantinos Daskalakis for providing the vision for this work, the many insightful discussions, his guidance and for his patience in working on this project. The constant zeal and enthusiasm displayed by him were instrumental to this work. I would like to thank my co-author on this work, Gautam Kamath, for his contribution and help throughout the duration of the project. I am immensely grateful to the people who have been there with me through every step of my journey, including my parents, sister and other members of my family. I would be amiss if I didn't thank Prof. Éva Tardos of Cornell University, Prof. Supratik Chakraborty and Prof. Pushpak Bhattacharya of Indian Institute of Technology Bombay for guiding my research journey during my undergraduate studies and for inspiring me to continue it after. Finally, I would like to thank Christos Tzamos, Manolis Zampetakis and other students in TOC group at MIT for the numerous insightful discussions and for making the theory lab at MIT a vibrant and fun place to work at.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Testing properties of objects is a quintessential task in science. The scientific question of testing is the following: Does the object O have the property P? One could ask this question in the case when the object of interest O is a probability distribution. Complete knowledge of the distribution would make this a purely computational problem. But in many cases, only partial access to the object is available. This partial access could be via samples drawn from the distribution, for instance. Such a setting occurs naturally in instances when we are given data from a study or an experiment and we wish to understand the properties of the underlying distribution. Under this style of a sampling model, testing properties if distributions has a long history in statistics, since the early days; for some old and some more recent references see, e.g., [9, 10, 11, 12]. Traditionally, the emphasis has been on the asymptotic analysis of tests, pinning down their error exponents as the number of samples tends to infinity [12, 13]. In the last two decades or so, distribution testing has also piqued the interest of theoretical computer scientists, where the emphasis has been different [14, 15, 16, 17, 18, 19, 20]. In contrast to much of the statistics literature, the goal has been to minimize the number of samples required for testing. Apart from being motivated by real world problems involving data, the field of distributional property testing has revealed that many interesting properties of distributions only require a few samples, often resulting in a sample complexity which is sub-linear in the size of the support of the distribution. This points to the idea that to perform property testing we do not need to learn the distribution in its entirety and there is something more efficient which can be done instead.

Before we state what property testing is in more formal terms we will talk a bit about the notion of a distribution being ε-far from another one. It is natural to define a distance between two distributions only if they are supported on the same domain. Many notions of distances exist in literature and one of the most common distance used is total variation distance (TV distance) which is also a metric. Other common distances such as Kullback-Liebler divergence (KL-divergence), Wasserstein distance are also considered in literature. Note that the KL divergence is not a metric (see Remark 1). We will describe these distance notions in greater detail in Section 2.1.

Next, we describe what a property means in a mathematical sense. Formally, a property is any set of distributions. All distributions q in the set are said to have the property. For instance, the property of uniformity is the singleton set containing the uniform distribution. Henceforth, when we refer to a property $\mathcal{P}$ we will be referring to the set of distributions q which have the property.

Using any distance measure, we can extend the notion of distance between distributions to define distance between a distribution and a property as follows:

Definition 1. *The distance $d(p, \mathcal{P})$ between a distribution X and a property $\mathcal{P}$ is defined as*

$$d(p, \mathcal{P}) \triangleq \inf_{q \in \mathcal{P}} d(p, q).$$

That is the distance between p and property $\mathcal{P}$ is the smallest possible distance between p and q, where q is any distribution having the property $\mathcal{P}$. The broad question we will be interested in is the following:

> Given $0 < \varepsilon \leq 1$ and i.i.d. sample access to an unknown distribution p supported over a known domain D, how many samples are required to test whether $p \in \mathcal{P}$ or is ε-*far* from every distribution which has property $\mathcal{P}$, i.e. $d(p, \mathcal{P}) \geq \varepsilon$ for some distance $d()$ of interest, with a probability of success at least $2/3$?

$\frac{2}{3}$ is an arbitrary choice of a constant, except that it is bounded away from $\frac{1}{2}$. It can always be boosted to some arbitrary $1 - \delta$ at the expense of a factor $O(\log 1/\delta)$ in the sample complexity. Our focus will be on discrete distributions supported over a finite domain. We will require our testers to be efficient in terms of both the number of samples they use and the amount of time they take. The focus in much of the literature has been on optimizing the sample complexity of the tester. With regards to the time complexity, we will be happy if our tests run in polynomial time and will not focus on optimizing this polynomial.

For the most general properties, it is folklore that $\Omega(|D|)$ samples are necessary to test. However many properties of interest can be tested using a number of samples which is sub-linear in $|D|$. Some of the well-known properties which have been studied in the above setting are uniformity, goodness-of-fit with respect to a known distribution q (also known as identity testing), monotonicity and log-concavity of distributions over an ordered domain D. It is known that all these properties can be tested with $O\left(\sqrt{|D|}/\varepsilon^2\right)$ samples from p [18].

There are many natural variants and generalizations of the testing question stated above based on variations in the underlying sampling model, the accuracy parameters in the problem and the power of the tester. We will describe the major ones here. In the following we will assume that access to the distribution of interest p is provided via some kind of an oracle ORACLE_p known as the sampling oracle.

Types of Sampling Oracles: The sampling model describes how we are allowed access to the distribution we wish to test. A common assumption to make is that we are given access via independent and identically distributed samples from p. That is each query the tester makes returns an i.i.d sample from p. Another popular model is conditional sampling known as the COND model. It was introduced independently by [5] and [6]. Under COND we can specify a subset Ω_S of the domain Ω with each query and the COND oracle returns an independent sample from p conditioned on it coming from Ω_S. The COND oracle gives more power to the tester leading to much more efficient testers for many problems. For instance, under the standard i.i.d sampling model testing goodness-of-fit under the total variation distance requires $\Theta(\sqrt{n}/\varepsilon^2)$ samples whereas under the COND model it requires only $\widetilde{O}(1/\varepsilon^2)$ samples [7].

Other sampling oracles break the independence assumption. For instance, one might consider an oracle which outputs a sequence of samples which come from an underlying Markov chain.

Tolerant vs Non-Tolerant Testing Testing is a promise problem, i.e., the tester is promised that the unknown distribution p either has a property $\mathcal{P}$, or is $\geq \varepsilon$-far from it. Tolerant testing, defined and formalized by [8] is a variant with a 'softer' promise as stated below.

Given $0 < \varepsilon_1, \varepsilon_2 \leq 1$ and i.i.d. sample access to an unknown distribution p supported over a known discrete domain D, how many samples are required to test whether $d(p, \mathcal{P}) \leq \varepsilon_1$ or $d(p, \mathcal{P}) \geq \varepsilon_2$ for some distance of interest $d()$, with a probability of success at least $2/3$?

Adaptivity of the tester: Another important variation is obtained based on whether the tester is allowed to be adaptive or not. A non-adaptive tester cannot use the answers given by the oracle to its previous queries to formulate the next query. In essence, it must list all its queries at once at the start of the algorithm. An adaptive tester on the other hand can formulate the n^{th} query based on the answers given by the oracle to its previous $n-1$ queries hence making it more powerful. Note that if the oracle is an i.i.d. sampler it doesn't matter whether our tester is adaptive or not. However under the COND model adaptive testers can potentially perform better than non-adaptive ones.

Although all the aforementioned variants offer interesting questions to study, in this thesis, our focus will be on the originally stated property testing question which is the non-tolerant, non-adaptive version with access to i.i.d samples. Most of the literature in this setting has focused on testing properties of single-dimensional distributions [18, 19]. Much less is known about property testing of high-dimensional distributions and this will be our focus in this thesis. The problems of interest for us will be high-dimensional goodness-of-fit and independence testing.

From this vantage point, our testing problems take the following form:

Goodness-of-fit (or Identity) Testing: Given i.i.d sample access to an unknown distribution p over Σ^n and a parameter $\varepsilon > 0$, the goal is to distinguish with probability at least 2/3 between $p = q$ and $d(p, q) > \varepsilon$, for some specific distribution q, from as few samples as possible.

Independence Testing: Given i.i.d sample access to an unknown distribution p over Σ^n and a parameter $\varepsilon > 0$, the goal is to distinguish with probability at least 2/3 between $p \in \mathcal{I}(\Sigma^n)$ and $d(p, \mathcal{I}(\Sigma^n)) > \varepsilon$, where $\mathcal{I}(\Sigma^n)$ is the set of product distributions over Σ^n, from as few samples as possible.

In these problem definitions, Σ is some discrete alphabet.

For both testing problems, recent work has identified tight upper and lower bounds on their sample complexity [15, 17, 18, 20]: when d is taken to be the total variation distance, the optimal sample complexity for both problems turns out to be $\Theta\left(\frac{|\Sigma|^{n/2}}{\varepsilon^2}\right)$, i.e. exponential in the dimension. As modern applications commonly involve high-dimensional data, this curse of dimensionality makes the above testing goals practically unattainable. Nevertheless, there *is* a sliver of hope, and it lies with the nature of all known sample-complexity lower bounds, which construct highly-correlated distributions that are hard to distinguish from the set of independent distributions [18, 20], or from a particular distribution q [15]. Worstcase analysis of this sort seems overly pessimistic, as these instances are unlikely to arise in real-world data. As such, we propose testing high-dimensional distributions which are *structured*, and thus could potentially rule out such adversarial distributions.

Motivated by the above considerations and the ubiquity of Markov Random Fields (MRFs) in the modeling of high-dimensional distributions (see [21] for the basics of MRFs and the references [22, 23] for a sample of applications), we initiate the study of distribution testing for the prototypical example of MRFs: *the Ising Model,* which captures all binary MRFs with node and edge potentials.[1]

Markov Random Fields: A Markov Random Field with pairwise potentials is a distribution defined by an undirected graph $G = (V, E)$ where associated with each vertex $v \in V$ is a random variable X_v taking values in some alphabet Σ. Also associated with each vertex $v \in V$ is a potential function $\phi_v : \Sigma \to [0, 1]$ and associated with each edge $e \in E$ is a potential function $\phi_e : \Sigma^2 \to [0, 1]$. The probability of a particular configuration of the nodes x_V is given by

$$p(x_V) \propto \prod_{v \in V} \phi_v(x_v) \prod_{e \in E} \phi_e(x_e),$$

where x_e refers to the restriction of the vector x to the nodes which share the edge e. MRFs defined as above have nice conditional independence properties captured by the graph structure. Specifically, if $S \subseteq V$ is a set of vertices such that every path from u to v passes through S, then conditioned on the vertices in S, X_u and X_v are independent. This conditional independence structure makes calculations tractable in many cases.

Note: All MRFs need not be restricted to only pairwise potentials. They can be defined with potential functions which take as input larger subsets of nodes as well. However in this thesis we will be interested in MRFs with pairwise potentials. Property testing on general MRFs is an interesting direction to pursue (refer to Section 8.1).

Ising models are a canonical example of MRFs where the random variables for each node are over a binary alphabet. They were originally used in physics for modeling magnetization of atoms in a lattice but have found applications in computer science and other fields where graphical models are used. For instance, the problem of learning the graph structure of an Ising model is a question of significant interest [24]. The potential functions ϕ_v and ϕ_e for Ising models take a very particular form yielding the following probability for a particular

[1] This follows trivially by the definition of MRFs, and elementary Fourier analysis of Boolean functions.

configuration x_V

$$p(x) = \exp\left(\sum_{v \in V} \theta_v x_v + \sum_{(u,v) \in E} \theta_{u,v} x_u x_v - \Phi(\vec{\theta})\right), \tag{1.1}$$

where $\Phi(\vec{\theta})$ is the log-partition function, ensuring that the distribution is normalized. Intuitively, there is a random variable X_v sitting on every node of G, which may be in one of two states, or spins: up (+1) or down (-1). The scalar parameter θ_v models a local (or "external") field at node v. The sign of θ_v represents whether this local field favors X_v taking the value $+1$, i.e. the up spin, when $\theta_v > 0$, or the value -1, i.e. the down spin, when $\theta_v < 0$, and its magnitude represents the strength of the local field. We will say a model is "without external field" when $\theta_v = 0$ for all $v \in V$. Similarly, $\theta_{u,v}$ represents the direct interaction between nodes u and v. Its sign represents whether it favors equal spins, when $\theta_{u,v} > 0$, or opposite spins, when $\theta_{u,v} < 0$, and its magnitude corresponds to the strength of the direct interaction. Of course, depending on the structure of the Ising model and the edge parameters, there may be indirect interactions between nodes, which may overwhelm local fields or direct interactions.

The Ising model has a rich history, starting with its introduction by statistical physicists as a probabilistic model to study phase transitions in spin systems [25]. Since then it has found a myriad of applications in diverse research disciplines, including probability theory, Markov chain Monte Carlo, computer vision, theoretical computer science, social network analysis, game theory, and computational biology [26, 27, 28, 29, 30, 31, 32]. The ubiquity of these applications motivate the problem of inferring Ising models from samples, or inferring statistical properties of Ising models from samples. This type of problem has enjoyed much study in statistics, machine learning, and information theory, see, i.e., [33, 34, 35, 36, 37, 38, 39, 24, 40, 41, 42, 43]. Much of prior work has focused on *parameter learning*, where the goal is to determine the parameters of an Ising model to which sample access is given. In contrast to this type of work, which focuses on discerning *parametrically* distant Ising models, our goal is to discern *statistically* distant Ising models, in the hopes of dramatic improvements in the sample complexity. (We will come to a detailed comparison between the two inference goals shortly, after we have stated our results.) To be precise, we study the following problems:

Ising Model Goodness-of-fit (or Identity) Testing: Given sample access to an unknown Ising model p (with unknown parameters over an unknown graph) and a parameter $\varepsilon > 0$, the goal is to distinguish with probability at least $2/3$ between $p = q$ and $d_{\mathrm{SKL}}(p, q) > \varepsilon$, for some specific Ising model q, from as few samples as possible.

Ising Model Independence Testing: Given sample access to an unknown Ising model p (with unknown parameters over an unknown graph) and a parameter $\varepsilon > 0$, the goal is to distinguish with probability at least $2/3$ between $p \in \mathcal{I}_n$ and $d_{\mathrm{SKL}}(p, \mathcal{I}_n) > \varepsilon$, where $\mathcal{I}_n$ are all product distributions over $\{-1, 1\}^n$, from as few samples as possible.

We note that there are several potential notions of statistical distance one could consider — classically, total variation distance and the Kullback-Leibler (KL) divergence have seen the most study. As our focus here is on upper bounds, we consider the symmetrized KL divergence d_{SKL}, which is a "harder" notion of distance than both: in particular, testers for d_{SKL} immediately imply testers for both total variation distance and the KL divergence. Moreover, by virtue of the fact that d_{SKL} upper-bounds KL in both directions, our tests offer useful information-theoretic interpretations of rejecting a model q, such as data differencing and large deviation bounds in both directions.

Sample Application: As an instantiation of our proposed testing problems for the Ising model one may maintain the study of strategic behavior on a social network. To offer a little bit of background, a body of work in economics has modeled strategic behavior on a social network as the evolution of the Glauber dynamics of an Ising model, whose graph is the social network, and whose parameters are related to the payoffs of the nodes under different selections of actions by them and their neighbors. For example, [31, 32] employ this model to study the adoption of competing technologies with network effects, e.g. iPhone versus Android phones. Glauber dynamics, as described in Section 2.4.2, is the canonical Markov chain for sampling an Ising model. Hence an observation of the actions (e.g. technologies) used by the nodes of the social network should offer us a sample from the corresponding Ising model (at least if the Glauber dynamics have mixed; see also Lemma 24 in Section C for a bound on the mixing time of Glauber dynamics). An analyst may not know the underlying social network or may know the social network but not the parameters of the underlying Ising model. In either case, how many independent observations would he need to test, e.g., whether the nodes are adopting technologies independently, or whether their adoptions conform to some conjectured parameters? Our results offer algorithms for testing such hypotheses in this stylized model of strategic behavior on a network. Similar applications can be found in other domains where Ising models have been a common modeling device, such as computer vision and computational biology.

1.1 Results and Techniques

The main result of this thesis is the following:

Theorem 1. *Both Ising Model Goodness-of-fit Testing and Ising Model Independence Testing can be solved from* $\text{poly}\left(n, \frac{1}{\varepsilon}\right)$ *samples in polynomial time.*

There are several variants of our testing problems, resulting from different knowledge that the analyst may have about the structure of the graph (connectivity, density), the nature of the interactions (attracting, repulsing, or mixed), as well as the temperature (low vs high). We proceed to discuss all these variants, instantiating the resulting polynomial sample complexity in the above theorem. We also illuminate the techniques involved to prove these theorems. This discussion should suffice in evaluating the merits of the results and techniques of this thesis.

A. Our Baseline Result. In the least favorable regime, i.e. when the analyst is oblivious to the structure of the Ising model p, the signs of the interactions, and their strength, the polynomial in Theorem 1 becomes $O\left(\frac{n^4\beta^2+n^2h^2}{\varepsilon^2}\right)$. In this expression, $\beta = \max\{|\theta^p_{u,v}|\}$ for independence testing, and $\beta = \max\{\max\{|\theta^p_{u,v}|\}, \max\{|\theta^q_{u,v}|\}\}$ for goodness-of-fit testing, while $h = 0$ for independence testing, and $h = \max\{\max\{|\theta^p_u|\}, \max\{|\theta^q_u|\}\}$ for goodness-of-fit testing; see Theorem 2. If the analyst has an upper bound on the maximum degree $d_{\max}$ (of all Ising models involved in the problem) the dependence improves to $O\left(\frac{n^2d^2_{\max}\beta^2+nd_{\max}h^2}{\varepsilon^2}\right)$, while if the analyst has an upper bound on the total number of edges m, then $\max\{m, n\}$ takes the role of $nd_{\max}$ in the previous bound; see Theorem 2.

Technical Discussion 1.0: "Testing via Localization." All the bounds mentioned so far are obtained via a simple localization argument showing that, whenever two Ising models p and q satisfy $d_{\text{SKL}}(p, q) > \varepsilon$, then "we can blame it on a node or an edge;" i.e. there exists a node with significantly different bias under p and q or a pair of nodes u, v whose covariance is significantly different under the the two models. Pairwise correlation tests are a simple screening that is often employed in practice. For our setting, there is a straighforward and elegant way to show that pair-wise (and not higher-order) correlation tests suffice; see Lemma 5.

For more details about our baseline localization tester see Section 3.

B. Anchoring Our Expectations. The next results aim at improving the afore-described baseline bound. Before stating these improvements, however, it is worth comparing the sample complexity of our baseline results to the sample complexity of learning. Indeed, one might expect and it is often the case that testing problems can be solved in a two-step fashion, by first learning a hypothesis $\hat{p}$ that is close to the true p and then using the learned hypothesis $\hat{p}$ as a proxy for p to determine whether it is close to or far from some q, or some set of distributions. Given that the KL divergence and its symmetrized version do not satisfy the triangle inequality, however, it is not clear how such an approach would work. Even if it could, the only algorithm that we are aware of for proper learning Ising models, which offers KL divergence guarantees but does not scale exponentially with the maximum degree and β, is a straightforward net-based algorithm. This algorithm, explained in Section B, requires $\Omega\left(\frac{n^6\beta^2+n^4h^2}{\varepsilon^2}\right)$ samples and is time inefficient. In particular, the baseline localization algorithm already beats this sample complexity and is also time-efficient. Alternatively, one could aim to parameter-learn p; see, e.g., [38, 24, 40] and their references. However, these algorithms require sample complexity that is exponential in the maximum degree [38], and they typically use samples exponential in β as well [24, 40]. For instance, if we use [40], which is one of the state-of-the-art algorithms, to do parameter learning prior to testing, we would need $\tilde{O}(\frac{n^4 \cdot 2^{\beta \cdot d_{\max}}}{\varepsilon^2})$ samples to learn p's parameters closely enough to be able to do the testing afterwards. Our baseline result beats this sample complexity, dramatically so if the degrees are unbounded.

D. The High-Temperature Regime. Motivated by phenomena in the physical world, the study of Ising models has identified phase transitions in the behavior of the model as its parameters vary. A common transition occurs as the temperature of the model changes from low to high. As the parameters $\vec{\theta}$ correspond to inverse (individualistic) temperatures, this corresponds to a transition of these parameters from low values (high temperature) to high values (low temperature). Often the transition to high temperature is identified with the satisfaction of Dobrushin-type conditions [44]. Under such conditions, the model enjoys a number of good properties, including rapid mixing of the Glauber dynamics, spatial mixing properties, and uniqueness of measure. For some background, in Section C, we show the rapid mixing of the Glauber dynamics, when $\max\{|\theta_{u,v}|\} = O(1/d_{\max})$, which corresponds to one of the most commonly studied high temperature regimes and the one we will adopt in this thesis.[2] We also show some basic facts about concentration of Lipschitz functions $f(X_V)$

[2] In fact, we show this for a more general condition stated in Definition 23. All our results for the high temperature regime can be extended to this more general condition, but we refrain from studying such generalizations to avoid making the notation in our proofs unnecessarily complicated.

of the variables X_V of an Ising model in the high temperature regime. Both the mixing time bound and the concentration result are easy adaptations of Chatterjee's framework [27] so we do not claim them as contributions of our work. They can also be skipped when reading this thesis, as they are only meant to provide background.

In the high-temperature regime, we show that we can improve our baseline result using a non-localization based argument, explained next. In particular, we show in Theorem 4 that under high temperature and with no external fields independence testing can be done computationally efficiently from $\tilde{O}\left(\max\left\{\frac{n^{10/3}}{\varepsilon^2 d_{\max}^2}, \frac{n^{11/3}}{\varepsilon^2 d_{\max}^{2.5}}\right\}\right)$ samples, which improves upon our baseline result if $d_{\max}$ is large enough. For instance, when $d_{\max} = \Omega(n)$, the sample complexity becomes $\tilde{O}\left(\frac{n^{4/3}}{\varepsilon^2}\right)$. Other tradeoffs between β, $d_{\max}$ and the sample complexity are explored in Theorem 3. Similar improvements hold when external fields are present (Theorem 6), as well as for identity testing, without and with external fields (Theorems 7 and 8).

We offer some intuition about the improvements in Figures 6-1 and 6-2 (appearing in Section 6), which are plotted for high temperature and no external fields. In Figure 6-1, we plot the number of samples required for testing Ising models with no external fields when $\beta = \Theta(\frac{1}{d_{\max}})$ as $d_{\max}$ varies. The horizontal axis is $\log_n d_{\max}$. We see that localization is the better algorithm for degrees smaller than $O(n^{2/3})$, above which its complexity can be improved. In particular, the sample complexity is $O(n^2/\varepsilon^2)$ until degree $d_{\max} = O(n^{2/3})$, beyond which it drops inverse quadratically in $d_{\max}$. In Figure 6-2, we consider a different tradeoff. We plot the number of samples required when $\beta = n^{-\alpha}$ and the degree of the graph varies. In particular, we see three regimes as a function of whether the Ising model is in high temperature ($d_{\max} = O(n^a)$) or low temperature ($d_{\max} = \omega(n^a)$), and also which of our techniques localization vs non-localization gives better sample complexity bounds.

Technical Discussion 3.0: "Testing via A Global Statistic." One way or another all our results so far had been obtained via localization, namely blaming the distance of p from independence, or from some distribution q to a node or an edge. Our improved bounds employ non-localized statistics that look at all the nodes of the Ising model simultaneously. Specifically, we employ statistics of the form $Z = \sum_{e=(u,v)\in E} c_e X_u X_v$ for some appropriately chosen signs c_e.

The first challenge we encounter here involves selecting the signs c_e in accordance with the sign of each edge marginal's expectation, $\mathbf{E}[X_u X_v]$. This is crucial to establish that the resulting statistic will be able to discern between the two hypotheses. While the necessary estimates of these signs could be computed independently for each edge, this would incur an unnecessary overhead of $O(n^2)$ in the number of samples. Instead we try to learn these signs from fewer samples. Despite the terms potentially having nasty correlations with each other, a careful analysis using anti-concentration calculations allows us to sidestep this cost and generate satisfactory estimates with a non-negligible probability, from fewer samples.

The second and more significant challenge involves bounding the variance of a statistic Z of the above form. Since Z's magnitude is at most $O(n^2)$, its variance can naively be bounded by $O(n^4)$. However, applying this bound in our algorithm gives a vacuous sample complexity. We require more work to arrive at useful bounds, and surprisingly, in fairly general regimes, we can show the variance to be $\tilde{O}(n^2)$. Stated another way, despite the complex correlations which may be present in the Ising model, the summands in Z behave roughly as if they were independent. In order to prove this result, we draw inspiration from the method of exchangeable pairs used in Chatterjee's thesis [27]. This method involves defining a coupling between two evolutions of the Glauber dynamics for the Ising model and demonstrating contraction of an appropriate statistic. Our analysis requires the definition of a new coupling and more careful contraction arguments, but allows us to show a variance which is up to a factor of $\tilde{O}(n)$ better than one would get by applying Chatterjee's arguments directly. We consider our techniques here to be a significant contribution, and we expect that they will be applied to analysis of other complex random structures which may be sampled by rapidly mixing Markov chains. Our technique is described in Section 5. Our variance bounds vary slightly depending on whether an external field is present and the bounds are given in Theorems 9 and 10.

E. Lower Bounds. The proof of our linear lower bound applies Le Cam's method [45]. Our construction is inspired by Paninski's lower bound for uniformity testing [15], which involves pairing up domain elements and jointly perturbing their probabilities. This style of construction is ubiquitous in univariate testing lower bounds. A naive application of this approach would involve choosing a fixed matching of the nodes and randomly perturbing the weight of the edges, which leads to an $\Omega(\sqrt{n})$ lower bound. To achieve the linear lower bound, we instead consider a *random* matching of the nodes. The analysis of this case turns out to be involved due to the complex structure of the probability function which corresponds to drawing k samples from an Ising model on a randomly chosen matching. Indeed, our proof turns out to have a significantly combinatorial flavor, and we believe that our techniques might be helpful for proving stronger lower bounds in combinatorial settings for multivariate distributions. See Theorem 15 for the formal statement of our main lower bound. As mentioned before, we also show that the sample complexity must depend on β

Testing Problem	**No External Field**	**Arbitrary External Field**
Independence using Localization	$\tilde{O}\left(\frac{n^2 d_{\max}^2 \beta^2}{\varepsilon^2}\right)$	$\tilde{O}\left(\frac{n^2 d_{\max}^2 \beta^2}{\varepsilon^2}\right)$
Identity using Localization	$\tilde{O}\left(\frac{n^2 d_{\max}^2 \beta^2}{\varepsilon^2}\right)$	$\tilde{O}\left(\frac{n^2 d_{\max}^2 \beta^2}{\varepsilon^2} + \frac{n^2 h^2}{\varepsilon^2}\right)$
Independence in high temperature using Learn-Then-Test	$\tilde{O}\left(\frac{n^{8/3} \max\{n^{2/3}, n\beta d_{\max}^{0.5}\}\beta^2}{\varepsilon^2}\right)$	$\tilde{O}\left(\frac{n^{8/3} \max\{n^{2/3}, n\beta^{2/3} d_{\max}^{1/3}\}\beta^2}{\varepsilon^2}\right)$
Identity in high temperature using Learn-Then-Test	$\tilde{O}\left(\frac{n^{8/3} \max\{n^{2/3}, n\beta d_{\max}^{0.5}\}\beta^2}{\varepsilon^2}\right)$	$\tilde{O}\left(\frac{n^{11/3}\beta^2}{\varepsilon^2} + \frac{n^{5/3} h^2}{\varepsilon^2}\right)$

Table 1.1: Summary of our results in terms of the sample complexity upper bounds for the various problems studied. n = number of nodes in the graph, $d_{\max}$ = maximum degree, β = maximum absolute value of edge parameters and h = maximum absolute value of node parameters (when applicable).

and h in certain cases, see Theorem 16 for a formal statement.

Table 1.1 summarizes our algorithmic results.

1.2 Thesis Organization

In Chapter 2, we discuss preliminaries. In Chapter 3, we give a simple localization-based algorithm. In Chapter 4, we describe our main algorithm. In Chapter 5, we discuss our technique for bounding the variance of statistics over the Ising model. In Chapter 6, we compare the localization and the learn-then-test algorithms and note the regimes under which one performs better than the other. Finally, in Chapter 7 our lower bound is presented. Some details in the above sections are deferred to the supplementary material.

Chapter 2

Preliminaries

In this chapter, we state some preliminaries which are intended as a background for the rest of the thesis. We will describe formally what an Ising model is and the terminology related to Ising models. We will also set down notation which will be used through the rest of this thesis. Knowledge of basic probability, random variables and Markov chains will be assumed. We start with a description of a folklore result about Rademacher random variables.

Rademacher random variables are binary random variables where *Rademacher*(p) takes value 1 with probability p, and -1 otherwise. We will use the following folklore result on estimating the parameter p of a Rademacher random variable.

Lemma 1. *Given iid random variables $X_1, \ldots, X_k \sim$ Rademacher(p) for $k = O(\log(1/\delta)/\varepsilon^2)$, there exists an algorithm which obtains an estimate $\hat{p}$ such that $|\hat{p} - p| \leq \varepsilon$ with probability $1 - \delta$.*

Next, we describe of some commonly used notions of distance between distributions.

2.1 Distance Between Distributions

A statistical distance function quantifies the distance between two statistical objects. Here the objects of interest are distributions. In the following sections we list some of the commonly used distance measures for distributions. We will focus on distributions supported over a discrete space. We denote the two distributions being considered as p and q and their common support by S (if they have different supports, take S to be the union of the two supports).

2.1.1 Total Variation Distance

This is one of the most widely used metric. Denoted by $d_{\mathrm{TV}}(.,.)$, it is defined as

$$d_{\mathrm{TV}}(p,q) = \frac{1}{2} \sum_{i \in S} |p(i) - q(i)| \,.$$

2.1.2 Kolmogorov Distance

Let P and Q represent the cumulative distribution functions of distributions p and q respectively. The Kolmogorov distance between p and q, $d_K(p,q)$, is defined as

$$d_K(p,q) = \sup_x |P(x) - Q(x)| \,.$$

It can be seen that the Kolmogorov distance lower bounds the total variation distance, i.e., $d_K(p,q) \le d_{\mathrm{TV}}(p,q)$.

2.1.3 Hellinger Distance

The Hellinger distance, denoted by $d_H(.,.)$ is defined as

$$d_H(p,q) = \sqrt{\frac{1}{2} \sum_{i \in S} \left(\sqrt{p(i)} - \sqrt{q(i)} \right)^2}.$$

It is closely related to the total variation distance as stated in the following inequalities

$$d_H^2(p,q) \le d_{\mathrm{TV}}(p,q) \le \sqrt{2} d_H(p,q).$$

2.1.4 Wasserstein Distance

Also known as the Earthmover distance in the discrete setting, the Wasserstein distance $d_W(.,.)$ is defined as the minimum cost required to move probability mass around in one distribution so as to make it identical to the second distribution, where cost is computed as the mass times the distance it has to be moved. Computing the Wasserstein distance between two distributions typically involves solving a linear program.

2.1.5 Kullback-Liebler Divergence

A popular distance in information theory literature is the non-symmetric KL-divergence. Measuring the relative entropy of p relative to q, the KL-divergence between p and q is defined as

$$d_{\mathrm{KL}}(p,q) = \mathbf{E}_p\left[\log\left(\frac{p}{q}\right)\right].$$

2.1.6 Symmetrized Kullback-Liebler Divergence

There are a number of ways to symmetrize KL-divergence. We present here the definition which is of interest to us in this thesis. Denoted by $d_{\mathrm{SKL}}(.,.)$, the symmetrized KL-divergence is defined as,

$$d_{\mathrm{SKL}}(p,q) = d_{\mathrm{KL}}(p,q) + d_{\mathrm{KL}}(q,p) = \mathbf{E}_p\left[\log\left(\frac{p}{q}\right)\right] + \mathbf{E}_q\left[\log\left(\frac{q}{p}\right)\right].$$

Remark 1. *Many distance functions are metrics. A distance function $d(.,.)$ on a set X is a metric if it satisfies the following properties. For all $x, y, z \in X$,*

- $d(x,y) \geq 0$ *(non-negativity)*
- $d(x,y) = 0 \iff x = y$ *(identity of indiscernibles)*
- $d(x,y) = d(y,x)$ *(symmetry)*
- $d(x,y) + d(y,z) \geq d(x,z)$ *(triangle inequality / sub-additivity)*

The total variation distance, Kolmogorov distance and Hellinger distance are metrics for instance.
Divergences are distance functions which satisfy only the first two properties in the above list. They can be asymmetric and may violate the triangle inequality. The KL-divergence is a popular example.

We will use without proof the following well-known result regarding relations between distance measures on probability distributions.

Lemma 2 (Pinsker's Inequality)**.** *For any two distributions p and q, we have the following relation between their total variation distance and their KL-divergence,*

$$2d_{\mathrm{TV}}^2(p,q) \leq d_{\mathrm{KL}}(p||q).$$

Also since $d_{\mathrm{KL}}(p||q) \geq 0$ for any distributions P and Q, we have

$$d_{\mathrm{SKL}}(p,q) \geq d_{\mathrm{KL}}(p||q) \geq 2d_{\mathrm{TV}}^2(p,q). \tag{2.1}$$

Hence the symmetric KL-divergence between two distributions upper bounds both the KL-divergence and total variation (TV) distance between them under appropriate scaling.

2.2 Concentration Inequalities

Concentration inequalities or tail inequalities bound the probability of a random variable being far from its mean in terms of its moments. We will state a couple of well-known tail inequalities which we use in this thesis.

The first is Chebyshev's inequality which gives a tail bound using the second moment of a random variable.

Lemma 3 (Chebyshev's Inequality)**.** *Let X be a random variable with a finite expected value $\mathbf{E}[X]$ and a finite non-zero variance $\mathbf{Var}[X]$, then for all $t > 0$*

$$\Pr\left[|X - \mathbf{E}[X]| \geq t\right] \leq \frac{\mathbf{Var}[X]}{t^2}.$$

The second is the stronger Chernoff bound which holds when looking at sums of independent random variables. There are many forms of the Chernoff bound which hold in different settings. We state one of them here.

Lemma 4 (Chernoff Bound)**.** *Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables (taking values in $\{0, 1\}$) and let $X = X_1 + X_2 + \ldots + X_n$. Also let $\mu = \mathbf{E}[X]$. Then,*

$$\Pr[|X - \mu| \geq t] \leq \exp\left(-\frac{t^2}{3\mu}\right)$$

2.3 Markov Chain Monte Carlo Sampling

Sampling from a high dimensional distribution can be an expensive task in practice. The reason being the exponential support size. A popular technique to get around this issue is Markov Chain Monte Carlo (MCMC) sampling wherein a Markov chain M is defined with one node for each possible support element of the distribution that we wish to sample from. M has the property that it is fast mixing and more crucially, the stationary distribution is the high dimensional distribution we wish to sample from. The sampling procedure simulates a run of the chain for mixing time number of steps and outputs the final state as a sample. This will represent a sample from a distribution which is statistically close to the desired distribution.

Apart from being a sampling tool, MCMC also enables us to prove some properties of the distribution in elegant ways. A popular Markov chain which performs MCMC sampling for Ising models is the Glauber dynamics which will be described in detail in Section 2.4.2. We will show some important properties of functions on the Ising model using the Glauber dynamics.

2.4 Ising Models

Ising models were first introduced in physics for modeling the spin interactions between atoms in a lattice. We do away with the lattice assumption in this thesis. We consider the Ising model on a graph $G = (V, E)$ with n nodes. It is a distribution over $\{\pm 1\}^n$, with a parameter vector $\vec{\theta} \in \mathbb{R}^{|V|+|E|}$. $\vec{\theta}$ has a parameter corresponding to each edge $e \in E$ and each node $v \in V$. The probability mass function assigned to a string x is

$$P(x) = \exp\left(\sum_{v \in V} \theta_v x_v + \sum_{e=(u,v) \in E} \theta_e x_u x_v - \Phi(\vec{\theta})\right),$$

where $\Phi(\vec{\theta})$ is the log-partition function for the distribution. The edge parameters signify the strength of the influence a node has on its neighbours. In physical systems, this interaction strength is inversely related to the temperature at which the lattice exists which leads to one of the important parameters related to an Ising model. Let β denote the maximum edge interaction strength (when looking at absolute values). $1/\beta$ is defined as the temperature of the Ising model. Hence a large value of β indicates a low temperature and vice-versa. A critical value of β at which the Ising model's behavior exhibits a phase transition is $\frac{\eta}{4d_{\max}}$ where η is a constant. If β is smaller than this critical value, we consider the Ising model to be in the high-temperature regime.

Definition 2. *In the* high-temperature regime*, for all $e \in E$, $\theta_e \leq \frac{\eta}{4d_{\max}}$, where $\eta < 1$ is a constant.*

Intuitively, in high temperature the atoms are in higher energy states and their spin alignments are less affected by their neighbours. This regime is a well-studied one as Ising models exhibit distinct behaviors depending on the temperature. In this thesis, we will see that in the high-temperature regime we can exploit the limited interaction strength to improve the sample complexities of our tests.

We will abuse notation, referring to both the probability distribution p and the random vector X that it samples in $\{\pm 1\}^V$ as the Ising model. That is, $X \sim p$. We will use X_u to denote the variable corresponding to node u in the Ising model X. When considering multiple samples from an Ising model X, we will use $X^{(l)}$ to denote the l^{th} sample. We will use h to denote the largest node parameter in absolute value and β to denote the largest edge parameter in absolute value. That is, $|\theta_v| \leq h$ for all $v \in V$ and $|\theta_e| \leq \beta$ for all $e \in E$. Depending on the setting, our results will depend on h and β. Furthermore, in this thesis we will use the convention that $E = \{(u, v) \mid u, v \in V \wedge u \neq v\}$ and θ_e may be equal to 0, indicating that edge e is not present in the graph. We use m to denote the number of edges with non-zero parameters in the graph, and $d_{\max}$ to denote the maximum degree of a node.

Throughout this thesis, we will use the notation $\mu_v \triangleq \mathbf{E}[X_v]$ for the marginal expectation of a node $v \in V$ (also called node marginal), and similarly $\mu_{uv} \triangleq \mathbf{E}[X_u X_v]$ for the marginal expectation of an edge $e = (u, v) \in E$ (also called edge marginal). In case a context includes multiple Ising models, we will use μ_e^p to refer to the marginal expectation of an edge e under the model p.

We will use $\mathcal{U}_n$ to denote the uniform distribution over $\{\pm 1\}^n$, which also corresponds to the Ising model with $\vec{\theta} = \vec{0}$. Similarly, we use $\mathcal{I}_n$ for the set of all product distributions over $\{\pm 1\}^n$.

When $\vec{p}$ and $\vec{q}$ are vectors, we will write $\vec{p} \leq \vec{q}$ to mean that $p_i \leq q_i$ for all i.

Definition 3. *In the setting with* no external field*,* $\theta_v = 0$ *for all* $v \in V$*.*

Definition 4. *In the* ferromagnetic *setting,* $\theta_e \geq 0$ *for all* $e \in E$*.*

2.4.1 Symmetric KL Divergence Between Two Ising Models

We note that the symmetric KL divergence between two Ising models p and q admits a very convenient expression [38]:

$$d_{\mathrm{SKL}}(p, q) = \sum_{v \in V} \left(\theta_v^p - \theta_v^q\right)\left(\mu_v^p - \mu_v^q\right) + \sum_{e=(u,v) \in E} \left(\theta_e^p - \theta_e^q\right)\left(\mu_e^p - \mu_e^q\right). \tag{2.2}$$

This expression will form the basis for all our algorithms.

2.4.2 Glauber Dynamics

Glauber dynamics is the canonical Markov chain for sampling from an Ising model. We consider the basic variant known as single-site Glauber dynamics here. The dynamics are a Markov chain defined on the set Σ^n where $\Sigma = \{-1, +1\}$. They proceed as follows:

1. Start at any state $X^{(0)} \in \Sigma^n$. Let $X^{(t)}$ denote the state of the dynamics at time t.
2. Let $N(u)$ denote the set of neighbors of node u. Pick a node u uniformly at random and update X as follows:

$$X_u^{(t+1)} = 1 \quad \text{w.p.} \quad \frac{\exp\left(\theta_u + \sum_{v \in N(u)} \theta_{uv} X_v^{(t)}\right)}{\exp\left(\theta_u + \sum_{v \in N(u)} \theta_{uv} X_v^{(t)}\right) + \exp\left(-\theta_u - \sum_{v \in N(u)} \theta_{uv} X_v^{(t)}\right)}$$

$$X_u^{(t+1)} = -1 \quad \text{w.p.} \quad \frac{\exp\left(-\theta_u - \sum_{v \in N(u)} \theta_{uv} X_v^{(t)}\right)}{\exp\left(\theta_u + \sum_{v \in N(u)} \theta_{uv} X_v^{(t)}\right) + \exp\left(-\theta_u - \sum_{v \in N(u)} \theta_{uv} X_v^{(t)}\right)}$$

$$X_v^{(t+1)} = X_v^{(t)} \quad \forall \quad v \neq u.$$

Glauber dynamics define a reversible Markov chain whose stationary distribution is identical to the corresponding Ising model. In many relevant settings, for instance, the high-temperature regime, the dynamics are fast mixing, i.e., they mix in time $O(n \log n)$ (Lemma 24) and hence offer an efficient way to sample from Ising models.

2.5 Input to Goodness-of-Fit Testers

To solve the goodness-of-fit testing or identity testing problem with respect to a discrete distribution q, a description of q is given as part of the input along with sample access to the distribution p which we are testing. In case q is an Ising model, its support has exponential size and specifying the vector of probability values at each point in its support is inefficient. Since q is characterized by the edge parameters between every pair of nodes and the node parameters associated with the nodes, a succinct description would be to specify the parameters vectors $\{\theta_{uv}\}, \{\theta_u\}$. In many cases, we are also interested in knowing the edge and node marginals of the model. Although these quantities can be computed from the parameter vectors, there is no efficient method known to compute the marginals exactly for general regimes. A common approach is to use MCMC sampling to generate samples from the Ising model. However, for this technique to be efficient we require that the mixing time of the Markov chain be small which is not true in general. Estimating and exact computation of the marginals of an Ising model is a well-studied problem but is not the focus of this thesis. Hence, to avoid such computational complications we will assume that for the identity testing problem the description of the Ising model q includes both the parameter vectors

$\{\theta_{uv}\}, \{\theta_u\}$ as well as the edge and node marginal vectors $\{\mu_{uv} = \mathbf{E}[X_u X_v]\}, \{\mu_u = \mathbf{E}[X_u]\}$.

Chapter 3

Localization Algorithm

Our first algorithm is a general purpose "localization" algorithm. While extremely simple, this serves as a proof-of-concept that testing on Ising models can avoid the curse of dimensionality, while simultaneously giving a very efficient algorithm for certain parameter regimes. The main observation which enables us to do a localization based approach is stated in the following Lemma, which allows us to "blame" a difference between models p and q on a discrepant node or edge.

Lemma 5. *Given two Ising models p and q, if $d_{\mathrm{SKL}}(p,q) \geq \varepsilon$, then either*

- *There exists an edge $e = (u,v)$ such that $\left(\theta^p_{uv} - \theta^q_{uv}\right)\left(\mu^p_{uv} - \mu^q_{uv}\right) \geq \frac{\varepsilon}{2m}$; or*
- *There exists a node u such that $\left(\theta^p_{u} - \theta^q_{u}\right)\left(\mu^p_{u} - \mu^q_{u}\right) \geq \frac{\varepsilon}{2n}$.*

Proof of Lemma 5: We have,

$$d_{\mathrm{SKL}}(p,q) = \sum_{e=(u,v)\in E} \left(\theta^p_e - \theta^q_e\right)\left(\mu^p_e - \mu^q_e\right) + \sum_{v\in V} \left(\theta^p_v - \theta^q_v\right)\left(\mu^p_v - \mu^q_v\right) \geq \varepsilon$$

$$\implies \sum_{e=(u,v)\in E} \left(\theta^p_e - \theta^q_e\right)\left(\mu^p_e - \mu^q_e\right) \geq \varepsilon/2 \quad \text{or} \quad \sum_{v\in V} \left(\theta^p_v - \theta^q_v\right)\left(\mu^p_v - \mu^q_v\right) \geq \varepsilon/2$$

In the first case, there has to exist an edge $e = (u,v)$ such that $\left(\theta^p_{uv} - \theta^q_{uv}\right)\left(\mu^p_{uv} - \mu^q_{uv}\right) \geq \frac{\varepsilon}{2m}$ and in the second case there has to exist a node u such that $\left(\theta^p_{u} - \theta^q_{u}\right)\left(\mu^p_{u} - \mu^q_{u}\right) \geq \frac{\varepsilon}{2n}$ thereby proving the lemma. □

Before giving a description of the localization algorithm, we state its guarantees.

Theorem 2. *Given $\tilde{O}\left(\frac{m^2\beta^2}{\varepsilon^2}\right)$ samples from an Ising model p, there exists a polynomial-time algorithm which distinguishes between the cases $p \in \mathcal{I}_n$ and $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$ with* *probability at least* $2/3$*. Furthermore, given* $\tilde{O}\left(\frac{m^2\beta^2}{\varepsilon^2} + \frac{n^2h^2}{\varepsilon^2}\right)$ *samples from an Ising model* p *and a description of an Ising model* q*, there exists a polynomial-time algorithm which distinguishes between the cases* $p = q$ *and* $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$ *with probability at least* $2/3$ *where* $\beta = \max\{|\theta_{uv}|\}$ *and* $h = \max\{|\theta_u|\}$*. If we are given as input the maximum degree of nodes in the graph* $d_{\max}$*,* m *in the above bounds is substituted by* $nd_{\max}$*.*

Note that the sample complexity achieved by the localization algorithm gets worse as the graph becomes denser. This is because as the number of possible edges in the graph grows, the contribution to the distance by any single edge grows smaller thereby making it harder to detect.

We describe the algorithm for independence testing in Section 3.1. The algorithm for testing identity is similar, its description and correctness proofs are given in Section 3.2.

3.1 Independence Test using Localization

We start with a high-level description of the algorithm. Given sample access to Ising model $X \sim p$ it will first obtain empirical estimates of the node marginals μ_u for each node $u \in V$ and edge marginals μ_{uv} for each pair of nodes (u, v). Denote these empirical estimates by $\hat{\mu}_u$ and $\hat{\mu}_{uv}$ respectively. Using these empirical estimates, the algorithm computes the empirical estimate for the covariance of each pair of variables in the Ising model. That is, it computes an empirical estimate of $\lambda_{uv} = \mathbf{E}[X_uX_v] - \mathbf{E}[X_u]\mathbf{E}[X_v]$ for all pairs (u, v). If they are all close to zero, then we can conclude that $p \in \mathcal{I}_n$. If there exists an edge for which λ_{uv} is far from 0, this indicates that p is far from $\mathcal{I}_n$. The reason for this follows from the expression Lemma 5 and is described in further detail in the proof of Lemma 7. A precise description of the test is given in in Algorithm 1 and its correctness is proven via Lemmas 6 and 7.

To prove correctness of Algorithm 1, we will require the following lemma, which allows us to detect pairs u, v for which λ_{uv} is far from 0.

Lemma 6. *Given* $O\left(\frac{\log n}{\varepsilon^2}\right)$ *samples from an Ising model* $X \sim p$*, there exists a polynomial-time algorithm which, with probability at least* $9/10$*, can identify all pairs of nodes* $(u, v) \in V^2$ *such that* $|\lambda_{uv}| \geq \varepsilon$*, where* $\lambda_{uv} = \mathbf{E}[X_uX_v] - \mathbf{E}[X_u]\mathbf{E}[X_v]$*.*

Proof. This lemma is a direct consequence of Lemma 1. Note that for any edge $e = (u, v) \in E$, $X_uX_v \sim$ *Rademacher*$((1 + \mu_e)/2)$. Also $X_u \sim$ *Rademacher*$((1 + \mu_u)/2)$ and

Algorithm 1 Test if an Ising model p is product

```
function LocalizationTest(sample access to Ising model $p$, accuracy parameter $\varepsilon, \beta, d_{\mathrm{max}}$)
    Draw $k = O\left(\frac{n^2 d_{\mathrm{max}}^2 \beta^2 \log n}{\varepsilon^2}\right)$ samples from $p$. Denote the samples by $X^{(1)}, \ldots, X^{(k)}$ .
    Compute empirical estimates $\hat{\mu}_u = \frac{1}{k}\sum_i X_u^{(i)}$ for each node $u \in V$ and $\hat{\mu}_{uv} =.$ $\frac{1}{k}\sum_i X_u^{(i)} X_v^{(i)}$ for each pair of nodes $(u, v)$
    Using the above estimates compute the covariance estimates $\hat{\lambda}_{uv} = \hat{\mu}_{uv} - \hat{\mu}_u \hat{\mu}_v$ for. each pair of nodes $(u, v)$
    If for any pair of nodes $(u, v)$, $\left|\hat{\lambda}_{uv}\right| \geq \frac{\varepsilon}{2n\beta d_{\mathrm{max}}}$ return that $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$ .
    Otherwise, return that $p \in \mathcal{I}_n$.
end function
```

$X_v \sim$ *Rademacher*$((1 + \mu_v)/2)$. We will use Lemma 1 to show that $O(\log n/\varepsilon^2)$ samples suffice to detect whether $\lambda_e = 0$ or $|\lambda_e| \geq \varepsilon$ with probability at least $1 - 1/10n^2$. With $O(\log n/\varepsilon^2)$ samples, Lemma 1 implies we can obtain estimates $\hat{\mu}_{uv}$, $\hat{\mu}_u$ and $\hat{\mu}_v$ for μ_{uv}, μ_u and μ_v respectively such that $|\hat{\mu}_{uv} - \mu_{uv}| \leq \frac{\varepsilon}{10}$, $|\hat{\mu}_u - \mu_u| \leq \frac{\varepsilon}{10}$ and $|\hat{\mu}_v - \mu_v| \leq \frac{\varepsilon}{10}$ with probability at least $1 - 1/10n^2$. Let $\hat{\lambda}_{uv} = \hat{\mu}_{uv} - \hat{\mu}_u \hat{\mu}_v$. Then from the above, it follows that $|\lambda_{uv} - \hat{\lambda}_{uv}| \leq \frac{3\varepsilon}{10} + \frac{\varepsilon^2}{100}$. It can be seen that in the case when the latter term in the previous inequality dominates the first, ε is large enough that $O(\log n)$ samples suffice to distinguish the two cases. In the more interesting case, $\frac{\varepsilon^2}{100} \leq \frac{\varepsilon}{10}$, and hence by the triangle inequality $|\lambda_{uv} - \hat{\lambda}_{uv}| \leq \frac{4\varepsilon}{10}$. Therefore if $|\lambda_{uv}| \geq \varepsilon$, then $\left|\hat{\lambda}_{uv}\right| \geq \frac{6\varepsilon}{10}$, and if $|\lambda_{uv}| = 0$, then $\left|\hat{\lambda}_{uv}\right| \leq \frac{4\varepsilon}{10}$ thereby implying that with probability at least $1 - 1/10n^2$ we can detect whether $\lambda_{uv} = 0$ or $|\lambda_{uv}| \geq \varepsilon$. Taking a union bound over all edges, the probability that we correctly identify all such edges is at least $9/10$. □

With this lemma in hand, we now prove the first part of Theorem 2.

Lemma 7. *Given $\tilde{O}\left(\frac{m^2\beta^2}{\varepsilon^2}\right)$ samples from an Ising model $X \sim p$, Algorithm 1 distinguishes between the cases $p \in \mathcal{I}_n$ and $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$ with probability at least* $2/3$.

Proof. We will run Algorithm 1 on all pairs X_u, X_v to identify any pair such that $|\lambda_{uv}|$ is large. If no such pair is identified, output that $p \in \mathcal{I}_n$, and otherwise, output that $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$. If $p \in \mathcal{I}_n$, we know that $\mathbf{E}[X_u X_v] = \mathbf{E}[X_u]\mathbf{E}[X_v]$ for all edges (u, v), and therefore, with probability $9/10$, there will be no edges for which the empirical estimate of $|\lambda_e| \geq \frac{\varepsilon}{2\beta m}$. On the other hand, if $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$, then $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$ for every $q \in \mathcal{I}_n$. In particular, consider the product distribution q on n nodes such that $\mu_u^q = \mu_u^p$ for all $u \in V$. For this particular product distribution q, by (2.2), there must exist some e^* such that $|\lambda_{e^*}| \geq \frac{\varepsilon}{2\beta m}$, and the algorithm will identify this edge. This is because

$$\sum_{v \in V} \left(\theta_v^p - \theta_v^q\right)\left(\mu_v^p - \mu_v^q\right) = 0 \tag{3.1}$$

$$\begin{aligned} \therefore\ & d_{\mathrm{SKL}}(p,q) \geq \varepsilon \\ & \implies \exists e^* = (u,v) \text{ s.t } \left(\theta_e^p - \theta_e^q\right)\left(\mu_e^p - \mu_e^q\right) \geq \frac{\varepsilon}{m} & (3.2) \\ & \implies \exists e^* = (u,v) \text{ s.t } |(\mu_e^p - \mu_e^q)| \geq \frac{\varepsilon}{2\beta m} & (3.3) \\ & \implies \exists e^* = (u,v) \text{ s.t } |\lambda_{e^*}| \geq \frac{\varepsilon}{2\beta m}. \end{aligned}$$

where (3.1) follows because $\mu_v^p = \mu_v^q$ for all $v \in V$, (3.2) follows from Lemma 5 and (3.3) follows because $|\theta_e^p - \theta_e^q| \leq 2\beta$. This completes the proof of the first part of Theorem 2. □

3.2 Identity Test using Localization

If one wishes to test for identity of p to an Ising model q, the quantities whose absolute values indicate that p is far from q are $\mu_{uv}^p - \mu_{uv}^q$ for all pairs u,v, and $\mu_u^p - \mu_u^q$ for all u, instead of λ_{uv}. Since μ_{uv}^q and μ_u^q are given as part of the description of q, we only have to identify whether $\mathbf{E}[X_u X_v] \geq c$ and $\mathbf{E}[X_u] \geq c$ for any constant $c \in [-1,1]$. A variant of Lemma 6 as stated in Lemma 8 achieves this goal. Algorithm 2 describes the localization based identity test. Its correctness proof will imply the second part of Theorem 2 and is similar in vein to that of Algorithm 1. It is omitted here.

Lemma 8. *Given $O\left(\frac{\log n}{\varepsilon^2}\right)$ samples from an Ising model p, there exists a polynomial-time algorithm which, with probability at least $9/10$, can identify all pairs of nodes $(u,v) \in V^2$ such that $|\mu_{uv}^p - c| \geq \varepsilon$ for any constant $c \in [-1,1]$. There exists a similar algorithm, with sample complexity $O\left(\frac{\log n}{\varepsilon^2}\right)$ which instead identifies all $v \in V$ such that $|\mu_v^p - c| \geq \varepsilon$, where $\mu_v^p = \mathbf{E}[X_v]$ for any constant $c \in [-1,1]$.*

Proof of Lemma 8: The proof follows along the same lines as Lemma 6. Let $X \sim p$. Then, for any pair of nodes (u,v), $X_u X_v \sim$ *Rademacher*$((1+\mu_e^p)/2)$. Also $X_u \sim$ *Rademacher*$((1+\mu_u^p)/2)$ for any node u. For any pair of nodes u,v, with $O(\log n/\varepsilon^2)$ samples, Lemma 1 implies we that the empirical estimate $\hat{\mu}_{uv}^p$ is such that $|\hat{\mu}_{uv}^p - \mu_{uv}^p| \leq \frac{\varepsilon}{10}$ with probability at least $1 - 1/10n^2$. By triangle inequality, we get $|\mu_{uv}^p - c| - \frac{\varepsilon}{10} \leq |\hat{\mu}_{uv}^p - c| \leq |\mu_{uv}^p - c| + \frac{\varepsilon}{10}$. Therefore if $|\mu_{uv}^p - c| = 0$, then $|\hat{\mu}_{uv}^p - c| \leq \frac{\varepsilon}{10}$ w.p. $\geq 1 - 1/10n^2$ and if $|\mu_{uv}^p - c| \geq \varepsilon$, then $|\hat{\mu}_{uv}^p - c| \geq \frac{9\varepsilon}{10}$ w.p. $\geq 1 - 1/10n^2$. Hence by comparing whether $|\hat{\mu}_{uv}^p - c|$ to $\varepsilon/2$ we can distinguish between the cases $|\mu_{uv}^p - c| = 0$ and $|\mu_{uv}^p - c| \geq \varepsilon$ w.p. $\geq 1 - 1/10n^2$. Taking a union bound over all edges, the probability that we correctly identify all such edges is at least 9/10. The second statement of the Lemma about the nodes follows similarly. □

Algorithm 2 Test if an Ising model p is identical to q

```
function LocalizationTestIdentity(sample access to Ising model $X \sim p$, description of Ising model $q$, accuracy parameter $\varepsilon$,$\beta$,$h$,$d_{\max}$)
    Draw $k = c\frac{\left(n^2 d_{\max}^2 \beta^2 + n^2 h^2\right) \log n}{\varepsilon^2}$ samples from $p$ for some constant $c$. Denote the. samples by $X^{(1)}, \ldots, X^{(k)}$
    Compute empirical estimates $\hat{\mu}_u^p = \frac{1}{k}\sum_i X_u^{(i)}$ for each node $u \in V$ and $\hat{\mu}_{uv}^p =. \frac{1}{k}\sum_i X_u^{(i)} X_v^{(i)}$ for each pair of nodes $(u, v)$
    If for any pair of nodes $(u, v)$, $|\hat{\mu}_{uv}^p - \mu_{uv}^q| \geq \frac{2\varepsilon}{n\beta d_{\max}}$ return that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$ .
    If for any node $u$, if $|\hat{\mu}_u^p - \mu_u^q| \geq \frac{2\varepsilon}{nhd_{\max}}$ return that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$ .
    Otherwise, return that $p = q$.
end function
```

The proof of correctness of Algorithm 2 follows along the same lines as that of Algorithm 1 and uses Lemma 8. We omit the proof here.

Chapter 4

Learn-then-Test Algorithm

In this chapter, we describe a framework for testing Ising models in the high temperature regime which results in algorithms which are more efficient than our baseline localization algorithm of Chapter 3 for dense graphs. This is the more technically involved part of the thesis and we modularize the description and analysis into different parts. We will give a high level overview of our approach here. Recall from Definition 2 that Ising models in the high temperature regime have a bound on the maximum allowed strength of edge interactions. To be precise, we have that $\beta \leq \frac{1}{4d_{\max}}$ where β is the maximum strength of the edge interactions.

The main approach we take in this chapter is to consider a global test statistic over all the variables on the Ising model in contrast to the localized statistics of Chapter 3. For ease of exposition, we first describe the approach for testing independence under no external field. We then describe the changes that need to be made to obtain tests for independence under an external field and goodness-of-fit in Section 4.5.

Note that testing independence under no external field boils down to testing uniformity as the only independent Ising model when there is no external field is the one corresponding to the uniform distribution. The intuition for the core of the algorithm is as follows. Suppose we are interested in testing uniformity of Ising model p with parameter vector $\vec{\theta}$. Note that for the uniform Ising model, $\theta_{uv} = \theta_u = 0$ for all $u, v \in V$. We start by obtaining an upper bound on the SKL between p and $\mathcal{U}_n$ which can be captured via a statistic that does not depend on $\vec{\theta}$. From (2.2), we have that under no external field ($\theta_u = 0$ for all $u \in V$),

$$\begin{aligned} d_{\mathrm{SKL}}(p, \mathcal{U}_n) &= \sum_{e=(u,v)\in E} \theta_{uv}\mu_{uv} \\ \implies d_{\mathrm{SKL}}(p, \mathcal{U}_n) &\leq \sum_{u\neq v} \beta \left|\mu_{uv}\right| \qquad (4.1) \\ \implies \frac{d_{\mathrm{SKL}}(p, \mathcal{U}_n)}{\beta} &\leq \sum_{u\neq v} \left|\mu_{uv}\right|. \qquad (4.2) \end{aligned}$$

where (4.1) holds because $|\theta_{uv}| \leq \beta$.

Given the above upper bound, we consider the statistic $Z = \sum_{u\neq v} \mathbf{sign}(\mu_{uv}) \cdot (X_u X_v)$, where $X \sim p$ and $\mathbf{sign}(\mu_{uv})$ is chosen arbitrarily if $\mu_{uv} = 0$.

$$\mathbf{E}[Z] = \sum_{u\neq v} \left|\mu_{uv}\right|.$$

If $X \in \mathcal{I}_n$, then $\mathbf{E}[Z] = 0$. On the other hand, by (4.2), we know that if $d_{\mathrm{SKL}}(X, \mathcal{I}_n) \geq \varepsilon$, then $\mathbf{E}[Z] \geq \varepsilon/\beta$. If the $\mathbf{sign}(\mu_e)$ parameters were known, we could simply plug them into Z, and using Chebyshev's inequality, distinguish these two cases using $\mathbf{Var}(Z)\beta^2/\varepsilon^2$ samples.

There are two main challenges here.

- First, the sign parameters, $\mathbf{sign}(\mu_{uv})$, are *not* known.
- Second, it is not obvious how to get a non-trivial bound for $\mathbf{Var}(Z)$.

One can quickly see that learning all the sign parameters might be prohibitively expensive. For example, if there is an edge e such that $|\mu_e| = 1/2^n$, there would be no hope of correctly estimating its sign with a polynomial number of samples. Instead, we perform a process we call *weak learning* – rather than trying to correctly estimate all the signs, we instead aim to obtain a $\vec{\Gamma}$ which is *correlated* with the vector $\mathbf{sign}(\mu_e)$. In particular, we aim to obtain $\vec{\Gamma}$ such that, in the case where $d_{\mathrm{SKL}}(p, \mathcal{U}_n) \geq \varepsilon$, $\mathbf{E}[\sum_{e=(u,v)\in E} \Gamma_e\, (X_u X_v)] \geq \varepsilon/\zeta\beta$, where $\zeta = \mathrm{poly}(n)$. That is we learn a sign vector $\vec{\Gamma}$ which is correlated enough with the true sign vector such that a sufficient portion of the signal from the d_{SKL} expression is still preserved. The main difficulty of analyzing this process is due to correlations between random variables $(X_u X_v)$. Naively, we could get an appropriate Γ_e for $(X_u X_v)$ by running a weak learning process independently for each edge. However, this incurs a prohibitive cost of $O(n^2)$ by iterating over all edges. We manage to sidestep this cost by showing that, despite these correlations, learning all Γ_e simultaneously succeeds with a probability which is $\geq 1/\operatorname{poly}(n)$, for a moderate polynomial in n. Thus, repeating this process several times, we can obtain a $\vec{\Gamma}$ which has the appropriate guarantee with sufficient constant probability.

At this point, we are in the setting as described above – we have a statistic Z' of the form:

$$Z' = \sum_{u \neq v} c_{uv} X_u X_v \tag{4.3}$$

where $c \in \{\pm 1\}^{\binom{V}{2}}$ represent the signs obtained from the weak learning procedure. $\mathbf{E}[Z'] = 0$ if $X \in \mathcal{I}_n$, and $\mathbf{E}[Z'] \geq \varepsilon/\zeta\beta$ if $d_{\mathrm{SKL}}(X, \mathcal{I}_n) \geq \varepsilon$. These two cases can be distinguished using $\mathbf{Var}(Z')\zeta^2\beta^2/\varepsilon^2$ samples, by Chebyshev's inequality. At this point, we run into the second issue mentioned above. Since the range of Z' is $\Omega(n^2)$, a crude bound for $\mathbf{Var}(Z')$ is $O(n^4)$, granting us no savings over the localization algorithm of Theorem 2. However, in the high temperature regime, we show the following bound on the variance of Z' (Theorem 9).

$$\mathbf{Var}(Z') = \tilde{O}(n^2) + \tilde{O}\left(n^3\beta^3 d_{\max}^{1.5}\right).$$

Surprisingly, for dense graphs in our high temperature regime, the above bound implies that $\mathbf{Var}(Z') = \tilde{O}(n^2)$. In other words, despite the potentially complex structure of the Ising model and potential correlations, the variables $X_u X_v$ contribute to the variance of Z' roughly as if they were all independent! We believe the result and techniques involved in the analysis of this variance bound are of independent interest outside the context of this algorithm, and describe them in Chapter 5. Given the tighter bound on the variance of our statistic, we run the Chebyshev-based test on all the hypotheses obtained in the previous learning step (with appropriate failure probability) to conclude our algorithm. Further details about the algorithm are provided in Sections 4.1-4.4.

We state the sample complexity achieved via our learn-then-test framework for independence testing under no external field here. The corresponding statements for independence testing under external fields and identity testing are given in Section 4.5.

Theorem 3 (Independence Testing using Learn-Then-Test, No External Field)**.** *Suppose p is an Ising model in the high temperature regime under no external field. Then, given $\tilde{O}\left(\max\left\{\frac{n^{10/3}\beta^2}{\varepsilon^2}, \frac{n^{11/3}\cdot\beta^3\cdot\sqrt{d_{\max}}}{\varepsilon^2}\right\}\right)$ i.i.d samples from p, the learn-then-test algorithm runs in polynomial time and distinguishes between the cases $p \in \mathcal{I}_n$ and $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$ with probability at least $9/10$.*

Next, we state a corollary of Theorem 3 with sample complexities we obtain when β is close to the high temperature threshold.

Theorem 4 (Independence Testing with β near the Threshold of High Temperature, No External Field)**.** *Suppose that p is an Ising model in the high temperature regime and suppose that $\beta = \frac{1}{4d_{\max}}$. That is, β is close to the high temperature threshold. Then:*

- *Given $\tilde{O}\left(\max\left\{\frac{n^{10/3}}{\varepsilon^2 d_{\max}^2}, \frac{n^{11/3}}{\varepsilon^2 d_{\max}^{2.5}}\right\}\right)$ i.i.d samples from p **with no external field**, the learn-then-test algorithm runs in polynomial time and distinguishes between the cases $p \in \mathcal{I}_n$ and $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$ with probability at least $2/3$. For testing identity of p to an Ising model q in the high temperature regime, we obtain the same sample complexity as above.*

Figure 6-1 shows the dependence of sample complexity of testing as $d_{\max}$ is varied in the regime of Theorem 4 for the case of no external field.

The description of our algorithm is presented in Algorithm 3. It contains a parameter τ, which we choose to be the value achieving the minimum in the sample complexity of Theorem 5. The algorithm follows a learn-then-test framework, which we outline here.

Algorithm 3 Test if an Ising model p under no external field is product using Learn-Then-Test

```
function LEARN-THEN-TEST-ISING(sample access to an Ising model $p, \beta, d_{\max}, \varepsilon, \tau$)
    Run the localization Algorithm 1 on $p$ with accuracy parameter $\frac{\varepsilon}{n^\tau}$. If it identifies. any edges, return that $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$
    for $\ell = 1$ to $O(n^{2-\tau})$ do
        Run the weak learning Algorithm 4 on $S = \{X_u X_v\}_{u \neq v}$ with parameters $\tau$ and $\varepsilon/\beta$. to generate a sign vector $\vec{\Gamma}^{(\ell)}$ where $\Gamma_{uv}^{(\ell)}$ is weakly correlated with $\mathbf{sign}\left(\mathbf{E}\left[X_{uv}\right]\right)$
    end for
    Using the same set of samples for all $\ell$, run the testing algorithm of Lemma 11. on each of the $\vec{\Gamma}^{(\ell)}$ with parameters $\tau_2 = \tau, \delta = O(1/n^{2-\tau})$. If any output that $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$, return that $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$. Otherwise, return that $p \in \mathcal{I}_n$
end function
```

Note: The first step in Algorithm 3 is to perform a localization test to check if $|\mu_e|$ is not too far away from 0 for all e. It is added to help simplify the analysis of the algorithm and is not necessary in principle. In particular, we use the first part of Algorithm 1, which checks if any edge looks far from uniform, to perform this first step, albeit with a smaller value of the accuracy parameter ε than before. Similar to before, if we find a single non-uniform edge, this is sufficient evidence to output $d_{\mathrm{SKL}}(X, \mathcal{I}_n) \geq \varepsilon$. If we do not find any edges which are verifiably far from uniform, we proceed onward, with the additional guarantee that $|\mu_e|$ is small for all $e \in E$.

A statement of the exact sample complexity achieved by our algorithm is given in Theorem 5. When optimized for the parameter τ, this yields Theorem 3.

Theorem 5. *Given $\tilde{O}\left(\min_{\tau>0}\left(n^{2+\tau}+n^{4-2\tau}\cdot\min\left\{n^3,\max\left\{n^2,n^3\cdot d_{\max}^{1.5}\cdot\beta^3\right\}\right\}\right)\frac{\beta^2}{\varepsilon^2}\right)$ i.i.d samples from an Ising model p in the high-temperature regime with no external field, there exists a polynomial-time algorithm which distinguishes between the cases $p\in\mathcal{I}_n$ and $d_{\mathrm{SKL}}(p,\mathcal{I}_n)\geq\varepsilon$ with probability at least $2/3$.*

The organization of the rest of the chapter is as follows. We describe and analyze our weak learning procedure in Section 4.1. Given a vector with the appropriate weak learning guarantees, we describe and analyze the testing procedure in Section 4.2. In Section 4.3, we describe how to combine all these ideas – in particular, our various steps have several parameters, and we describe how to balance the complexities to obtain the sample complexity stated in Theorem 5. Finally, in Section 4.4, we optimize the sample complexities from Theorem 5 for the parameter τ and filter out cleaner statement of Theorem 3. We compare the performance of our localization and learn-then-test algorithms and describe the best sample complexity achieved in different regimes in Section 6.

4.1 Weak Learning

Our overall goal of this section is "weakly learn" the sign of $\mu_e=\mathbf{E}[X_uX_v]$ for all edges $e=(u,v)$. More specifically, we wish to output a vector $\vec{\Gamma}$ with the following guarantee:

$$\mathbf{E}_X\left[\sum_{e=(u,v)\in E}\Gamma_e X_u X_v\right]\geq\frac{c\varepsilon}{2\beta n^{2-\tau_2}},$$

for some constant $c>0$ and parameter τ_2 to be specified later. Note that the "best" Γ, for which $\Gamma_e=\mathbf{sign}(\mu_e)$, has this guarantee with $\tau_2=2$ – by relaxing our required learning guarantee, we can reduce the sample complexity in this stage.

The first step will be to prove a simple but crucial lemma answering the following question: Given k samples from a Rademacher random variable with parameter p, how well can we estimate the sign of its expectation? This type of problem is well studied in the regime where $k=\Omega(1/p^2)$, in which we have a constant probability of success (see, i.e. Lemma 1), but we analyze the case when $k\ll 1/p^2$ and prove how much better one can do versus randomly guessing the sign. See Lemma 22 in Section A for more details.

With this lemma in hand, we proceed to describe the weak learning procedure. Given parameters τ,ε and sample access to a set S of 'Rademacher-like' random variables which may be *arbitrarily correlated* with each other, the algorithm draws $\tilde{O}\left(\frac{n^{2\tau}}{\varepsilon^2}\right)$ samples from each random variable in the set and computes their empirical expected values and outputs a signs of thus obtained empirical expectations. The procedure is described in Algorithm 4.

Algorithm 4 Weakly Learn Signs of the Expectations of a set of Rademacher-like random variables

```
function WeakLearning(sample access to set $S = \{Z_i\}_i$ of random variables where $|S| = O(n^s)$ and where $Z_i \in \{-1, 0, +1\}$ and can be arbitrarily correlated,$\varepsilon$, $\tau$,).
    Draw $k = \tilde{O}\left(\frac{n^{2\tau}}{\varepsilon^2}\right)$ samples from each $Z_i$. Denote the samples by $Z_i^{(1)}, \ldots, Z_i^{(k)}$ .
    Compute the empirical expectation for each $Z_i$: $\hat{Z}_i = \frac{1}{k}\sum_{l=1}^{k} Z_i^{(l)}$.
    Output $\vec{\Gamma}$ where $\Gamma_i = \mathbf{sign}(\hat{Z}_i)$.
end function
```

We now turn to the setting of the Ising model, discussed in Section 4.1. We invoke the weak-learning procedure of Algorithm 4 on the set $S = \{X_u X_v\}_{u \neq v}$ with parameters ε/β and $0 \leq \tau \leq 2$. By linearity of expectations and Cauchy-Schwarz, it is not hard to see that we can get a guarantee of the form we want in expectation (see Lemma 9). However, the challenge remains to obtain this guarantee with constant probability. Carefully analyzing the range of the random variable and using this guarantee on the expectation allows us to output an appropriate vector $\vec{\Gamma}$ with probability inversely polynomial in n (see Lemma 10). Repeating this process several times will allow us to generate a collection of candidates $\{\vec{\Gamma}^{(\ell)}\}$, at least one of which has our desired guarantees with constant probability.

Weak Learning the Edges of an Ising Model

We now turn our attention to weakly learning the edge correlations in the Ising model. To recall, our overall goal is to obtain a vector $\vec{\Gamma}$ such that

$$\mathbf{E}_X\left[\sum_{e=(u,v)\in E} \Gamma_e X_u X_v\right] \geq \frac{c\varepsilon}{2\beta n^{2-\tau_2}}.$$

We start by proving that such a bound holds in expectation. The following is fairly straightforward from Lemma 22 and linearity of expectations.

Lemma 9. *Given $k = O\left(\frac{n^{2\tau_2}\beta^2}{\varepsilon^2}\right)$ samples from an Ising model X such that $d_{\mathrm{SKL}}(X, \mathcal{I}_n) \geq \varepsilon$ and $|\mu_e| \leq \frac{\varepsilon}{\beta n^{\tau_2}}$ for all $e \in E$, there exists an algorithm which outputs $\vec{\Gamma} = \{\Gamma_e\} \in \{\pm 1\}^{|E|}$*

such that

$$\mathbf{E}_{\vec{\Gamma}}\left[\mathbf{E}_X\left[\sum_{e=(u,v)\in E}\Gamma_e X_u X_v\right]\right] \geq \frac{c\beta}{\varepsilon n^{2-\tau_2}}\left(\sum_{e\in E}|\mu_e|\right)^2,$$

for some constant $c > 0$.

Proof. Since for all $e = (u, v) \in E$, $|\mu_e| \leq \frac{\varepsilon}{\beta n^{\tau_2}}$, and by our upper bound on k, all of the random variables $X_u X_v$ fall into the first case of Lemma 22 (the "small k" regime). Hence, we get that

$$\Pr\left[\Gamma_e = \mathbf{sign}(\mu_e)\right] \geq \frac{1}{2} + \frac{c_1|\mu_e|\sqrt{k}}{2}$$

which implies that

$$\begin{aligned}\mathbf{E}_{\Gamma_e}\left[\Gamma_e\mu_e\right] &\geq \left(\frac{1}{2} + \frac{c_1|\mu_e|\sqrt{k}}{2}\right)|\mu_e| + \left(\frac{1}{2} - \frac{c_1|\mu_e|\sqrt{k}}{2}\right)(-|\mu_e|) \\ &= c_1|\mu_e|^2\sqrt{k}\end{aligned}$$

Summing up the above bound over all edges, we get

$$\begin{aligned}\mathbf{E}_{\vec{\Gamma}}\left[\sum_{e\in E}\Gamma_e\mu_e\right] &\geq c_1\sqrt{k}\sum_{e\in E}|\mu_e|^2 \\ &\geq \frac{c_1' n^{\tau_2}\beta}{\varepsilon}\sum_{e\in E}|\mu_e|^2,\end{aligned}$$

for some constant $c_1' > 0$. Applying the Cauchy-Schwarz inequality gives us

$$\mathbf{E}_{\vec{\Gamma}}\left[\sum_{e\in E}\Gamma_e\mu_e\right] \geq \frac{c\beta}{\varepsilon n^{2-\tau_2}}\left(\sum_{e\in E}|\mu_e|\right)^2,$$

as desired. □

Next, we prove that the desired bound holds with sufficiently high probability. The following lemma follows by a careful analysis of the extreme points of the random variable's range.

Lemma 10. *Given $k = O\left(\frac{n^{2\tau_2}\beta^2}{\varepsilon^2}\right)$ samples from an Ising model X such that $d_{\mathrm{SKL}}(X, \mathcal{I}_n) \geq \varepsilon$ and $|\mu_e| \leq \frac{\varepsilon}{\beta n^{\tau_2}}$ for all $e \in E$, there exists an algorithm which outputs $\vec{\Gamma} = \{\Gamma_e\} \in \{\pm 1\}^{|E|}$.*

Define χ_{τ_2} to be the event that

$$\mathbf{E}_X\left[\sum_{e=(u,v)\in E}\Gamma_e X_u X_v\right] \geq \frac{c\varepsilon}{2\beta n^{2-\tau_2}},$$

for some constant $c > 0$. We have that

$$\mathbf{Pr}_\Gamma\left[\chi_{\tau_2}\right] \geq \frac{c}{4n^{2-\tau_2}}.$$

Proof. We introduce some notation which will help in the elucidation of the argument which follows. Let $p = \mathbf{Pr}_\Gamma\left[\chi_{\tau_2}\right]$. Let

$$T = \frac{c\beta}{2\varepsilon n^{2-\tau_2}}\left(\sum_{e\in E}|\mu_e|\right)^2.$$

Let Y be the random variable defined as follows

$$Y = \mathbf{E}_X\left[\sum_{e=(u,v)\in E}\Gamma_e X_u X_v\right],$$

$$U = \mathbf{E}_{\vec{\Gamma}}\left[Y|Y > T\right] \text{ and}$$

$$L = \mathbf{E}_{\vec{\Gamma}}\left[Y|Y \leq T\right]$$

Then we have

$$\begin{aligned} pU + (1-p)L &= 2T \text{ (From Lemma 9)} \\ \implies p &= \frac{2T - L}{U - L} \end{aligned}$$

Since $U \leq \sum_{e\in E}|\mu_e|$, we have

$$p \geq \frac{2T - L}{\left(\sum_{e\in E}|\mu_e|\right) - L}$$

Since $L \geq -\sum_{e\in E}|\mu_e|$,

$$p \geq \frac{2T - L}{2\left(\sum_{e\in E}|\mu_e|\right)}$$

Since $L \leq T$, we get

$$p \geq \frac{T}{2\left(\sum_{e\in E}|\mu_e|\right)}$$

Substituting in the value for T we get

$$p \geq \frac{c\beta \left(\sum_{e \in E} |\mu_e|\right)^2}{4\varepsilon n^{2-\tau_2} \left(\sum_{e \in E} |\mu_e|\right)}$$

$$\implies p \geq \frac{c\beta \left(\sum_{e \in E} |\mu_e|\right)}{4\varepsilon n^{2-\tau_2}}$$

Since $d_{\mathrm{SKL}}(X, \mathcal{I}_n) \geq \varepsilon$, this implies $\left(\sum_{e \in E} |\mu_e|\right) \geq \varepsilon/\beta$ and thus

$$p \geq \frac{c}{4n^{2-\tau_2}},$$

as desired. □

4.2 Testing Our Learned Hypothesis

In this section, we assume that we were successful in weakly learning a vector $\vec{\Gamma}$ which is "good" (i.e., it satisfies χ_{τ_2}). With such a $\vec{\Gamma}$, we show that we can distinguish between $X \in \mathcal{I}_n$ and $d_{\mathrm{SKL}}(X, \mathcal{I}_n) \geq \varepsilon$.

Lemma 11. *Let X be an Ising model, and let σ^2 be such that, for any $\vec{\gamma} = \{\gamma_e\} \in \{\pm 1\}^{|E|}$,*

$$\mathbf{Var}\left(\sum_{e=(u,v) \in E} \gamma_e X_u X_v\right) \leq \sigma^2.$$

Given $k = O\left(\sigma^2 \cdot \frac{n^{4-2\tau_2}\beta^2 \log(1/\delta)}{\varepsilon^2}\right)$ samples from X, which satisfies either $X \in \mathcal{I}_n$ or $d_{\mathrm{SKL}}(X, \mathcal{I}_n) \geq \varepsilon$, and $\vec{\Gamma} = \{\Gamma_e\} \in \{\pm 1\}^{|E|}$ which satisfies χ_{τ_2} (as defined in Lemma 10) in the case that $d_{\mathrm{SKL}}(X, \mathcal{I}_n) \geq \varepsilon$, then there exists an algorithm which distinguishes these two cases with probability $\geq 1 - \delta$.

Proof. We prove this lemma with failure probability $1/3$ – by standard boosting arguments, this can be lowered to δ by repeating the test $O(\log(1/\delta))$ times and taking the majority result.

Denote the ith sample as $X^{(i)}$. The algorithm will compute the statistic

$$Z = \frac{1}{k}\left(\sum_{i=1}^{k} \sum_{e=(u,v) \in E} \Gamma_e X_u^{(i)} X_v^{(i)}\right).$$

If $Z \leq \frac{c\varepsilon}{4\beta n^{2-\tau_2}}$, then the algorithm will output that $X \in \mathcal{I}_n$. Otherwise, it will output that $d_{\mathrm{SKL}}(X, \mathcal{I}_n) \geq \varepsilon$.

By our assumptions in the lemma statement, in either case,

$$\mathbf{Var}\,(Z) \leq \frac{\sigma^2}{k}.$$

If $X = \mathcal{I}_n$, then we have that

$$\mathbf{E}[Z] = 0.$$

By Chebyshev's inequality, this implies that

$$\Pr\left[Z \geq \frac{\varepsilon}{4\beta n^{2-\tau_2}}\right] \leq \frac{16\sigma^2\beta^2 n^{4-2\tau_2}}{kc^2\varepsilon^2}.$$

Substituting the value of k gives the desired bound in this case. The case where $d_{\mathrm{SKL}}(X, \mathcal{I}_n) \geq \varepsilon$ follows similarly, but additionally using the fact that χ_{τ_2} implies that

$$\mathbf{E}[Z] \geq \frac{c\varepsilon}{2\beta n^{2-\tau_2}}.$$

□

4.3 Combining Learning and Testing

In this section, we combine lemmas from the previous sections to complete the proof of Theorem 5. Lemma 10 gives us that a single iteration of the weak learning step gives a "good" $\vec{\Gamma}$ with probability at least $\Omega\left(\frac{1}{n^{2-\tau_2}}\right)$. We repeat this step $O(n^{2-\tau_2})$ times, generating $O(n^{2-\tau_2})$ hypotheses $\vec{\Gamma}^{(\ell)}$. By standard tail bounds on geometric random variables, this will imply that at least one hypothesis is good (i.e. satisfying χ_{τ_2}) with probability at least 9/10. We then run the algorithm of Lemma 11 on each of these hypotheses, with failure probability $\delta = O(1/n^{2-\tau_2})$. If $p \in \mathcal{I}_n$, all the tests will output that $p \in \mathcal{I}_n$ with probability at least 9/10. Similarly, if $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$, conditioned on at least one hypothesis $\vec{\Gamma}^{(\ell^*)}$ being good, the test will output that $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$ for this hypothesis with probability at least 9/10. This proves correctness of our algorithm.

To conclude our proof, we analyze its sample complexity. Combining the complexities of Lemmas 6, 10, and 11, the overall sample complexity is

$$O\left(\frac{n^{2\tau_1}\beta^2 \log n}{\varepsilon^2}\right) + O\left(\frac{n^{2+\tau_2}\beta^2}{\varepsilon^2}\right) + O\left(\sigma^2 \frac{n^{4-2\tau_2}\beta^2}{\varepsilon^2} \log n\right).$$

Noting that the first term is always dominated by the second term we can simplify the complexity to the following expression.

$$O\left(\frac{n^{2+\tau_2}\beta^2}{\varepsilon^2}\right)+O\left(\sigma^2\frac{n^{4-2\tau_2}\beta^2}{\varepsilon^2}\log n\right). \tag{4.4}$$

Plugging in the variance bounds from Section 5, Theorems 9 and 10 gives Theorem 5.

4.4 Balancing Weak Learning and Testing

s

The sample complexities in the statement of Theorem 5 arise from a combination of two separate algorithms and from a variance bound for our multi-linear statistic which depends on β and $d_{\max}$. To balance for the optimal value of τ in various regimes of β and $d_{\max}$ we use Claim 1 which can be easily verified and arrive at Lemma 12.

Claim 1. *Let $S = \tilde{O}\left(\left(n^{2+\tau}+n^{4-2\tau}\cdot\sigma^2\right)\frac{\beta^2}{\varepsilon^2}\right)$. Let $\sigma^2 = O(n^s)$. The value of τ which minimizes S is $\frac{2+s}{3}$.*

Lemma 12. *Suppose p is an Ising model in the high temperature regime and under no external field. Then, given S i.i.d samples from p, the learn-then-test algorithm runs in polynomial time and distinguishes between the cases $p \in \mathcal{I}_n$ and $d_{\mathrm{SKL}}(p,\mathcal{I}_n) \geq \varepsilon$ with probability $\geq 9/10$ where*

- *$S = \tilde{O}\left(\frac{n^{11/3}\cdot\beta^3\cdot\sqrt{d_{\max}}}{\varepsilon^2}\right)$ if $\beta\sqrt{d_{\max}} = \Omega(n^{-1/3})$, and*
- *$S = \tilde{O}\left(n^{10/3}\frac{\beta^2}{\varepsilon^2}\right)$ if $\beta\sqrt{d_{\max}} = o(n^{-1/3})$.*

Lemma 12 can be condensed to give Theorem 3.

4.5 Changes Required for General Independence and Identity Testing

We describe the modifications that need to be done to the learn-then-test approach described in Sections 4.1-4.4 to obtain testers for independence under an arbitrary external field (Section 4.5), identity without an external field (Section 4.5), and identity under an external field (Section 4.5).

Independence Testing under an External Field

Under an external field, the statistic we considered in Section 4 needs to be modified. Suppose we are interested in testing independence of an Ising model p defined on a graph $G = (V, E)$ with a parameter vector $\vec{\theta}^p$. Let $X \sim p$. We have that $d_{\mathrm{SKL}}(p, \mathcal{I}_n) = \min_{q \in \mathcal{I}_n} d_{\mathrm{SKL}}(p, q)$. In particular, we consider q to be the independent Ising model on graph $G' = (V, E')$ with parameter vector $\vec{\theta}^q$ such that $E' = \phi$ and θ^q_u is such that $\mu^q_u = \mu^p_u$ for all $u \in V$. Then,

$$\begin{aligned} d_{\mathrm{SKL}}(p, \mathcal{I}_n) &\leq d_{\mathrm{SKL}}(p, q) \qquad (4.5) \\ &= \sum_{e=(u,v) \in E} \theta^p_{uv} \left(\mu^p_{uv} - \mu^q_{uv}\right) \\ &= \sum_{e=(u,v) \in E} \theta^p_{uv} \left(\mu^p_{uv} - \mu^p_u \mu^p_v\right) \\ &\leq \sum_{e=(u,v) \in E} \beta \left|\mu^p_{uv} - \mu^p_u \mu^p_v\right| \\ \implies \frac{d_{\mathrm{SKL}}(p, \mathcal{I}_n)}{\beta} &\leq \sum_{e=(u,v) \in E} \left|\mu^p_{uv} - \mu^p_u \mu^p_v\right|. \end{aligned}$$

The above inequality suggests a statistic Z such that $\mathbf{E}[Z] = \sum_{e=(u,v) \in E} |\lambda^p_{uv}|$ where $\lambda^p_{uv} = \mu^p_{uv} - \mu^p_u \mu^p_v$. We consider $Z = \sum_{u \neq v} \mathbf{sign}(\lambda_{uv}) \left(X^{(1)}_u - X^{(2)}_u\right) \left(X^{(1)}_v - X^{(2)}_v\right)$ where $X^{(1)}, X^{(2)} \sim p$ are two independent samples from p. It can be seen that Z has the desired expectation. However, we have the same issue as before that we don't know the $\mathbf{sign}(\lambda_{uv})$ parameters. Luckily, it turns out that our weak learning procedure is general enough to handle this case as well. Consider the following random variable: $Z_{uv} = \frac{1}{4} \left(X^{(1)}_u - X^{(2)}_u\right) \left(X^{(1)}_v - X^{(2)}_v\right)$. Z_{uv} takes on values in $\{-1, 0, +1\}$. Consider an associated Rademacher variable Z'_{uv} defined as follows: $\Pr[Z_{uv} = -1] = \Pr[Z_{uv} = -1] + 1/2 \Pr[Z_{uv} = 0]$. It is easy to simulate a sample from Z'_{uv} given access to a sample from Z_{uv}. If $Z_{uv} = 0$, toss a fair coin to decide whether $Z'_{uv} = -1$ or $+1$. $\mathbf{E}[Z'_{uv}] = \mathbf{E}[Z_{uv}] = \frac{\lambda_{uv}}{2}$. Hence $Z'_{uv} \sim \textit{Rademacher}\left(\frac{1}{2} + \frac{\lambda_{uv}}{4}\right)$ and by Lemma 22 with k copies of the random variable Z_{uv} we get a success probability of $1/2 + c_1\sqrt{k}\,|\lambda_{uv}|$ of estimating $\mathbf{sign}(\lambda_{uv})$ correctly. Given this guarantee, the rest of the weak learning argument of Lemmas 9 and 10 follows analogously by replacing μ_e with λ_e.

After we have *weakly learnt* the signs, we are left with a statistic Z'_{cen} of the form:

$$Z'_{cen} = \sum_{u \neq v} c_{uv} \left(X_u^{(1)} - X_u^{(2)} \right) \left(X_v^{(1)} - X_v^{(2)} \right) \tag{4.6}$$

where the subscript *cen* denotes that the statistic is a centered one and $c \in \{\pm 1\}^{\binom{V}{2}}$. We need to obtain a bound on $\mathbf{Var}(Z'_{cen})$. We again employ the technique of exchangeable pairs described in Section 5 to obtain a non-trivial bound on $\mathbf{Var}(Z'_{cen})$ in the high-temperature regime. The statement of the variance result is given in Theorem 10 and the details are in Section 5.3. Combining the weak learning part and the variance bound gives us the following sample complexity for independence testing under an external field:

$$\begin{aligned} &\tilde{O}\left(\frac{(n^{2+\tau} + n^{4-2\tau}\sigma^2)\beta^2}{\varepsilon^2} \right) \\ =&\tilde{O}\left(\frac{(n^{2+\tau} + n^{4-2\tau}\max\{n^2, n^3 \cdot \beta^2 \cdot d_{\max}\})\beta^2}{\varepsilon^2} \right) \end{aligned}$$

Balancing for the optimal value of the τ parameter gives Theorem 6.

Theorem 6 (Independence Testing using Learn-Then-Test, Arbitrary External Field)**.** *Suppose p is an Ising model in the high temperature regime under an arbitrary external field. The learn-then-test algorithm takes in $\tilde{O}\left(\frac{n^{2/3}\max\{n^{2/3}, n\beta^{2/3}d_{\max}^{1/3}\}\beta^2}{\varepsilon^2} \right)$ i.i.d. samples from p and distinguishes between the cases $p \in \mathcal{I}_n$ and $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$ with probability $\geq 9/10$.*

The tester is formally described in Algorithm 5.

Algorithm 5 Test if an Ising model p under arbitrary external field is product

```
function LEARN-THEN-TEST-ISING(sample access to an Ising model $p, \beta, d_{\max}, \varepsilon, \tau$)
    Run the localization Algorithm 1 with accuracy parameter $\frac{\varepsilon}{n^\tau}$. If it identifies any.
    edges, return that $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$
    for $\ell = 1$ to $O(n^{2-\tau})$ do
        Run the weak learning Algorithm 4 on $S = \{(X_u^{(1)} - X_u^{(2)})(X_v^{(1)} - X_v^{(2)})\}_{u \neq v}$ with.
        parameters $\tau_2 = \tau$ and $\varepsilon/\beta$ to generate a sign vector $\vec{\Gamma}^{(\ell)}$ where $\Gamma_{uv}^{(\ell)}$ is weakly
        correlated with $\mathbf{sign}\left(\mathbf{E}\left[(X_u^{(1)} - X_u^{(2)})(X_v^{(1)} - X_v^{(2)})\right]\right)$
    end for
    Using the same set of samples for all $\ell$, run the testing algorithm of Lemma 11.
    on each of the $\vec{\Gamma}^{(\ell)}$ with parameters $\tau_2 = \tau, \delta = O(1/n^{2-\tau})$. If any output that
    $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$, return that $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$. Otherwise, return that $p \in \mathcal{I}_n$
end function
```

Identity Testing under No External Field

We first look at the changes needed for identity testing under no external field. Similar to before, we start by obtaining an upper bound on the SKL between the Ising models p and q. We get that,

$$\begin{aligned} d_{\mathrm{SKL}}(p,q) &= \sum_{(u,v)\in E} \left(\theta^p_{uv} - \theta^q_{uv}\right)\left(\mu^p_{uv} - \mu^q_{uv}\right) \\ \implies \frac{d_{\mathrm{SKL}}(p,q)}{2\beta} &\le \sum_{u\neq v} \left|\left(\mu^p_{uv} - \mu^q_{uv}\right)\right| \end{aligned}$$

Since we know μ^q_{uv} for all pairs u, v, the above upper bound suggests the statistic Z of the form

$$Z = \sum_{u\neq v} \mathbf{sign}\left(\mu^p_{uv} - \mu^q_{uv}\right)\left(X_u X_v - \mu^q_{uv}\right)$$

If $p = q$, $\mathbf{E}[Z] = 0$ and if $d_{\mathrm{SKL}}(p,q) \ge \varepsilon$, $\mathbf{E}[Z] \ge \varepsilon/2\beta$. As before, there are two things we need to do: learn a sign vector which is weakly correlated with the right sign vector and obtain a bound on $\mathbf{Var}(Z)$. By separating out the part of the statistic which is just a constant, we obtain that

$$\mathbf{Var}(Z) \le \mathbf{Var}\left(\sum_{u\neq v} c_{uv} X_u X_v\right)$$

where $c \in \{\pm 1\}^{\binom{V}{2}}$. Hence, the variance bound of Theorem 9 holds for $\mathbf{Var}(Z)$.

As for the weakly learning the signs, using Corollary 1 of Lemma 22 we get that for each pair u, v, with k samples, we can achieve a success probability $1/2 + c_1\sqrt{k}\left|\mu^p_{uv} - \mu^q_{uv}\right|$ of correctly estimating $\mathbf{sign}(\mu^p_{uv} - \mu^q_{uv})$. Following this up with analogous proofs of Lemmas 9 and 10 where μ_e is replaced by $\mu^p_e - \mu^q_e$, we achieve our goal of weakly learning the signs with a sufficient success probability.

By making these changes we arrive at the following theorem for testing identity to an Ising model under no external field.

Theorem 7 (Identity Testing using Learn-Then-Test, No External Field)**.** *Suppose p and q are Ising models in the high temperature regime under no external field. The learn-then-test algorithm takes in $\tilde{O}\left(\frac{n^{2/3}\max\{n^{2/3}, n\beta d_{\max}^{0.5}\}\beta^2}{\varepsilon^2}\right)$ i.i.d. samples from p and distinguishes* *between the cases $p = q$ and $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$ with probability $\geq 9/10$.*

The tester is formally described in Algorithm 6.

Algorithm 6 Test if an Ising model p under no external field is identical to q

```
function TestIsing(sample access to an Ising model $p, \beta, d_{\max}, \varepsilon, \tau$, description of Ising model $q$ under no external field)
    Run the localization Algorithm 2 with accuracy parameter $\frac{\varepsilon}{n^{\tau}}$. If it identifies any. edges, return that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$
    for $\ell = 1$ to $O(n^{2-\tau})$ do
        Run the weak learning Algorithm 4 on $S = \{X_u X_v - \mu^q_{uv}\}_{u \neq v}$ with parameters. $\tau_2 = \tau$ and $\varepsilon/\beta$ to generate a sign vector $\vec{\Gamma}^{(\ell)}$ where $\Gamma^{(\ell)}_{uv}$ is weakly correlated with $\mathbf{sign}\left(\mathbf{E}\left[X_{uv} - \mu^q_{uv}\right]\right)$
    end for
    Using the same set of samples for all $\ell$, run the testing algorithm of Lemma 11. on each of the $\vec{\Gamma}^{(\ell)}$ with parameters $\tau_2 = \tau, \delta = O(1/n^{2-\tau})$. If any output that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$, return that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$. Otherwise, return that $p = q$
end function
```

Identity Testing under an External Field

When an external field is present, two things change. Firstly, the terms corresponding to nodes of the Ising model in the SKL expression no longer vanish and have to be accounted for. Secondly, the statistic we use is not appropriately centered and can have a variance of $O(n^3)$. This worsens the sample complexity slightly. We will describe the first change in more detail now. Again, we start by considering an upper bound on the SKL between Ising models p and q.

$$d_{\mathrm{SKL}}(p, q) = \sum_{v \in V} \left(\theta^p_v - \theta^q_v\right)\left(\mu^p_v - \mu^q_v\right) + \sum_{(u,v) \in E} \left(\theta^p_{uv} - \theta^q_{uv}\right)\left(\mu^p_{uv} - \mu^q_{uv}\right)$$

$$\implies d_{\mathrm{SKL}}(p, q) \leq 2h \sum_{v \in V} \left|\left(\mu^p_v - \mu^q_v\right)\right| + 2\beta \sum_{u \neq v} \left|\left(\mu^p_{uv} - \mu^q_{uv}\right)\right|$$

Hence if $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$, then either

- $2h \sum_{v \in V} |(\mu^p_v - \mu^q_v)| \geq \varepsilon/2$ or
- $2\beta \sum_{u \neq v} |(\mu^p_{uv} - \mu^q_{uv})| \geq \varepsilon/2$.

Moreover, if $p = q$, then both $2h \sum_{v \in V} |(\mu^p_v - \mu^q_v)| = 0$ and $2\beta \sum_{u \neq v} |(\mu^p_{uv} - \mu^q_{uv})| = 0$. Our tester will first test for case (i) and if that test doesn't declare that the two Ising models are far, then proceeds to test whether case (ii) holds.

We will first describe the test to detect whether $\sum_{v\in V} |(\mu_v^p - \mu_v^q)| = 0$ or is $\geq \varepsilon/2h$. We observe that the random variables X_v are Rademachers and hence we can use the weak-learning framework we developed so far to accomplish this goal. The statistic we consider is $Z = \sum_{v\in V} \mathbf{sign}(\mu_v^p)\,(X_v - \mu_v^q)$. Again, as before, we face two challenges: we don't know the signs of the node expectations μ_v^p and we need a bound on $\mathbf{Var}(Z)$.

We employ the weak-learning framework described in Sections 4.1-4.4 to weakly learn a sign vector correlated with the true sign vector. In particular, since $X_v \sim \mathit{Rademacher}(1/2+\mu_v/2)$, from Corollary 1, we have that with k samples we can correctly estimate $\mathbf{sign}(\mu_v^p - \mu_v^q)$ with probability $1/2 + c_1\sqrt{k}\,|\mu_v^p - \mu_v^q|$. The rest of the argument for obtaining a sign vector which, with sufficient probability, preserves a sufficient amount of signal from the expected value of the statistic, proceeds in a similar way as before. However since the total number of terms we have in our expression is only linear we get some savings in the sample complexity.

And from Lemma 13, we have the following bound on functions $f_c(.)$ of the form $f_c(X) = \sum_{v\in V} c_v X_v$ (where $c \in \{\pm 1\}^V$) on the Ising model:

$$\mathbf{Var}(f_c(X)) = O(n).$$

By performing calculations analogous to the ones in Sections 4.3 and 4.4, we obtain that by using $\tilde{O}\left(\frac{n^{5/3}h^2}{\varepsilon^2}\right)$ samples we can test whether $\sum_{v\in V} |(\mu_v^p - \mu_v^q)| = 0$ or is $\geq \varepsilon/4h$ with probability $\geq 19/20$. If the tester outputs that $\sum_{v\in V} |(\mu_v^p - \mu_v^q)| = 0$, then we proceed to test whether $\sum_{u\neq v} |(\mu_{uv}^p - \mu_{uv}^q)| = 0$ or $\geq \varepsilon/4\beta$.

To perform this step, we begin by looking at the statistic Z used in Section 4.5:

$$Z = \sum_{u\neq v} \mathbf{sign}\left(\mu_{uv}^p - \mu_{uv}^q\right)\left(X_u X_v - \mu_{uv}^q\right)$$

as Z has the right expected value. We learn a sign vector which is weakly correlated with the true sign vector. However we need to obtain a variance bound on functions of the form $f_c(X) = \sum_{u\neq v} c_{uv}(X_u X_v - \mu_{uv}^q)$ where $c \in \{\pm 1\}^{\binom{V}{2}}$. By ignoring the constant term in $f_c(X)$, we get that,

$$\mathbf{Var}(f_c(X)) = \mathbf{Var}\left(\sum_{u\neq v} c_{uv} X_u X_v\right)$$

which can be $\Omega(n^3)$ as it is not appropriately centered. We employ this slightly worse variance bound to get a sample complexity of $\tilde{O}\left(\frac{n^{11/3}\beta^2}{\varepsilon^2}\right)$ for this part.

Theorem 8 captures the total sample complexity of our identity tester under the presence of external fields.

Theorem 8 (Identity Testing using Learn-Then-Test, Arbitrary External Field)**.** *Suppose p and q are Ising models in the high temperature regime under arbitrary external fields. The learn-then-test algorithm takes in $\tilde{O}\left(\frac{n^{5/3}h^2+n^{11/3}\beta^2}{\varepsilon^2}\right)$ i.i.d. samples from p and distinguishes between the cases $p = q$ and $d_{\mathrm{SKL}}(p,q) \geq \varepsilon$ with probability $\geq 9/10$.*

The tester is formally described in Algorithm 7.

Algorithm 7 Test if an Ising model p under an external field is identical to Ising model q

```
function TestIsing(sample access to an Ising model $p, \beta, d_{\max}, \varepsilon, \tau_1, \tau_2$, description of Ising model $q$)
    Run the localization Algorithm 2 on the nodes with accuracy parameter $\frac{\varepsilon}{2n^{\tau_1}}$. If it. identifies any nodes, return that $d_{\mathrm{SKL}}(p,q) \geq \varepsilon$
    for $\ell = 1$ to $O(n^{1-\tau_1})$ do
        Run the weak learning Algorithm 4 on $S = \{(X_u - Y_u\}_{u \in V}$, where $Y_u \sim$. $Rademacher(1/2+\mu_u^q/2)$, with parameters $\tau_1$ and $\varepsilon/2h$ to generate a sign vector $\vec{\Gamma}^{(\ell)}$ where $\Gamma_u^{(\ell)}$ is weakly correlated with $\mathbf{sign}\left(\mathbf{E}\left[X_u - \mu_u^q\right]\right)$
    end for
    Using the same set of samples for all $\ell$, run the testing algorithm of Lemma 11. on each of the $\vec{\Gamma}^{(\ell)}$ with parameters $\tau_3 = \tau_1, \delta = O(1/n^{1-\tau_1})$. If any output that $d_{\mathrm{SKL}}(p,q) \geq \varepsilon$, return that $d_{\mathrm{SKL}}(p,q) \geq \varepsilon$
    ____________________
    Run the localization Algorithm 2 on the edges with accuracy parameter $\frac{\varepsilon}{2n^{\tau_2}}$. If it. identifies any edges, return that $d_{\mathrm{SKL}}(p,q) \geq \varepsilon$
    for $\ell = 1$ to $O(n^{2-\tau_2})$ do
        Run the weak learning Algorithm 4 on $S = \{(X_uX_v - Y_{uv}\}_{u \neq v}$, where $Y_{uv} \sim$. $Rademacher(1/2+\mu_{uv}^q/2)$, with parameters $\tau_2$ and $\varepsilon/2\beta$ to generate a sign vector $\vec{\Gamma}^{(\ell)}$ where $\Gamma_{uv}^{(\ell)}$ is weakly correlated with $\mathbf{sign}\left(\mathbf{E}\left[X_uX_v - \mu_{uv}^q\right]\right)$
    end for
    Using the same set of samples for all $\ell$, run the testing algorithm of Lemma 11. on each of the $\vec{\Gamma}^{(\ell)}$ with parameters $\tau_4 = \tau_2, \delta = O(1/n^{2-\tau_2})$. If any output that $d_{\mathrm{SKL}}(p,q) \geq \varepsilon$, return that $d_{\mathrm{SKL}}(p,q) \geq \varepsilon$. Otherwise, return that $p = q$
end function
```

Chapter 5

Bounding the Variance of Functions of the Ising Model in the High-Temperature Regime

In this chapter, we describe our technique for bounding the variance of our statistics on the Ising model in high temperature. As the structure of Ising models can be quite complex, it can be challenging to obtain non-trivial bounds on the variance of even relatively simple statistics. In particular, to apply our learn-then-test framework of Chapter 4, we must bound the variance of statistics of the form $Z' = \sum_{u \neq v} c_{uv} X_u X_v$ (from (4.3)) and $Z'_{cen} = \sum_{u \neq v} c_{uv} \left(X_u^{(1)} - X_u^{(2)}\right)\left(X_v^{(1)} - X_v^{(2)}\right)$ (from (4.6)). While the variance for both the statistics is easily seen to be $O(n^2)$ if the graph has no edges, it proves challenging to prove variance bounds better than the trivial $O(n^4)$ for general graphs. In order to do this, we use the technique of exchangeable pairs, inspired by Chatterjee's thesis [27]. While a straightforward application of his result gives an improved bound of $O(n^3)$, we must extend his framework to achieve tighter bounds. We believe this technique may be of independent interest when analyzing statistics of distributions which exhibit such rich and complex structure. We state the main results of this chapter now. Our first result, Theorem 9, bounds the variance of functions of the form $\sum_{u \neq v} c_{uv} X_u X_v$ under no external field which captures the statistic used for testing independence and identity by the learn-then-test framework of Chapter 4 in the absence of an external field.

Theorem 9 (High Temperature Variance Bound, No External Field). *Let $c \in [-1,1]^{\binom{V}{2}}$ and define $f_c : \{\pm 1\}^V \to \mathbb{R}$ as follows: $f_c(x) = \sum_{i \neq j} c_{\{i,j\}} x_i x_j$. Let also X be distributed according to an Ising model, without node potentials (i.e. $\theta_v = 0$, for all v), in the high* *temperature regime of Definition 2. Then*

$$\mathbf{Var}\left(f_c(X)\right) = \tilde{O}(n^{1.5} \cdot \max_v |c_{\cdot v}|_2) + O\left(n^{2.5} \cdot \max_v |c_{\cdot v}|_2 \cdot d_{\max}^{1.5} \cdot \beta^3\right).$$

In particular, since $\beta \le 1/4d_{\max}$ and $\max_v |c_{\cdot v}|_2 \le \sqrt{n}$ for the function corresponding to our statistic of interest, the above bound is always $\tilde{O}(n^2) + \tilde{O}\left(\frac{n^3}{d_{\max}^{1.5}}\right)$. For dense graphs it is $\tilde{O}(n^2)$.

Our second result of this chapter, Theorem 10, bounds the variance of functions of the form $\sum_{u \ne v} c_{uv}(X_u^{(1)} - X_u^{(2)})(X_v^{(1)} - X_v^{(2)})$ which captures the statistic of interest for independence testing using the learn-then-test framework of Chapter 4 under an external field. Intuitively, this modification is required to "recenter" the random variables. Here, we view the two samples from Ising model p over graph $G = (V, E)$ as coming from a single Ising model $p^{\otimes 2}$ over a graph $G^{(1)} \cup G^{(2)}$ where $G^{(1)}$ and $G^{(2)}$ are identical copies of G.

Theorem 10 (High Temperature Variance Bound, Arbitrary External Field)**.** *Let $c \in [-1,1]^{\binom{V}{2}}$ and let X be distributed according to Ising model $p^{\otimes 2}$ over graph $G^{(1)} \cup G^{(2)}$ in the high temperature regime of Definition 2 and define $f_c : \{\pm 1\}^{V \cup V'} \to \mathbb{R}$ as follows: $f_c(x) = \sum_{\substack{u,v \in V \\ s.t.\ u \ne v}} c_{uv}(x_{u^{(1)}} - x_{u^{(2)}})(x_{v^{(1)}} - x_{v^{(2)}})$. Then*

$$\mathbf{Var}(f_c(X)) = \tilde{O}\left(n^{1.5} \max_v |c_{\cdot v}|_2\right) + \tilde{O}(n^{2.5} \max_v |c_{\cdot v}|_2 \cdot d_{\max} \cdot \beta^2).$$

In particular, since $\beta \le 1/4d_{\max}$ and $\max_v |c_{\cdot v}|_2 \le \sqrt{n}$, the above bound is always $\tilde{O}(n^2) + \tilde{O}\left(\frac{n^3}{d_{\max}}\right)$. For dense graphs it is $\tilde{O}(n^2)$.

5.1 Overview of the Technique

We will present an overview of the technique used to obtain the aforementioned results by considering the statistic of interest under the absence of an external field, Z'. The result for Z'_{cen} uses an extension of the same technique and is presented in greater detail in Section 5.3.

Given some $c \in [-1,1]^{\binom{V}{2}}$, we define $f_c : \{\pm 1\}^V \to \mathbb{R}$ as follows: $f_c(x) = \sum_{i \ne j} c_{\{i,j\}} x_i x_j$. To ease our notation, we will set $c_{ij} = c_{ji} = c_{\{i,j\}}$. We are interested to bound the variance of $f_c(X)$, when X is sampled from an Ising model p on graph $G = (V, E)$ with a parameter vectore $\vec{\theta}$. An obvious bound on the variance is $O(n^4 \max\{c_{ij}\}^2)$. On the other hand, if the Ising model was a product distribution, then the variance would be bounded by

$O(n^2 \max\{c_{ij}\}^2)$. Our goal is to match the variance bound for product distributions, in the high temperature regime. We will do this using exchangeable pairs. Our proof is inspired by Chapter 4 of Chatterjee's thesis [27], but it has significant differences from that development. Using technology lifted off from Chatterjee's thesis we can quite straightforwardly obtain a variance bound of $O(n^3 \max\{c_{ij}\}^2)$. Lemma 13 states the variance bound we get from Chatterjee's thesis [27]:

Lemma 13. *Consider any function $f(X)$ on the variables of the Ising model. Let c_i be the Lipschitz constant of $f(.)$ corresponding to variable X_i. That is,*

$$\frac{1}{2}\left|f(X_1,\dots,X_i,\dots,X_n) - f(X_1,\dots,X_i',\dots,X_n)\right| \le c_i$$

for any X_i, X_i' and for all possible values of $X_1,\dots,X_{i-1},X_{i+1},\dots,X_n$. In the high temperature regime,

$$\mathbf{Var}(f(X)) \le \sum_i c_i^2.$$

Our function of interest on the Ising model has a Lipschitz constant of $O(n)\max\{c_{ij}\}$. Hence by Lemma 13, in the high temperature regime

$$\mathbf{Var}\left(\sum_{i\neq j} c_{ij}X_iX_j\right) \le \max\{c_{ij}\}^2 \times n \times n^2 = O(n^3)\max\{c_{ij}\}^2. \tag{5.1}$$

To push the variance down further we need to develop new machinery, involving a different coupling and more delicate contraction arguments. We discuss these differences as we develop our bounds.

On with our argument, we consider an exchangeable pair (X, X') defined as follows: we sample a state X from the Ising model, and let X' be the state reached after one step of the Glauber dynamics from X. In particular, X' is obtained by choosing a node $v \in V$ uniformly at random, and sampling X_v' from the marginal distribution of the Ising model at v conditioning the state of v's neighbors to be $X_{N(v)}$. For all other nodes $u \neq v$, we set $X_u' = X_u$.

We are now seeking an antisymmetric function $F(x, x')$ such that:

$$\mathbf{E}\left[F(X,X')\middle|X\right] = f_c(X) - \mathbf{E}\left[f_c(X)\right]. \tag{5.2}$$

To identify one, we consider the evolution $(X_t)_t$ of the Glauber dynamics starting at some arbitrary state $X_0 = x$ and a coupled evolution $(X_t')_t$ of the Glauber dynamics starting at some state $X'_0 = x'$. Besides being a faithful coupling, our coupling should also satisfy the following property:

> **P**: For every initial values (x, x') and every t, the marginal distribution of X_t depends only on x and the marginal distribution of X'_t depends only on x'.

If our coupling satisfies property **P** and additionally

$$\forall (x, x') : \sum_{t=0}^{\infty} \left| \mathbf{E}\left[f_c(X_t) - f_c(X'_t) \middle| X_0 = x, X'_0 = x' \right] \right| < \infty, \tag{5.3}$$

then we can define our antisymmetric function F, satisfying (5.2), as follows:

$$F(x, x') = \sum_{t=0}^{\infty} \mathbf{E}\left[f_c(X_t) - f_c(X'_t) \middle| X_0 = x, X'_0 = x' \right], \tag{5.4}$$

i.e. we are summing the expected differences of our function applied to the trajectories of our coupled dynamics. That F, defined as above, satisfies (5.2) under Conditions **P** and (5.3), is simple and can be found as Lemma 4.2 in Chatterjee's thesis [27]. In terms of our exchangeable pair (X, X') and function F defined as above, we can express the variance of $f_c(X)$ as follows:

$$\mathbf{Var}\left(f_c(X)\right) = \frac{1}{2} \cdot \mathbf{E}\left[\left(f_c(X) - f_c(X')\right) \cdot F(X, X') \right]. \tag{5.5}$$

Henceforth, to bound the variance of $f_c(X)$ we will bound the RHS of (5.5). We shall do this in a few steps.

5.1.1 Choosing a Coupling

We will be considering the following coupling of $(X_t)_t$ and $(X'_t)_t$. At every time step $t > 0$, to set (X_t, X'_t) in terms of (X_{t-1}, X'_{t-1}), we choose to update the same (uniformly randomly chosen) node v in both chains. However, we will set this node in X_t and X'_t independently. We call our coupling the "generous coupling," in contrast to the "greedy coupling" used by Chatterjee, where the state of node v in the two chains is set so as to maximize the probability of agreement. Intuitively, a greedy coupling appears effective, as our ultimate goal is to bound the RHS of (5.5). Given that $F(X, X')$ involves a summation over the differences of $f_c(\cdot)$ applied to the trajectories of the two chains, as per (5.4), a reasonable approach is to bias the coupling towards minimizing the Hamming distance between X_t and X'_t. Despite this intuition, we elect not to use the greedy coupling for our analysis. Using our generous coupling, enables us to improve by a factor of $\Omega(n)$ the variance bounds obtained in Section 4.2 of Chatterjee's thesis, and by factor of $\Omega(n^2)$ the naive bounds.

5.1.2 Establishing Contraction and Completing the Proof

At this point, we could follow Chatterjee's recipe and obtain a variance bound of $O(n^3)$ as follows. First, expanding out the expression for $F(x, x')$, we get that

$$\mathbf{Var}\left(f_c(X)\right) = \frac{1}{2}\sum_{t=0}^{\infty}\mathbf{E}\left[(f_c(X_0) - f_c(X_0')) \cdot \mathbf{E}\left[f_c(X_t) - f_c(X_t') \middle| X_0, X_0'\right]\right].$$

Since the mixing time of this chain is $t^* = O(n \log n)$, the sum of the contributions of terms $t > t^*$ is negligible, and thus, we must bound $|f_c(X_t) - f_c(X_t')|$ only for $t = O(n \log n)$. We note that $|f_c(X) - f_c(X')| \leq \sum_i n \mathbb{1}_{\{\sum_j X_j \neq X_j'\}}$. Chatterjee shows that if f satisfies such a Lipschitz condition, it implies the bound $\mathbf{Var}(f_c(X)) \leq \sum_i n^2 = O(n^3)$.

We diverge from his strategy, and apply a more careful argument. First, instead of showing that a specific function contracts, we must show that a family of related multilinear functions, with different coefficients, contracts simultaneously. Secondly, since we are not using Hamming distance as a measure of progress, and we are doing a generous coupling instead of Chatterjee's greedy coupling, we need to deal more directly with the non-linearities of the Glauber updates. This involves linearizing the tanh function, which comes at the cost of quadratic or cubic error terms which accumulate as we backpropagate our contraction bound from time t^* to time 0. To control these error terms, we must bootstrap the concentration of *linear* functions of the Ising model, which can be proven by appealing directly to Chatterjee's results without loss. Ultimately, our variance bounds also imply tight concentration results for multilinear functions of the Ising model, which are similarly better by a factor of $O(n)$ in comparison to Chatterjee.

Our variance bound for the relevant statistics of interest in the presence of external fields is slightly worse. More details on the proof of Theorem 9 are given in Section 5.2. Theorem 10 is proven in Section 5.3.

5.2 Bounding Variance of $f_c(\cdot)$, No External Field

In this section, we prove Theorem 9. We recall the statement of Theorem 9,

Theorem 9 (High Temperature Variance Bound, No External Field)**.** *Let $c \in [-1, 1]^{\binom{V}{2}}$ and define $f_c : \{\pm 1\}^V \to \mathbb{R}$ as follows: $f_c(x) = \sum_{i \neq j} c_{\{i,j\}} x_i x_j$. Let also X be distributed according to an Ising model, without node potentials (i.e. $\theta_v = 0$, for all v), in the high* *temperature regime of Definition 2. Then*

$$\mathbf{Var}\left(f_c(X)\right) = \tilde{O}(n^{1.5} \cdot \max_v |c_{\cdot v}|_2) + O\left(n^{2.5} \cdot \max_v |c_{\cdot v}|_2 \cdot d_{\max}^{1.5} \cdot \beta^3\right).$$

In particular, since $\beta \leq 1/4d_{\max}$ and $\max_v |c_{\cdot v}|_2 \leq \sqrt{n}$ for the function corresponding to our statistic of interest, the above bound is always $\tilde{O}(n^2) + \tilde{O}\left(\frac{n^3}{d_{\max}^{1.5}}\right)$. For dense graphs it is $\tilde{O}(n^2)$.

5.2.1 Establishing Contraction

We now need to show that as our coupled dynamics evolve, the $f_c(X_t) - f_c(X'_t)$ contracts. We first establish a one-step contraction in the following statement. The terms involving function $e(\cdot)$ are error terms.

Lemma 14. *Consider the vector function $g(\cdot)$ mapping a vector $c \in \mathbb{R}^{\binom{V}{2}}$ to the following vector: $g(c)_{\{u,w\}} := \sum_{v \in N(w)} c_{uv}\theta_{wv} + \sum_{v \in N(u)} c_{wv}\theta_{uv}$, for all $w \neq u$. Consider also a pair of coupled executions $(X_t)_t$, $(X'_t)_t$ of the Glauber dynamics on some Ising model, starting from a pair of arbitrary states X_0, X'_0. Suppose these executions are coupled using the generous coupling of Section 5.1.1. If the Ising model has no node potentials (i.e. $\theta_v = 0, \forall v$), then for all t and point-wise with respect to X_t, X'_t:*

$$\mathbf{E}\left[f_c(X_{t+1}) - f_c(X'_{t+1}) \mid X_t, X'_t\right] = \left(1 - \frac{2}{n}\right)\left(f_c(X_t) - f_c(X'_t)\right) + \frac{1}{n}\left(f_{g(c)}(X_t) - f_{g(c)}(X'_t)\right) \pm e(c, X_t) \pm e(c, X'_t),$$

where $e(\cdot)$ is the non-negative function defined as follows:

$$e(c, X_t) = \frac{1}{3n}\sum_v \left|\sum_{u \neq v} c_{uv} X_{t,u}\right| \left|\sum_{w \in N(v)} \theta_{wv} X_{t,w}\right|^3.$$

Proof of Lemma 14: For all X_t, X'_t:

$$\begin{aligned}
&\mathbf{E}\left[f_c(X_{t+1}) - f_c(X'_{t+1}) \,\middle|\, X_t, X'_t\right] = \\
&\quad = \frac{1}{n}\sum_v \mathbf{E}\left[f_c(X_{t+1}) - f_c(X'_{t+1}) \,\middle|\, X_t, X'_t, \text{node } v \text{ is chosen in step } t+1\right] \\
&\quad = \frac{1}{n}\sum_v \left(f_c(X_t) - \sum_{u \neq v} c_{uv} X_{t,u} X_{t,v} - f_c(X'_t) + \sum_{u \neq v} c_{uv} X'_{t,u} X'_{t,v} \right) + \qquad (5.6) \\
&\qquad + \frac{1}{n}\sum_v \left(\sum_{u \neq v} c_{uv} X_{t,u} \tanh\left(\sum_{w \in N(v)} \theta_{wv} X_{t,w} \right) - \sum_{u \neq v} c_{uv} X'_{t,u} \tanh\left(\sum_{w \in N(v)} \theta_{wv} X'_{t,w} \right) \right) \qquad (5.7) \\
&\quad = \left(1 - \frac{2}{n}\right)\left(f_c(X_t) - f_c(X'_t)\right) \qquad (5.8) \\
&\qquad + \frac{1}{n}\sum_v \sum_{u \neq v} c_{uv} X_{t,u} \tanh\left(\sum_{w \in N(v)} \theta_{wv} X_{t,w} \right) - \frac{1}{n}\sum_v \sum_{u \neq v} c_{uv} X'_{t,u} \tanh\left(\sum_{w \in N(v)} \theta_{wv} X'_{t,w} \right) \qquad (5.9)
\end{aligned}$$

where Line (5.6) accounts for the terms of $f_c(X_{t+1})$ and $f_c(X'_{t+1})$ that stay untouched when we randomly chose to update node v in our coupled dynamics, while Line (5.7) accounts for the terms that do change. Given our generous coupling, the values of $X_{t+1,v}$ and $X'_{t+1,v}$ are set independently from their marginal distributions conditioning on X_t and X'_t respectively, and their expectations are the expressions involving $\tanh(\cdot)$ in Line (5.7). Finally, in (5.8) we rewrote (5.6) more neatly, emphasizing a contraction that takes place, while (5.9) just replicates (5.7).

Our goal next is to get rid of the tanh's. We start with a trivial claim:

Claim 2. *$|\tanh(x) - x| \le \frac{|x|^3}{3}$ for all $x \in \mathbb{R}$.*

Using derivation (5.6)-(5.9), and Claim 2 we get that

$$\begin{aligned}
\mathbf{E}\left[f_c(X_{t+1}) - f_c(X'_{t+1}) \mid X_t, X'_t\right] &= \left(1-\frac{2}{n}\right)\left(f_c(X_t) - f_c(X'_t)\right) \\
&\quad + \frac{1}{n}\sum_v \sum_{u\neq v} c_{uv}X_{t,u} \sum_{w\in N(v)} \theta_{wv}X_{t,w} - \frac{1}{n}\sum_v \sum_{u\neq v} c_{uv}X'_{t,u} \sum_{w\in N(v)} \theta_{wv}X'_{t,w} \qquad (5.10) \\
&\quad \pm \frac{1}{3n}\sum_v \left|\sum_{u\neq v} c_{uv}X_{t,u}\right| \left|\sum_{w\in N(v)} \theta_{wv}X_{t,w}\right|^3 \pm \frac{1}{3n}\sum_v \left|\sum_{u\neq v} c_{uv}X'_{t,u}\right| \left|\sum_{w\in N(v)} \theta_{wv}X'_{t,w}\right|^3 \\
&= \left(1-\frac{2}{n}\right)\left(f_c(X_t) - f_c(X'_t)\right) \\
&\quad + \frac{1}{n}\sum_{u\neq w}\left(\sum_{v\in N(w)} c_{uv}\theta_{wv} + \sum_{v\in N(u)} c_{wv}\theta_{uv}\right)\left(X_{t,u}X_{t,w} - X'_{t,u}X'_{t,w}\right) \qquad (5.11) \\
&\qquad + \frac{1}{n}\sum_u \left(\sum_{v\in N(u)} c_{uv}\theta_{uv}\right)\left(X^2_{t,u} - X'^2_{t,u}\right) \qquad (5.12) \\
&\quad \pm e(c, X_t) \pm e(c, X'_t) \\
&= \left(1-\frac{2}{n}\right)\left(f_c(X_t) - f_c(X'_t)\right) + \frac{1}{n}\left(f_{g(c)}(X_t) - f_{g(c)}(X'_t)\right) \pm e(c, X_t) \pm e(c, X'_t), \qquad (5.13)
\end{aligned}$$

where the sum of (5.11) and (5.12) is a rewriting of (5.10), (5.12) is actually 0, and $g(\cdot)$, $e(\cdot)$ are defined as in the statement of the lemma. □

Using Lemma 14, we can establish a multi-step contraction. The terms involving function $e_2(\cdot)$ in the statement, encapsulate the error that is being accumulated and needs to be controlled:

Lemma 15. *Consider the same setup as that of Lemma 14. Then, for all t and point-wise with respect to X_0, X'_0:*

$$\begin{aligned}
\mathbf{E}\left[f_c(X_t) - f_c(X'_t) \mid X_0, X'_0\right] &= \sum_{\ell=0}^{t}\binom{t}{\ell}\left(1-\frac{2}{n}\right)^{t-\ell}\left(\frac{1}{n}\right)^{\ell}\cdot\left(f_{g^{\circ\ell}(c)}(X_0) - f_{g^{\circ\ell}(c)}(X'_0)\right) \\
&\quad \pm e_2^t(c, X_0) \pm e_2^t(c, X'_0),
\end{aligned}$$

where $g^{\circ\ell}(\cdot)$ denotes the ℓ-fold composition of g with itself, and $e_2^t(\cdot)$ is the non-negative function defined as follows in terms of function $e(\cdot)$ of the statement of Lemma 14:

$$e_2^t(c, X_0) = \sum_{\ell=0}^{t-1}\sum_{q=0}^{t-1-\ell}\binom{t-1-\ell}{q}\left(1-\frac{2}{n}\right)^{t-1-\ell-q}\left(\frac{1}{n}\right)^{q}\mathbf{E}\left[e(g^{\circ q}(c), X_\ell)|X_0\right].$$

Proof of Lemma 15: The proof uses Lemma 14, and property **P** of our coupling, and proceeds by induction. It is straightforward to verify that the base case for induction, $t = 1$, follows from Lemma 14. Assume the statement holds for some $t > 1$. We will show that it holds for $t + 1$ as well. First, from the law of iterated expectations, we have,

$$\mathbf{E}\left[f_c(X_{t+1}) - f_c(X'_{t+1}) \mid X_0, X'_0\right] = \mathbf{E}\left[\mathbf{E}\left[f_c(X_{t+1}) - f_c(X_{t+1}) \mid X_t, X'_t\right] \mid X_0, X'_0\right]$$

Therefore from Lemma 14, we get

$$\begin{aligned}
\mathbf{E}\left[f_c(X_{t+1}) - f_c(X'_{t+1}) \mid X_0, X'_0\right] &= \left(1 - \frac{2}{n}\right)\mathbf{E}\left[f_c(X_t) - f_c(X'_t) \mid X_0, X'_0\right] + \\
&+ \frac{1}{n}\mathbf{E}\left[f_{g(c)}(X_t) - f_{g(c)}(X'_t) \mid X_0, X'_0\right] \pm \mathbf{E}\left[e(c, X_t) \mid X_0\right] \pm \mathbf{E}\left[e(c, X'_t) \mid X'_0\right] \\
&= \left(1 - \frac{2}{n}\right)\left(\sum_{\ell=0}^{t}\binom{t}{\ell}\left(1 - \frac{2}{n}\right)^{t-\ell}\left(\frac{1}{n}\right)^{\ell} \cdot \left(f_{g^{\circ\ell}(c)}(X_0) - f_{g^{\circ\ell}(c)}(X'_0)\right)\right) \\
&+ \frac{1}{n}\left(\sum_{\ell=0}^{t}\binom{t}{\ell}\left(1 - \frac{2}{n}\right)^{t-\ell}\left(\frac{1}{n}\right)^{\ell} \cdot \left(f_{g^{\circ\ell+1}(c)}(X_0) - f_{g^{\circ\ell+1}(c)}(X'_0)\right)\right) \\
&\pm \left(1 - \frac{2}{n}\right)e_2^t(c, X_0) \pm \left(1 - \frac{2}{n}\right)e_2^t(c, X'_0) \pm \frac{1}{n}e_2^t(c, X_0) \\
&\pm \frac{1}{n}e_2^t(c, X'_0) \pm \mathbf{E}\left[e(c, X_t) \mid X_0\right] \pm \mathbf{E}\left[e(c, X'_t) \mid X'_0\right] \\
&= \sum_{\ell=0}^{t+1}\binom{t+1}{\ell}\left(1 - \frac{2}{n}\right)^{t+1-\ell}\left(\frac{1}{n}\right)^{\ell} \cdot \left(f_{g^{\circ\ell}(c)}(X_0) - f_{g^{\circ\ell}(c)}(X'_0)\right) \\
&\pm \left(\left(1 - \frac{2}{n}\right)e_2^t(c, X_0) + \frac{1}{n}e_2^t(c, X_0) + \mathbf{E}\left[e(c, X_t) \mid X_0\right]\right) \\
&\pm \left(\left(1 - \frac{2}{n}\right)e_2^t(c, X'_0) + \frac{1}{n}e_2^t(c, X'_0) + \mathbf{E}\left[e(c, X'_t) \mid X'_0\right]\right).
\end{aligned}$$

It can be verified that $\left(1 - \frac{2}{n}\right)e_2^t(c, X_0) + \frac{1}{n}e_2^t(c, X_0) + \mathbf{E}\left[e(c, X_t) \mid X_0\right] = e_2^{t+1}(c, X_0)$ using the definition of $e_2^t(.)$ from the statement of the Lemma. Therefore, by induction, this shows the statement is true for all $t \geq 1$. □

5.2.2 Bounding the Variance of $f_c(\cdot)$ under High Temperature, No External Field

We are now ready to bound the variance of $f_c(\cdot)$, using (5.5), (5.4), our generous coupling of Section 5.1.1 and our recently established contraction property achieved by this coupling (Lemma 15). We prove Theorem 9.

Proof of Theorem 9: (5.5) and (5.4) give

$$\begin{aligned}\mathbf{Var}\left(f_c(X)\right) &= \frac{1}{2}\cdot \mathbf{E}\left[(f_c(X)-f_c(X'))\cdot F(X,X')\right]\\ &= \frac{1}{2}\sum_{t=0}^{\infty}\mathbf{E}\left[(f_c(X_0)-f_c(X_0'))\cdot \mathbf{E}\left[f_c(X_t)-f_c(X_t')\middle|X_0,X_0'\right]\right]. \qquad (5.14)\end{aligned}$$

In Lemma 24, we establish that the mixing time of the Ising model under high temperature is $O(n\log n)$. In fact, it follows from our proof of Lemma 24 that, for all t^*, if we start the Glauber dynamics from an arbitrary state X_0, then the total variation between the state, X_{t^*}, of the dynamics at time t^* and a random sample from the Ising model is bounded by $n\left(1-\frac{1-\eta}{n}\right)^{t^*}$. Hence, for large enough $t^*=\Omega(n\log n)$:

$$|\mathbf{E}\left[f_c(X_t)-f_c(X_t')\middle|X_0,X_0'\right]| \le ne^{-(1-\eta)\frac{t^*}{n}}4n^2\max|c_{ij}| = 4e^{-(1-\eta)\frac{t^*}{n}}n^3\max|c_{ij}|,$$

where $n^2\max|c_{ij}|$ is a trivial bound on the maximum absolute value of $f_c(\cdot)$. Hence, for large enough $t^*=\Omega(n\log n)$:

$$\mathbf{E}\left[|f_c(X_0)-f_c(X_0')|\cdot|\mathbf{E}\left[f_c(X_t)-f_c(X_t')\middle|X_0,X_0'\right]|\right] \le 8e^{-(1-\eta)\frac{t^*}{n}}n^5\max|c_{ij}|^2.$$

This implies that for large enough $t^*=\Omega(n\log n)$:

$$\begin{aligned}\frac{1}{2}\sum_{t=t^*}^{\infty}\mathbf{E}\left[(f_c(X_0)-f_c(X_0'))\cdot\mathbf{E}\left[f_c(X_t)-f_c(X_t')\middle|X_0,X_0'\right]\right] &\le 4n^5\max|c_{ij}|^2\sum_{t=t^*}^{\infty}e^{-(1-\eta)\frac{t^*}{n}}\\ &\le 4n^5\max|c_{ij}|^2e^{-(1-\eta)\frac{t^*}{n}}\frac{1}{1-e^{-(1-\eta)\frac{1}{n}}}\\ &\le 4n^5\max|c_{ij}|^2e^{-(1-\eta)\frac{t^*}{n}}\frac{n}{1-\eta}\le\max|c_{ij}|^2\le 1. \qquad (5.15)\end{aligned}$$

The above shows that we only need to bound (5.14) for t ranging from 0 to some $t^*=O(n\log n)$. It also shows that Condition 5.3, required for our anti-symmetric function $F()$ to be well-defined, holds.

Bounding (5.14) for t ranging from 0 to some $t^*=O(n\log n)$ requires more work. Let us take one of the terms, and plug in our bound from Lemma 15. Given that the bound of the lemma holds point-wise and $e_2()$ is non-negative we have:

$$\begin{aligned}
&\mathbf{E}\left[(f_c(X_0) - f_c(X_0')) \cdot \mathbf{E}\left[f_c(X_t) - f_c(X_t') \middle| X_0, X_0'\right]\right] \leq \\
&\quad \sum_{\ell=0}^{t} \binom{t}{\ell} \left(1 - \frac{2}{n}\right)^{t-\ell} \left(\frac{1}{n}\right)^{\ell} \cdot \mathbf{E}\left[\left|f_c(X_0) - f_c(X_0')\right| \left|f_{g^{\circ\ell}(c)}(X_0) - f_{g^{\circ\ell}(c)}(X_0')\right|\right] \qquad (5.16)\\
&\qquad + \mathbf{E}\left[\left|f_c(X_0) - f_c(X_0')\right| e_2^t(c, X_0)\right] + \mathbf{E}\left[\left|f_c(X_0) - f_c(X_0')\right| e_2^t(c, X_0')\right]. \qquad (5.17)
\end{aligned}$$

Now, recall that the pair (X_0, X_0') is sampled as follows: X_0 is a sample from the Ising model, and X_0' is one step of the Glauber dynamics from X_0. So:

$$\left|f_c(X_0) - f_c(X_0')\right| \leq 2 \max_v \left|\sum_{u \neq v} c_{uv} X_{0,u}\right|.$$

It follows from Lemma 25 that, for all v, a sample X_0 from an Ising model (without node potentials that we are analyzing) satisfies:

$$\Pr\left[\left|\sum_{u \neq v} c_{uv} X_{0,u}\right| \geq t\right] \leq 2e^{-\frac{(1-\eta)t^2}{4\sum_{u \neq v} c_{uv}^2}},$$

where η is the constant from Definition 2. So for sufficiently large $t = \Omega(\sqrt{\log n} \cdot |c_{\cdot v}|_2)$, with probability at least $1 - \frac{1}{8n^3}$: $\left|\sum_{u \neq v} c_{uv} X_{0,u}\right| < t$. It follows that, with probability at least $1 - 1/8n^2$, $\max_v \left|\sum_{u \neq v} c_{uv} X_{0,u}\right| = O(\sqrt{\log n} \cdot \max_v |c_{\cdot v}|_2)$. Hence, with probability at least $1 - 1/8n^2$:

$$\left|f_c(X_0) - f_c(X_0')\right| \leq O(\sqrt{\log n} \cdot \max_v |c_{\cdot v}|_2).$$

By a similar token, for any fixed ℓ, with probability at least $1 - 1/8n^2$:

$$\left|f_{g^{\circ\ell}(c)}(X_0) - f_{g^{\circ\ell}(c)}(X_0')\right| \leq O(\sqrt{\log n} \cdot \max_v \left|g^{\circ\ell}(c)_{\cdot v}\right|_2).$$

At the same time, the maximum that $2\max_v \left|\sum_{u \neq v} c_{uv} X_{0,u}\right|$ (and hence $|f_c(X_0) - f_c(X_0')|$) can possibly be is $2\max_v |c_{\cdot v}|_1 \leq 2n$. Notice that in the regime of Definition 2, function g maps points in $[-1, 1]^{\binom{V}{2}}$ to the same set. Hence the maximum that $\left|f_{g^{\circ\ell}(c)}(X_0) - f_{g^{\circ\ell}(c)}(X_0')\right|$ can possibly be is also at most $2n$, for any ℓ.

It follows from the above calculations that:

$$\mathbf{E}\left[\left|f_c(X_0) - f_c(X_0')\right| \left|f_{g^{\circ \ell}(c)}(X_0) - f_{g^{\circ \ell}(c)}(X_0')\right|\right] \leq O(\log n \cdot \max_v |c_{\cdot v}|_2 \cdot \max_v \left|g^{\circ \ell}(c)_{\cdot v}\right|_2)$$

$$\leq O(\sqrt{n}\log n \cdot \max_v |c_{\cdot v}|_2) \tag{5.18}$$

Given that this bound holds for any ℓ, and recognizing the binomial expansion in (5.16), we obtain the bound:

$$(5.16) \leq O(\sqrt{n}\log n \cdot \max_v |c_{\cdot v}|_2).$$

It remains to bound the error terms (5.17). For a fixed ℓ and q let us try to bound the term $\mathbf{E}\left[e(g^{\circ q}(c), X_\ell) | X_0\right]$, involved in the definition of $e_2^t(c, X_0)$. For convenience set $c' = g^{\circ q}(c)$, and recall (as we have pointed out above) that $c' \in [-1,1]^{\binom{V}{2}}$. Recalling the definition of $e()$ from the statement of Lemma 14, we have that:

$$\mathbf{E}\left[e(c', X_\ell) | X_0\right] = \mathbf{E}\left[\frac{1}{3n}\sum_v \left|\sum_{u \neq v} c'_{uv} X_{\ell,u}\right| \left|\sum_{w \in N(v)} \theta_{wv} X_{\ell,w}\right|^3 \middle| X_0\right].$$

Given that X_0 is sampled from the Ising model, and X_ℓ is the state reached after ℓ steps of the Glauber dynamics from X_0, it follows that X_ℓ is also a sample from the Ising model. So a similar analysis as the one we did earlier implies that for a fixed v, with probability at least $1 - \frac{1}{2n^{21}}$: $\left|\sum_{u \neq v} c'_{uv} X_{\ell,u}\right| < O(\sqrt{\log n} \cdot |c'_{\cdot v}|_2)$. So, with probability at least $1 - \frac{1}{2n^{20}}$, simultaneously for all v:

$$\left|\sum_{u \neq v} c'_{uv} X_{\ell,u}\right| \leq O(\sqrt{\log n} \cdot \sqrt{n}).$$

Via similar arguments, it can be shown that, with probability at least $1 - \frac{1}{2n^{20}}$, simultaneously for all v:

$$\left|\sum_{w \in N(v)} \theta_{wv} X_{\ell,w}\right| \leq O\left(\sqrt{\log n \cdot d_{\max}} \cdot \beta\right),$$

where we used that our working regime is the high-temperature regime of Definition 2.

So it follows from the above that, with probability at least $1 - 1/n^{20}$, it holds that:

$$\frac{1}{3n}\sum_v \left|\sum_{u \neq v} c'_{uv} X_{\ell,u}\right| \left|\sum_{w \in N(v)} \theta_{wv} X_{\ell,w}\right|^3 \leq O\left(\sqrt{n}\log^2 n \cdot d_{\max}^{1.5} \cdot \beta^3\right).$$

Let us call the event that the above holds $\mathcal{E}$. We want to view this event as a function

$\mathcal{E} = \mathcal{E}(X_0, G_\ell)$ of X_0 and the decisions G_ℓ that the Glauber dynamics made in the first ℓ steps. Indeed, we want to view X_0 and G_ℓ as independent random variables. G_ℓ samples independently of X_0 which nodes it will update, together with ℓ uniform $[0,1]$ random variables. Then the Glauber dynamics are a deterministic function of X_0 and G_ℓ. With this perspective in mind, we have from the above that:

$$\Pr_{X_0,G_\ell}[\mathcal{E}(X_0,G_\ell)] \geq 1 - \frac{1}{n^{20}}.$$

From this it follows that

$$\Pr_{X_0}\left[\Pr_{G_\ell}[\mathcal{E}(X_0,G_\ell)] \geq 1 - 1/n^9\right] \geq 1 - 1/n^9.$$

In turn, the above implies that

$$\Pr_{X_0}\left[\mathbf{E}\left[\frac{1}{3n}\sum_v \left|\sum_{u \neq v} c'_{uv} X_{\ell,u}\right| \left|\sum_{w \in N(v)} \theta_{wv} X_{\ell,w}\right|^3 \middle| X_0\right] \leq O\left(\sqrt{n}\log^2 n \cdot d_{\max}^{1.5} \cdot \beta^3\right)\right] \geq 1 - 1/n^9.$$

i.e.

$$\Pr_{X_0}\left[\mathbf{E}\left[e(c', X_\ell) | X_0\right] \leq O\left(\sqrt{n}\log^2 n \cdot d_{\max}^{1.5} \cdot \beta^3\right)\right] \geq 1 - 1/n^9. \tag{5.19}$$

From similar analysis to the one we did earlier we also have:

$$\Pr_{X_0}\left[\left|f_c(X_0) - f_c(X'_0)\right| \leq O(\sqrt{\log n} \cdot \max_v |c_{\cdot v}|_2))\right] \geq 1 - 1/n^9. \tag{5.20}$$

So (5.19) and (5.20) imply:

$$\mathbf{E}\left[\left|f_c(X_0) - f_c(X'_0)\right| \mathbf{E}\left[e(c', X_\ell) | X_0\right]\right] \leq O\left(\sqrt{n}\log^{2.5} n \cdot \max_v |c_{\cdot v}|_2 \cdot d_{\max}^{1.5} \cdot \beta^3\right). \tag{5.21}$$

Now the definition of function $e_2^t(\cdot)$ in the statement of Lemma 15 and (5.21) imply that:

$$\mathbf{E}\left[\left|f_c(X_0) - f_c(X'_0)\right| e_2^t(c, X_0)\right] \leq O\left(t \cdot \sqrt{n}\log^{2.5} n \cdot \max_v |c_{\cdot v}|_2 \cdot d_{\max}^{1.5} \cdot \beta^3\right).$$

The same bound applies to $\mathbf{E}\left[|f_c(X_0) - f_c(X'_0)|\, e_2^t(c, X'_0)\right]$. So we have successfully bounded (5.17).

Using our bounds for (5.16) and (5.17), we get that:

$$\mathbf{E}\left[(f_c(X_0) - f_c(X_0')) \cdot \mathbf{E}\left[f_c(X_t) - f_c(X_t') | X_0, X_0'\right]\right] \tag{5.22}$$
$$\leq O(\sqrt{n}\log n \cdot \max_v |c_{\cdot v}|_2) + O\left(t \cdot \sqrt{n}\log^{2.5} n \cdot \max_v |c_{\cdot v}|_2 \cdot d_{\max}^{1.5} \cdot \beta^3\right). \tag{5.23}$$

So we can go back to (5.14) to bound the first t^* terms of the summation, for $t^* = O(n \log n)$ as set earlier. We get:

$$\frac{1}{2}\sum_{t=0}^{t^*} \mathbf{E}\left[(f_c(X_0) - f_c(X_0')) \cdot \mathbf{E}\left[f_c(X_t) - f_c(X_t') | X_0, X_0'\right]\right] =$$
$$= O(n^{1.5}\log^2 n \cdot \max_v |c_{\cdot v}|_2) + O\left(n^{2.5}\log^{4.5} n \cdot \max_v |c_{\cdot v}|_2 \cdot d_{\max}^{1.5} \cdot \beta^3\right) \tag{5.24}$$

Plugging (5.24) and (5.15) into (5.14), we bound the variance as follows:

$$\mathbf{Var}\left(f_c(X)\right) = \tilde{O}(n^{1.5} \cdot \max_v |c_{\cdot v}|_2) + O\left(n^{2.5} \cdot \max_v |c_{\cdot v}|_2 \cdot d_{\max}^{1.5} \cdot \beta^3\right).$$

□

5.3 Bounding the Variance of $f_c(\cdot)$, Arbitrary External Field

Extending our techniques from Section 5.2, we obtain a variance bound for the centered multi-linear function on arbitrary Ising models. Firstly, we note that the non-centered function $\sum_{u \neq v} X_u X_v$ can have a variance $O(n^3)$ even in the case the Ising model is product, i.e. has no edges. This is because the function $\sum_{u \neq v} X_u X_v$ is not appropriately centered when external fields are present. We show a better variance bound on our centered statistic for independence testing under an external field, as stated in equation (4.6). Recall from (4.6), that $Z'_{cen} = \sum_{u \neq v} c_{uv}\left(X_u^{(1)} - X_u^{(2)}\right)\left(X_v^{(1)} - X_v^{(2)}\right)$ is a function of two independent samples from an Ising model p. Together, the two samples can be viewed as a single sample from an Ising model which consists of two copies of p put next to each other. The new Ising model $p^{\otimes 2}$ has the underlying graph $G^{(1)} + G^{(2)}$, where $G^{(1)}$ and $G^{(2)}$ are identical copies of G. Note that $p^{\otimes 2}$ is also in the high temperature regime. The statistic Z'_{cen} now becomes a multi-linear function of the variables in the Ising model $p^{\otimes 2}$. We can then apply the exchangeable pairs technique described in Section 5.1 to $p^{\otimes 2}$ to show a variance bound for functions of the form

$$f_c(X) = \sum_{u \neq v} c_{uv} \left(X_{u^{(1)}} - X_{u^{(2)}}\right) \left(X_{v^{(1)}} - X_{v^{(2)}}\right)$$

where $c \in [-1,1]^{\binom{V}{2}}$. This will directly imply a bound for $\mathbf{Var}(Z'_{cen})$. The proof will again proceed by considering two coupled executions $\{X_t\}_t, \{X'_t\}_t$ of the Glauber dynamics on the two sample Ising model $\pi^{\otimes 2}$.

Our bound for $\mathbf{Var}(f_c(X))$, stated in Theorem 10, is only slightly worse than the one without node potentials (from Theorem 9):

Theorem 10 (High Temperature Variance Bound, Arbitrary External Field)**.** *Let $c \in [-1,1]^{\binom{V}{2}}$ and let X be distributed according to Ising model $p^{\otimes 2}$ over graph $G^{(1)} \cup G^{(2)}$ in the high temperature regime of Definition 2 and define $f_c : \{\pm 1\}^{V \cup V'} \to \mathbb{R}$ as follows: $f_c(x) = \sum_{\substack{u,v \in V \\ s.t.\ u \neq v}} c_{uv}(x_{u^{(1)}} - x_{u^{(2)}})(x_{v^{(1)}} - x_{v^{(2)}})$. Then*

$$\mathbf{Var}(f_c(X)) = \tilde{O}\left(n^{1.5} \max_v |c_{\cdot v}|_2\right) + \tilde{O}(n^{2.5} \max_v |c_{\cdot v}|_2 \cdot d_{\max} \cdot \beta^2).$$

In particular, since $\beta \leq 1/4d_{\max}$ and $\max_v |c_{\cdot v}|_2 \leq \sqrt{n}$, the above bound is always $\tilde{O}(n^2) + \tilde{O}\left(\frac{n^3}{d_{\max}}\right)$. For dense graphs it is $\tilde{O}(n^2)$.

The proof of Theorem 10 follows along similar lines as the proof of Theorem 9. The first step would be to establish contraction of our coupled dynamics $f_c(X_t) - f_c(X'_t)$ as t grows. We show this in the following statement. The terms involving function $e(.)$ are error terms.

Lemma 16. *Consider the vector function $g(\cdot)$ mapping a vector $c \in \mathbb{R}^{\binom{V}{2}}$ to the following vector: $g(c)_{\{u,w\}} := \sum_{v \in N(w)} c_{uv} \operatorname{sech}^2(\sigma_v)\theta_{wv} + \sum_{v \in N(u)} c_{wv} \operatorname{sech}^2(\sigma_v)\theta_{uv}$, for all $w \neq u$, where $\sigma_v = \theta_v + \sum_{w \in N(v)} \theta_{wv}\mu_w$. Consider also a pair of coupled executions $(X_t)_t$, $(X'_t)_t$ of the Glauber dynamics on some Ising model, starting from a pair of arbitrary states X_0, X'_0. Suppose these executions are coupled using the generous coupling of Section 5.1.1. Then for all t and point-wise with respect to X_t, X'_t:*

$$\mathbf{E}\left[f_c(X_{t+1}) - f_c(X'_{t+1}) \mid X_t, X'_t\right] = \left(1 - \frac{1}{n}\right)\left(f_c(X_t) - f_c(X'_t)\right) + \frac{1}{n}\left(f_{g(c)}(X_t) - f_{g(c)}(X'_t)\right)$$
$$\pm e(c, X_t) \pm e(c, X'_t),$$

where $e(\cdot)$ is the non-negative function defined as follows:

$$\begin{aligned} e(c, X_t) = &\frac{1}{2n} \sum_{v \in V} \left| \sum_{u \neq v} c_{uv} (X_{t,u^{(1)}} - X_{t,u^{(2)}}) \right| \left| \tanh(\sigma_v) \operatorname{sech}^2(\sigma_v) \right| \left| \sum_{w \in N(v)} \theta_{wv} (X_{t,w^{(1)}} - \mu_w) \right|^2 + \\ &+ \frac{1}{2n} \sum_{v \in V} \left| \sum_{u \neq v} c_{uv} (X_{t,u^{(1)}} - X_{t,u^{(2)}}) \right| \left| \tanh(\sigma_v) \operatorname{sech}^2(\sigma_v) \right| \left| \sum_{w \in N(v)} \theta_{wv} (X_{t,w^{(2)}} - \mu_w) \right|^2 . \end{aligned}$$

Proof of Lemma 16: For all X_t, X_t':

$$\mathbf{E}\left[f_c(X_{t+1}) - f_c(X'_{t+1}) \,\middle|\, X_t, X'_t\right] =$$

$$= \frac{1}{2n} \sum_{v^{(1)} \in V^{(1)}} \mathbf{E}\left[f_c(X_{t+1}) - f_c(X'_{t+1}) \,\middle|\, X_t, X'_t, \text{node } v^{(1)} \text{ is chosen in step } t+1\right]$$

$$+ \frac{1}{2n} \sum_{v^{(2)} \in V^{(2)}} \mathbf{E}\left[f_c(X_{t+1}) - f_c(X'_{t+1}) \,\middle|\, X_t, X'_t, \text{node } v^{(2)} \text{ is chosen in step } t+1\right]$$

$$= \frac{1}{2n} \sum_{v^{(1)} \in V^{(1)}} \left(f_c(X_t) - \sum_{u \neq v} c_{uv} \left(X_{t,u^{(1)}} - X_{t,u^{(2)}}\right) \left(X_{t,v^{(1)}} - X_{t,v^{(2)}}\right) \right) - \tag{5.25}$$

$$- \frac{1}{2n} \sum_{v^{(1)} \in V^{(1)}} \left(f_c(X'_t) - \sum_{u \neq v} c_{uv} \left(X'_{t,u^{(1)}} - X'_{t,u^{(2)}}\right) \left(X'_{t,v^{(1)}} - X'_{t,v^{(2)}}\right) \right) +$$

$$+ \frac{1}{2n} \sum_{v^{(2)} \in V^{(2)}} \left(f_c(X_t) - \sum_{u \neq v} c_{uv} \left(X_{t,u^{(1)}} - X_{t,u^{(2)}}\right) \left(X_{t,v^{(1)}} - X_{t,v^{(2)}}\right) \right) - \tag{5.26}$$

$$- \frac{1}{2n} \sum_{v^{(2)} \in V^{(2)}} \left(f_c(X'_t) - \sum_{u \neq v} c_{uv} \left(X'_{t,u^{(1)}} - X'_{t,u^{(2)}}\right) \left(X'_{t,v^{(1)}} - X'_{t,v^{(2)}}\right) \right)$$

$$+ \frac{1}{2n} \sum_{v^{(1)} \in V^{(1)}} \sum_{u^{(1)} \neq v^{(1)}} c_{uv} \left(X_{t,u^{(1)}} - X_{t,u^{(2)}}\right) \left(\tanh\left(\theta_v + \sum_{w \in N(v)} \theta_{wv} X_{t,w^{(1)}} \right) - X_{t,v^{(2)}} \right) \tag{5.27}$$

$$- \frac{1}{2n} \sum_{v^{(1)} \in V^{(1)}} \sum_{u^{(1)} \neq v^{(1)}} c_{uv} \left(X'_{t,u^{(1)}} - X'_{t,u^{(2)}}\right) \left(\tanh\left(\theta_v + \sum_{w \in N(v)} \theta_{wv} X'_{t,w^{(1)}} \right) - X'_{t,v^{(2)}} \right) + \tag{5.28}$$

$$+ \frac{1}{2n} \sum_{v^{(2)} \in V^{(2)}} \sum_{u^{(2)} \neq v^{(2)}} c_{uv} \left(X_{t,u^{(1)}} - X_{t,u^{(2)}}\right) \left(X_{t,v^{(1)}} - \tanh\left(\theta_v + \sum_{w \in N(v)} \theta_{wv} X_{t,w^{(2)}} \right) \right) \tag{5.29}$$

$$- \frac{1}{2n} \sum_{v^{(2)} \in V^{(2)}} \sum_{u^{(2)} \neq v^{(2)}} c_{uv} \left(X'_{t,u^{(1)}} - X'_{t,u^{(2)}}\right) \left(X'_{t,v^{(1)}} - \tanh\left(\theta_v + \sum_{w \in N(v)} \theta_{wv} X'_{t,w^{(2)}} \right) \right) \tag{5.30}$$

The above expression on simplification yields the following:

$$\begin{aligned}
&\left(1-\frac{1}{n}\right)\left(f_c(X_t)-f_c(X_t')\right)+ &(5.31)\\
&+\frac{1}{2n}\sum_{v\in V}\sum_{u\neq v}c_{uv}\left(X_{t,u^{(1)}}-X_{t,u^{(2)}}\right)\tanh\left(\theta_v+\sum_{w\in N(v)}\theta_{wv}X_{t,w^{(1)}}\right)\\
&-\frac{1}{2n}\sum_{v\in V}\sum_{u\neq v}c_{uv}\left(X_{t,u^{(1)}}-X_{t,u^{(2)}}\right)\tanh\left(\theta_v+\sum_{w\in N(v)}\theta_{wv}X_{t,w^{(2)}}\right)\\
&-\frac{1}{2n}\sum_{v\in V}\sum_{u\neq v}c_{uv}\left(X'_{t,u^{(1)}}-X'_{t,u^{(2)}}\right)\tanh\left(\theta_v+\sum_{w\in N(v)}\theta_{wv}X'_{t,w^{(1)}}\right)\\
&+\frac{1}{2n}\sum_{v\in V}\sum_{u\neq v}c_{uv}\left(X'_{t,u^{(1)}}-X'_{t,u^{(2)}}\right)\tanh\left(\theta_v+\sum_{w\in N(v)}\theta_{wv}X'_{t,w^{(2)}}\right).
\end{aligned}$$

In the above derivation, we have followed the same strategy as the one in Lemma 14 where we first split $f_c(X_{t+1})-f_c(X'_{t+1})$ into terms which stay untouched when we randomly choose to update nodes v or v' in our coupled dynamics and the terms which do change. Given our generous coupling, the values of $X_{t+1,v}$ and $X'_{t+1,v}$ are set independently from their marginal distributions conditioning on X_t and X'_t respectively, and their expectations are the expressions involving $\tanh(\cdot)$ in Lines (5.27)-(5.30).

Our goal next is to get rid of the tanh's. We will use the following claim which follows from Taylor's theorem:

Claim 3. $\left|\tanh(x+a)-\tanh(a)-\operatorname{sech}^2(a)x\right|\leq\tanh(a)\operatorname{sech}^2(a)|x|^2$ *for all* $x\in\mathbb{R}$.

Note that all the tanh expressions involved in the above derivation have the same expected value $\sigma_v:=\theta_v+\sum_{w\in N(v)}\theta_{wv}\mathbf{E}[X_w]$. We perform a Taylor approximation of the tanhs around σ_v. Using derivation (5.27)-(5.31), and Claim 3 we get that,

$$
\begin{aligned}
&\mathbf{E}\left[f_c(X_{t+1}) - f_c(X'_{t+1}) \mid X_t, X'_t\right] = \left(1 - \frac{1}{n}\right)\left(f_c(X_t) - f_c(X'_t)\right) + \\
&+ \frac{1}{2n}\sum_{v\in V}\sum_{u\neq v} c_{uv}\left(X_{t,u^{(1)}} - X_{t,u^{(2)}}\right)\left(\operatorname{sech}^2(\sigma_v)\sum_{w\in N(v)}\theta_{wv}\left(X_{t,w^{(1)}} - X_{t,w^{(2)}}\right)\right) \\
&- \frac{1}{2n}\sum_{v\in V}\sum_{u\neq v} c_{uv}\left(X'_{t,u^{(1)}} - X'_{t,u^{(2)}}\right)\left(\operatorname{sech}^2(\sigma_v)\sum_{w\in N(v)}\theta_{wv}\left(X'_{t,w^{(1)}} - X'_{t,w^{(2)}}\right)\right) \\
&\pm \frac{1}{2n}\sum_{v\in V}\left|\sum_{u\neq v} c_{uv}(X_{t,u^{(1)}} - X_{t,u^{(2)}})\right|\left|\tanh(\sigma_v)\operatorname{sech}^2(\sigma_v)\right|\left|\sum_{w\in N(v)}\theta_{wv}(X_{t,w^{(1)}} - \mu_w)\right|^2 \\
&\pm \frac{1}{2n}\sum_{v\in V}\left|\sum_{u\neq v} c_{uv}(X_{t,u^{(1)}} - X_{t,u^{(2)}})\right|\left|\tanh(\sigma_v)\operatorname{sech}^2(\sigma_v)\right|\left|\sum_{w\in N(v)}\theta_{wv}(X_{t,w^{(2)}} - \mu_w)\right|^2 \\
&\pm \frac{1}{2n}\sum_{v\in V}\left|\sum_{u\neq v} c_{uv}(X'_{t,u^{(1)}} - X'_{t,u^{(2)}})\right|\left|\tanh(\sigma_v)\operatorname{sech}^2(\sigma_v)\right|\left|\sum_{w\in N(v)}\theta_{wv}(X'_{t,w^{(1)}} - \mu_w)\right|^2 \\
&\pm \frac{1}{2n}\sum_{v\in V}\left|\sum_{u\neq v} c_{uv}(X'_{t,u^{(1)}} - X'_{t,u^{(2)}})\right|\left|\tanh(\sigma_v)\operatorname{sech}^2(\sigma_v)\right|\left|\sum_{w\in N(v)}\theta_{wv}(X'_{t,w^{(2)}} - \mu_w)\right|^2 \\
&= \left(1 - \frac{1}{n}\right)\left(f_c(X_t) - f_c(X'_t)\right) + \frac{1}{n}\left(f_{g(c)}(X_t) - f_{g(c)}(X'_t)\right) \pm e(c, X_t) \pm e(c, X'_t).
\end{aligned}
$$

□

Using Lemma 16, we now establish a multi-step contraction. The terms involving function $e_2^t(\cdot)$ in the statement, encapsulate the error that is being accumulated and needs to be controlled:

Lemma 17. *Consider the same setup as that of Lemma 16. Then for all t and point wise with respect to X_0, X'_0:*

$$
\begin{aligned}
\mathbf{E}\left[f_c(X_t) - f_c(X'_t) \mid X_0, X'_0\right] = &\sum_{\ell=0}^{t}\binom{t}{\ell}\left(1 - \frac{1}{n}\right)^{t-\ell}\left(\frac{1}{n}\right)^{\ell}\cdot\left(f_{g^{\circ\ell}(c)}(X_0) - f_{g^{\circ\ell}(c)}(X'_0)\right) \\
&\pm e_2^t(c, X_0) \pm e_2^t(c, X'_0),
\end{aligned}
$$

where $g^{\circ\ell}(\cdot)$ denotes the ℓ-fold composition of g with itself, and $e_2^t(\cdot)$ is the non-negative *function defined as follows in terms of function* $e(\cdot)$ *of the statement of Lemma 16:*

$$e_2^t(c, X_0) = \sum_{\ell=0}^{t-1} \sum_{q=0}^{t-1-\ell} \binom{t-1-\ell}{q} \left(1 - \frac{1}{n}\right)^{t-1-\ell-q} \left(\frac{1}{n}\right)^q \mathbf{E}\left[e(g^{\circ q}(c), X_\ell)|X_0\right].$$

The proof of Lemma 17 uses induction and follows along similar lines to that of Lemma 15, hence it is skipped here.

We are now ready to bound the variance of $f_c(\cdot)$ and prove Theorem 10:

Proof of Theorem 10: (5.5) and (5.4) give

$$\begin{aligned}\mathbf{Var}\left(f_c(X)\right) &= \frac{1}{2} \cdot \mathbf{E}\left[(f_c(X) - f_c(X')) \cdot F(X, X')\right] \\ &= \frac{1}{2} \sum_{t=0}^{\infty} \mathbf{E}\left[(f_c(X_0) - f_c(X_0')) \cdot \mathbf{E}\left[f_c(X_t) - f_c(X_t')|X_0, X_0'\right]\right]. \qquad (5.32)\end{aligned}$$

Using the same argument as in the proof of Theorem 9 it follows that for large enough $t^* = \Omega(n \log n)$:

$$\frac{1}{2} \sum_{t=t^*}^{\infty} \mathbf{E}\left[(f_c(X_0) - f_c(X_0')) \cdot \mathbf{E}\left[f_c(X_t) - f_c(X_t')|X_0, X_0'\right]\right] \leq 1. \qquad (5.33)$$

The above shows that we only need to bound (5.32) for t ranging from 0 to some $t^* = O(n \log n)$. It also shows that Condition 5.3, required for our anti-symmetric function $F()$ to be well-defined, holds.

To bound (5.32) for t ranging from 0 to $t^* = O(n \log n)$, let us take one of the terms, and plug in the bound from Lemma 17. Given that the bound of the lemma holds point-wise and $e_2()$ is non-negative we have:

$$\begin{aligned}&\mathbf{E}\left[(f_c(X_0) - f_c(X_0')) \cdot \mathbf{E}\left[f_c(X_t) - f_c(X_t')|X_0, X_0'\right]\right] \leq \\ &\quad \sum_{\ell=0}^{t} \binom{t}{\ell} \left(1 - \frac{1}{n}\right)^{t-\ell} \left(\frac{1}{n}\right)^{\ell} \cdot \mathbf{E}\left[\left|f_c(X_0) - f_c(X_0')\right| \left|f_{g^{\circ \ell}(c)}(X_0) - f_{g^{\circ \ell}(c)}(X_0')\right|\right] \qquad (5.34) \\ &\qquad + \mathbf{E}\left[\left|f_c(X_0) - f_c(X_0')\right| e_2^t(c, X_0)\right] + \mathbf{E}\left[\left|f_c(X_0) - f_c(X_0')\right| e_2^t(c, X_0')\right]. \qquad (5.35)\end{aligned}$$

Now, recall that the pair (X_0, X_0') is sampled as follows: X_0 is a sample from the Ising model, and X_0' is one step of the Glauber dynamics from X_0. So:

$$\left|f_c(X_0) - f_c(X_0')\right| \leq 2\max_v \left|\sum_{u\neq v} c_{uv}(X_{0,u^{(1)}} - X_{0,u^{(2)}})\right|.$$

Since $\mathbf{E}\left[\sum_{u\neq v} c_{uv}(X_{0,u^{(1)}} - X_{0,u^{(2)}})\right] = 0$, it follows from Lemma 25 that, for all v, a sample X_0 from $p^{\otimes 2}$ satisfies:

$$\Pr\left[\left|\sum_{u\neq v} c_{uv}(X_{0,u^{(1)}} - X_{0,u^{(2)}})\right| \geq t\right] \leq 2e^{-\frac{(1-\eta)t^2}{8\sum_{u\neq v} c_{uv}^2}},$$

where η is the constant from Definition 2. So for sufficiently large $t = \Omega(\sqrt{\log n}\cdot |c_{\cdot v}|_2)$, with probability at least $1-\frac{1}{8n^3}$: $\left|\sum_{u\neq v} c_{uv}(X_{0,u^{(1)}} - X_{0,u^{(2)}})\right| < t$. It follows that, with probability at least $1-1/8n^2$, $\max_v \left|\sum_{u\neq v} c_{uv}(X_{0,u^{(1)}} - X_{0,u^{(2)}})\right| = O(\sqrt{\log n}\cdot \max_v |c_{\cdot v}|_2)$. Hence, with probability at least $1-1/8n^2$:

$$\left|f_c(X_0) - f_c(X_0')\right| \leq O(\sqrt{\log n}\cdot \max_v |c_{\cdot v}|_2).$$

Notice that in the regime of Definition 2, function g maps points in $[-1,1]^{\binom{V}{2}}$ to the same set. Hence by a similar token, for any fixed ℓ, with probability at least $1-1/8n^2$:

$$\left|f_{g^{\circ\ell}(c)}(X_0) - f_{g^{\circ\ell}(c)}(X_0')\right| \leq O(\sqrt{\log n}\cdot \max_v \left|g^{\circ\ell}(c)_{\cdot v}\right|_2).$$

At the same time, the maximum that $2\max_v \left|\sum_{u\neq v} c_{uv}(X_{0,u^{(1)}} - X_{0,u^{(2)}})\right|$ (and hence $|f_c(X_0) - f_c(X_0')|$) can possibly be is $4\max_v |c_{\cdot v}|_1 \leq 4n$. Similarly, the maximum that $\left|f_{g^{\circ\ell}(c)}(X_0) - f_{g^{\circ\ell}(c)}(X_0')\right|$ can possibly be is also at most $4n$, for any ℓ.

It follows from the above calculations that:

$$\begin{aligned}\mathbf{E}\left[\left|f_c(X_0) - f_c(X_0')\right|\left|f_{g^{\circ\ell}(c)}(X_0) - f_{g^{\circ\ell}(c)}(X_0')\right|\right] &\leq O\left(\log n\cdot \max_v |c_{\cdot v}|_2\cdot \max_v \left|g^{\circ\ell}(c)_{\cdot v}\right|_2\right)\\ &\leq O(\sqrt{n}\log n\cdot \max_v |c_{\cdot v}|_2) \qquad (5.36)\end{aligned}$$

Given that this bound holds for any ℓ, and recognizing the binomial expansion in (5.34), we obtain the bound:

$$(5.34) \leq O(\sqrt{n}\log n\cdot \max_v |c_{\cdot v}|_2).$$

It remains to bound the error terms (5.35). For a fixed ℓ and q let us try to bound the term $\mathbf{E}\left[e(g^{\circ q}(c), X_\ell) | X_0\right]$, involved in the definition of $e_2^t(c, X_0)$. For convenience set $c' = g^{\circ q}(c)$, and recall (as we have pointed out above) that $c' \in [-1,1]^{\binom{V}{2}}$. Recalling the definition of $e()$ from the statement of Lemma 16, we have that:

$$\begin{aligned}
&\mathbf{E}\left[e(c', X_\ell) | X_0\right] = \\
&= \mathbf{E}\left[\sum_v \frac{\left|\tanh(\sigma_v)\,\mathrm{sech}^2(\sigma_v)\right|}{2n}\left|\sum_{u\neq v} c'_{uv}(X_{\ell,u^{(1)}} - X_{\ell,u^{(2)}})\right|\left|\sum_{w\in N(v)} \theta_{wv}(X_{\ell,w^{(1)}} - \sigma_v)\right|^2 \middle| X_0\right] \\
&+ \mathbf{E}\left[\sum_v \frac{\left|\tanh(\sigma_v)\,\mathrm{sech}^2(\sigma_v)\right|}{2n}\left|\sum_{u\neq v} c'_{uv}(X_{\ell,u^{(1)}} - X_{\ell,u^{(2)}})\right|\left|\sum_{w\in N(v)} \theta_{wv}(X_{\ell,w^{(2)}} - \sigma_v)\right|^2 \middle| X_0\right].
\end{aligned}$$

Given that X_0 is sampled from the Ising model, and X_ℓ is the state reached after ℓ steps of the Glauber dynamics from X_0, it follows that X_ℓ is also a sample from the Ising model. So a similar analysis as the one we did earlier implies that for a fixed v, with probability at least $1 - \frac{1}{2n^{21}}$: $\left|\sum_{u\neq v} c'_{uv}(X_{\ell,u^{(1)}} - X_{\ell,u^{(2)}})\right| < O(\sqrt{\log n}\cdot |c'_{\cdot v}|_2)$. So, with probability at least $1 - \frac{1}{2n^{20}}$, simultaneously for all v:

$$\left|\sum_{u\neq v} c'_{uv}(X_{\ell,u^{(1)}} - X_{\ell,u^{(2)}})\right| \leq O(\sqrt{\log n}\cdot\sqrt{n}).$$

Via similar arguments, it can be shown that, with probability at least $1 - \frac{1}{4n^{20}}$, simultaneously for all $v^{(1)} \in V^{(1)}$:

$$\left|\sum_{w^{(1)}\in N(v^{(1)})} \theta_{wv}(X_{\ell,w^{(1)}} - \mu_w)\right| \leq O\left(\sqrt{\log n \cdot d_{\max}}\cdot\beta\right),$$

and for all $v^{(2)} \in V^{(2)}$:

$$\left|\sum_{w^{(2)}\in N(v^{(2)})} \theta_{wv}(X_{\ell,w^{(2)}} - \mu_w)\right| \leq O\left(\sqrt{\log n \cdot d_{\max}}\cdot\beta\right).$$

So it follows from the above that, with probability at least $1 - 1/n^{20}$, it holds that:

$$\frac{1}{2n}\sum_v \left|\tanh(\sigma_v)\operatorname{sech}^2(\sigma_v)\right| \left|\sum_{u \neq v} c'_{uv}(X_{\ell,u^{(1)}} - X_{\ell,u^{(2)}})\right| \left|\sum_{w \in N(v)} \theta_{wv}(X_{\ell,w^{(1)}} - \mu_w)\right|^2$$
$$\leq O\left(\sqrt{n}\log^2 n \cdot d_{\max} \cdot \beta^2\right)$$
$$, \frac{1}{2n}\sum_v \left|\tanh(\sigma_v)\operatorname{sech}^2(\sigma_v)\right| \left|\sum_{u \neq v} c'_{uv}(X_{\ell,u^{(1)}} - X_{\ell,u^{(2)}})\right| \left|\sum_{w \in N(v)} \theta_{wv}(X_{\ell,w^{(2)}} - \mu_w)\right|^2$$
$$\leq O\left(\sqrt{n}\log^2 n \cdot d_{\max} \cdot \beta^2\right)$$

Let us call the event that the above two statements hold $\mathcal{E}$. We want to view this event as a function $\mathcal{E} = \mathcal{E}(X_0, G_\ell)$ of X_0 and the decisions G_ℓ that the Glauber dynamics made in the first ℓ steps. Indeed, we want to view X_0 and G_ℓ as independent random variables. G_ℓ samples independently of X_0 which nodes it will update, together with ℓ uniform $[0, 1]$ random variables. Then the Glauber dynamics are a deterministic function of X_0 and G_ℓ. With this perspective in mind, we have from the above that:

$$\Pr_{X_0, G_\ell}[\mathcal{E}(X_0, G_\ell)] \geq 1 - \frac{1}{n^{20}}.$$

From this it follows that

$$\Pr_{X_0}\left[\Pr_{G_\ell}[\mathcal{E}(X_0, G_\ell)] \geq 1 - 1/n^9\right] \geq 1 - 1/n^9.$$

In turn, the above implies that

$$\Pr_{X_0}\left[\mathbf{E}\left[e(c', X_\ell) \middle| X_0\right] \leq O\left(\sqrt{n}\log^2 n \cdot d_{\max} \cdot \beta^2\right)\right] \geq 1 - 1/n^9. \tag{5.37}$$

From similar analysis to the one we did earlier we also have:

$$\Pr_{X_0}\left[\left|f_c(X_0) - f_c(X'_0)\right| \leq O(\sqrt{\log n} \cdot \max_v |c_{\cdot v}|_2)\right] \geq 1 - 1/n^9. \tag{5.38}$$

So (5.37) and (5.38) imply:

$$\mathbf{E}\left[\left|f_c(X_0) - f_c(X'_0)\right| \mathbf{E}\left[e(c', X_\ell) \middle| X_0\right]\right] \leq O\left(\sqrt{n}\log^{2.5} n \cdot \max_v |c_{\cdot v}|_2 \cdot d_{\max} \cdot \beta^2\right). \tag{5.39}$$

Now the definition of function $e_2^t(\cdot)$ in the statement of Lemma 17 and (5.39) imply that:

$$\mathbf{E}\left[\left|f_c(X_0) - f_c(X_0')\right| e_2^t(c, X_0)\right] \leq O\left(t \cdot \sqrt{n} \log^{2.5} n \cdot \max_v |c_{\cdot v}|_2 \cdot d_{\max} \cdot \beta^2\right).$$

The same bound applies to $\mathbf{E}\left[|f_c(X_0) - f_c(X_0')| \, e_2^t(c, X_0')\right]$. So we have successfully bounded (5.35).

Using our bounds for (5.34) and (5.35), we get that:

$$\mathbf{E}\left[(f_c(X_0) - f_c(X_0')) \cdot \mathbf{E}\left[f_c(X_t) - f_c(X_t') \middle| X_0, X_0'\right]\right] \tag{5.40}$$
$$\leq O(\sqrt{n} \log n \cdot \max_v |c_{\cdot v}|_2) + O\left(t \cdot \sqrt{n} \log^{2.5} n \cdot \max_v |c_{\cdot v}|_2 \cdot d_{\max} \cdot \beta^2\right). \tag{5.41}$$

So we can go back to (5.32) to bound the first t^* terms of the summation, for $t^* = O(n \log n)$ as set earlier. We get:

$$\frac{1}{2} \sum_{t=0}^{t^*} \mathbf{E}\left[(f_c(X_0) - f_c(X_0')) \cdot \mathbf{E}\left[f_c(X_t) - f_c(X_t') \middle| X_0, X_0'\right]\right] =$$
$$= O(n^{1.5} \log^2 n \cdot \max_v |c_{\cdot v}|_2) + O\left(n^{2.5} \log^{4.5} n \cdot \max_v |c_{\cdot v}|_2 \cdot d_{\max} \cdot \beta^2\right) \tag{5.42}$$

Plugging (5.42) and (5.33) into (5.32), we bound the variance as follows:

$$\mathbf{Var}\left(f_c(X)\right) = \tilde{O}(n^{1.5} \cdot \max_v |c_{\cdot v}|_2) + O\left(n^{2.5} \cdot \max_v |c_{\cdot v}|_2 \cdot d_{\max} \cdot \beta^2\right).$$

□

Chapter 6

Comparing Localization and Learn-then-Test Algorithms

At this point, we now have two algorithms: the localization algorithm of Chapter 3 and the learn-then-test algorithm of Chapter 4. We note that their sample complexities differ in their dependence on β and $d_{\max}$. In this chapter, we offer some intuition as to why the difference arises and state the best sample complexities we achieve for our testing problems by combining these two approaches.

First, the localization algorithm gets worse as $d_{\max}$ increases. As noted in Chapter 3, the reason for this worsening is that the contribution to the distance by any single edge grows smaller thereby making it harder to detect. However, when we are in the high-temperature regime a larger $d_{\max}$ implies a tighter bound on the strength of the edge interactions β and the variance bound of Chapter 5 exploits this tighter bound to get savings in sample complexities when the degree is large enough.

We combine the sample complexities obtained by the localization and the learn-then-test algorithms and summarize in the following theorems the best sample complexities we can achieve for testing independence and identity by noting the parameter regimes in which of the above two algorithms gives better sample complexity. In both of the following theorems we fix β to be $n^{-\alpha}$ for some α and present which algorithm dominates as $d_{\max}$ ranges from a constant to n.

Theorem 11 (Best Sample Complexity Achieved, No External Field)**.** *Suppose p is an Ising model under **no external field**.*

- *if $\beta = O(n^{-2/3})$, then for the range $d_{\max} \le n^{2/3}$, localization performs better, for both* *independence and identity testing. For the range $n^{2/3} \leq d_{\max} \leq \frac{1}{4\beta}$, learn-then-test performs better than localization for both independence and identity testing yielding a sample complexity which is independent of $d_{\max}$. If $d_{\max} \geq \frac{1}{4\beta}$, then we are no longer in the high temperature regime.*

- *if $\beta = \omega(n^{-2/3})$, then for the entire range of $d_{\max}$ localization performs at least as well as the learn-then-test algorithm for both independence and identity testing.*

The theorem stated above is summarized in Figure 6-2 for the regime when $\beta = O(n^{-2/3})$.

The comparison for independence testing under the presence of an external field is a bit more complex and is presented in Theorem 12.

Theorem 12 (Best Sample Complexity Achieved for Independence Testing, Arbitrary External Field)**.** *Suppose p is an Ising model under* ***an arbitrary external field****.*

- *if $\beta^2 d_{\max} = O(1/n)$ and $\beta = O(n^{-5/6})$, then for $d_{\max} = \Omega(n^{2/3})$ learn-then-test performs better than localization.*
- *if $\beta^2 d_{\max} = \omega(1/n)$ and $\beta^{-1} d_{\max}^{5/2} = \Omega(n^{5/2})$, learn-then-test performs better than localization.*
- *In all other regimes, localization performs at least as well as learn-then-test.*

Finally, we note in Theorem 13, the parameter regimes when learn-then-test performs better for identity testing under an external field.

Theorem 13 (Best Sample Complexity Achieved for Identity Testing, Arbitrary External Field)**.** *Suppose p is an Ising model under* ***an arbitrary external field****.*

- *if $\beta = O(n^{-5/6})$, then for the range $n^{2/3} \leq d_{\max} \leq \frac{1}{4\beta}$, learn-then-test performs better than localization for identity testing yielding a sample complexity which is independent of $d_{\max}$. If $d_{\max} \geq \frac{1}{4\beta}$, then we are no longer in the high temperature regime.*
- *if $\beta = \omega(n^{-5/6})$, then for the entire range of $d_{\max}$ localization performs at least as well as the learn-then-test algorithm for identity.*

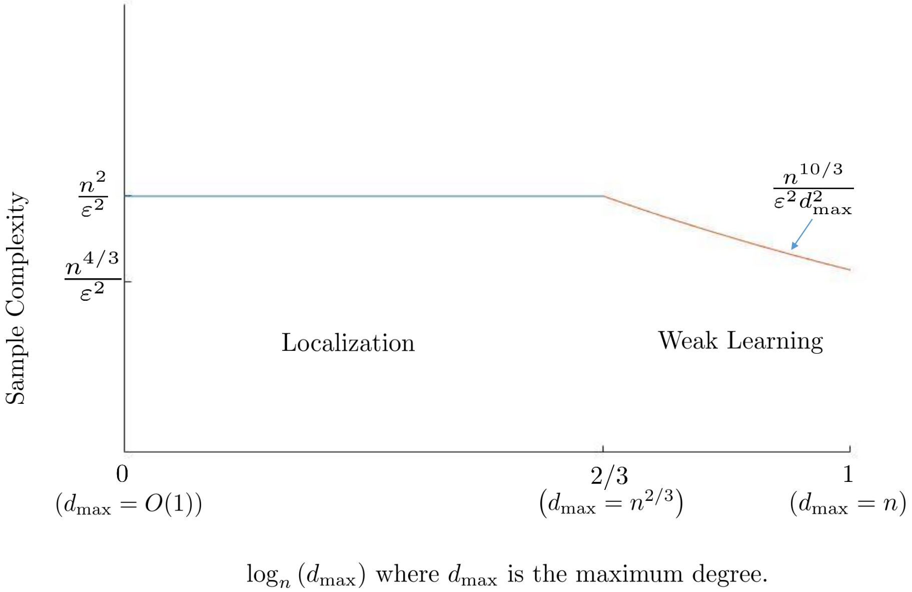

Figure 6-1: Localization vs Learn-Then-Test: A plot of the sample complexity of testing identity under no external field when $\beta = \frac{1}{4d_{\max}}$ is close to the threshold of high temperature. Note that throughout the range of values of $d_{\max}$ we are in high temperature regime in this plot.

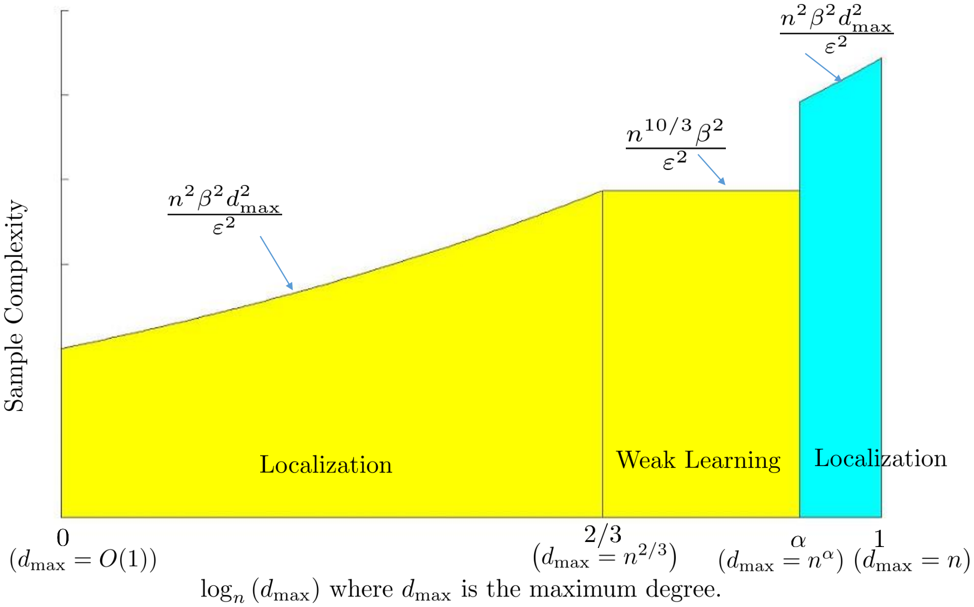

Figure 6-2: Localization vs Learn-Then-Test: A plot of the sample complexity of testing identity under no external field when $\beta \leq n^{-2/3}$. The regions shaded yellow denote the high temperature regime while the region shaded blue denotes the low temperature regime. The algorithm which achieves the better sample complexity is marked on the corresponding region.

Chapter 7

Lower Bounds

In this chapter we describe lower bound constructions for the testing problems studied in this thesis and state the main results.

7.1 Dependences on n

Our first lower bounds show dependences on n, the number of nodes, in the complexity of testing Ising models.

To start, we prove that uniformity testing on product measures over a binary alphabet requires $\Omega(\sqrt{n}/\varepsilon)$ samples. Note that a binary product measure corresponds to the case of an Ising model with no edges. This implies the same lower bound for identity testing, but *(*not) independence testing, as a product measure always has independent marginals, so the answer is trivial.

Theorem 14. *There exists a constant $c > 0$ such that any algorithm, given sample access to an Ising model p with no edges (i.e., a product measure over a binary alphabet), which distinguishes between the cases $p = \mathcal{U}_n$ and $d_{\mathrm{SKL}}(p, \mathcal{U}_n) \geq \varepsilon$ with probability at least $99/100$ requires $k \geq c\sqrt{n}/\varepsilon$ samples.*

Next, we show that any algorithm which tests uniformity of an Ising model requires $\Omega(n/\varepsilon)$ samples. In this case, it implies the same lower bounds for independence and identity testing.

Theorem 15. *There exists a constant $c > 0$ such that any algorithm, given sample access to an Ising model p, which distinguishes between the cases $p = \mathcal{U}_n$ and $d_{\mathrm{SKL}}(p, \mathcal{U}_n) \geq \varepsilon$ with probability at least $99/100$ requires $k \geq cn/\varepsilon$ samples. This remains the case even if p is known to have a tree structure and only ferromagnetic edges.*

The lower bounds use Le Cam's two point method which constructs a family of distributions $\mathcal{P}$ such that the distance between any $P \in \mathcal{P}$ and a particular distribution Q is large (at least ε). But given a $P \in \mathcal{P}$ chosen uniformly at random, it is hard to distinguish between P and Q with at least $2/3$ success probability unless we have sufficiently many samples.

Our construction for product measures is inspired by Paninski's lower bound for uniformity testing [15]. We start with the uniform Ising model and perturb each node positively or negatively by $\sqrt{\varepsilon/n}$, resulting in a model which is ε-far in d_{SKL} from $\mathcal{U}_n$. The proof appears in Section 7.3.1.

Our construction for the linear lower bound builds upon this style of perturbation. In the previous construction, instead of perturbing the node potentials, we could have left the node marginals to be uniform and perturbed the edges of some fixed, known matching to obtain the same lower bound. To get a linear lower bound, we instead choose a *random* matching, which turns out to require quadratically more samples to test. Interestingly, we only need ferromagnetic edges (i.e., positive perturbations), as the randomness in the choice of matching is sufficient to make the problem harder. Our proof is significantly more complicated for this case, and it uses a careful combinatorial analysis involving graphs which are unions of two perfect matchings. The lower bound is described in detail in Section 7.3.2.

Remark 2. *Similar lower bound constructions to those of Theorems 14 and 15 also yield* $\Omega(\sqrt{n}/\varepsilon^2)$ *and* $\Omega(n/\varepsilon^2)$ *for the corresponding testing problems when* d_{SKL} *is replaced with* d_{TV}*. In our constructions, we describe families of distributions which are* ε*-far in* d_{SKL}*. This is done by perturbing certain parameters by a magnitude of* $\Theta(\sqrt{\varepsilon/n})$*. We can instead describe families of distributions which are* ε*-far in* d_{TV} *by performing perturbations of* $\Theta(\varepsilon/\sqrt{n})$*, and the rest of the proofs follow similarly.*

7.2 Dependences on h, β

Finally, we show that dependences on the h and β parameters are, in general, necessary for independence and identity testing. Recall that h and β are upper bounds on the absolute values of the node and edge parameters, respectively. Our constructions are fairly simple, involving just one or two nodes, and the results are stated in Theorem 16.

Theorem 16. *There is a linear lower bound on the parameters* h *and* β *for testing problems on Ising models. More specifically,*

- *There exists a constant $c > 0$ such that, for all $\varepsilon < 1$ and $\beta \geq 0$, any algorithm, given sample access to an Ising model p, which distinguishes between the cases $p \in \mathcal{I}_n$ and $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$ with probability at least $99/100$ requires $k \geq c\beta/\varepsilon$ samples.*

- *There exists constants $c_1, c_2 > 0$ such that, for all $\varepsilon < 1$ and $\beta \geq c_1 \log(1/\varepsilon)$, any algorithm, given a description of an Ising model q with no external field (i.e., $h = 0$) and has sample access to an Ising model p, and which distinguishes between the cases $p = q$ and $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$ with probability at least $99/100$ requires $k \geq c_2\beta/\varepsilon$ samples.*

- *There exists constants $c_1, c_2 > 0$ such that, for all $\varepsilon < 1$ and $h \geq c_1 \log(1/\varepsilon)$, any algorithm, given a description of an Ising model q with no edge potentials(i.e., $\beta = 0$) and has sample access to an Ising model p, and which distinguishes between the cases $p = q$ and $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$ with probability at least $99/100$ requires $k \geq c_2 h/\varepsilon$ samples.*

The construction and analysis appears in Section 7.3.3.

This lower bound shows that the dependence on β parameters by our algorithms cannot be avoided in general, though it may be sidestepped in certain cases.

7.3 Lower Bound Proofs

7.3.1 Proof of Theorem 14

This proof will follow via an application of Le Cam's two-point method. More specifically, we will consider two classes of distributions $\mathcal{P}$ and $\mathcal{Q}$ such that:

1. $\mathcal{P}$ consists of a single distribution $p \triangleq \mathcal{U}_n$;

2. $\mathcal{Q}$ consists of a family of distributions such that for all distributions $q \in \mathcal{Q}$, $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$;

3. There exists some constant $c > 0$ such that any algorithm which distinguishes p from a uniformly random distribution $q \in \mathcal{Q}$ with probability $\geq 2/3$ requires $\geq c\sqrt{n}/\varepsilon$ samples.

The third point will be proven by showing that, with $k < c\sqrt{n}/\varepsilon$ samples, the following two processes have miniscule total variation distance, and thus no algorithm can distinguish them:

- The process $p^{\otimes k}$, which draws k samples from p;

- The process $\bar{q}^{\otimes k}$, which selects q from $\mathcal{Q}$ uniformly at random, and then draws k samples from q.

We will let $p_i^{\otimes k}$ be the process $p^{\otimes k}$ restricted to the ith coordinate of the random vectors sampled, and $\bar{q}_i^{\otimes k}$ is defined similarly.

We proceed with a description of our construction. Let $\delta = \sqrt{3\varepsilon/2n}$. As mentioned before, $\mathcal{P}$ consists of the single distribution $p \triangleq \mathcal{U}_n$, the Ising model on n nodes with 0 potentials on every node and edge. Let $\mathcal{M}$ be the set of all 2^n vectors in the set $\{\pm\delta\}^n$. For each $M \in \mathcal{M}$, we define a corresponding $q_M \in \mathcal{Q}$ where the node potential M_i is placed on node i.

Proposition 1. *For each $q \in \mathcal{Q}$, $d_{\mathrm{SKL}}(q,\mathcal{U}_n) \geq \varepsilon$.*

Proof. Recall that

$$d_{\mathrm{SKL}}(q,\mathcal{U}_n) = \sum_{v\in V} \delta \tanh(\delta).$$

Note that $\tanh(\delta) \geq 2\delta/3$ for all $\delta \leq 1$, which can be shown using a Taylor expansion. Therefore

$$d_{\mathrm{SKL}}(q,\mathcal{U}_n) \geq n \cdot \delta \cdot 2\delta/3 = 2n\delta^2/3 = \varepsilon.$$

□

The goal is to upper bound $d_{\mathrm{TV}}(p^{\otimes k}, \bar{q}^{\otimes k})$. We will use the following lemma from [46], which follows from Pinsker's and Jensen's inequalities:

Lemma 18. *For any two distributions p and q,*

$$2d_{\mathrm{TV}}^2(p,q) \leq \log \mathbf{E}_q \left[\frac{q}{p}\right].$$

Applying this lemma, the fact that $\mathcal{Q}$ is a family of product distributions, and that we can picture $\bar{q}^{\otimes k}$ as the process which picks a $q \in \mathcal{Q}$ by selecting a parameter for each node in an iid manner, we have that

$$2d_{\mathrm{TV}}^2(p^{\otimes k}, \bar{q}^{\otimes k}) \leq n \log \mathbf{E}_{\bar{q}_1^{\otimes k}} \left[\frac{\bar{q}_1^{\otimes k}}{p_1^{\otimes k}}\right].$$

We proceed to bound the right-hand side. To simplify notation, let $p_+ = e^{\delta}/(e^{\delta} + e^{-\delta})$ be the probability that a node with parameter δ takes the value 1. Note that a node with parameter $-\delta$ takes the value 1 with probability $1 - p_+$. We will perform a sum over all realizations k_1 for the number of times that node 1 is observed to be 1.

$$
\begin{aligned}
\mathbf{E}_{\bar{q}_1^{\otimes k}}\left[\frac{\bar{q}_1^{\otimes k}}{p_1^{\otimes k}}\right] &= \sum_{k_1=0}^{k} \frac{(\bar{q}_1^{\otimes k}(k_1))^2}{p_1^{\otimes k}(k_1)} \\
= \quad & \sum_{k_1=0}^{k} \frac{\left(\frac{1}{2}\binom{k}{k_1}(p_+)^{k_1}(1-p_+)^{k-k_1} + \frac{1}{2}\binom{k}{k-k_1}(p_+)^{k_1}(1-p_+)^{k_1}\right)^2}{\binom{k}{k_1}(1/2)^k} \\
= \quad & \frac{2^k}{4}\sum_{k_1=0}^{k}\binom{k}{k_1}\left((p_+)^{2k_1}(1-p_+)^{2(k-k_1)} + (p_+)^{2(k-k_1)}(1-p_+)^{2k_1} + 2(p_+(1-p_+))^k\right) \\
= \quad & \frac{2^k}{2}(p_+(1-p_+))^k \sum_{k_1=0}^{k}\binom{k}{k_1} + 2\cdot\frac{2^k}{4}\sum_{k_1=0}^{k}\left(\binom{k}{k_1}(p_+^2)^{k_1}((1-p_+)^2)^{k-k_1}\right)
\end{aligned}
$$

where the second equality uses the fact that $\bar{q}_1^{\otimes k}$ chooses the Ising model with parameter on node 1 being δ and $-\delta$ each with probability 1/2. Using the identity $\sum_{k_1=0}^{k}\binom{k}{k_1}a^{k_1}b^{k-k_1} = (a+b)^k$ gives that

$$
\mathbf{E}_{\bar{q}_1^{\otimes k}}\left[\frac{\bar{q}_1^{\otimes k}}{p_1^{\otimes k}}\right] = \frac{4^k}{2}(p_+(1-p_+))^k + \frac{2^k}{2}\left(2p_+^2 + 1 - 2p_+\right)^k.
$$

Substituting in the value for p_+ and applying hyperbolic trigenometric identities, the above expression simplifies to

$$
\begin{aligned}
& \frac{1}{2}\left(\left(\operatorname{sech}^2(\delta)\right)^k + \left(1+\tanh^2(\delta)\right)^k\right) \\
\leq \quad & 1 + \binom{k}{2}\delta^4 \\
= \quad & 1 + \binom{k}{2}\frac{9\varepsilon^2}{4n^2}
\end{aligned}
$$

where the inequality follows by a Taylor expansion.

This gives us that

$$
2d_{\mathrm{TV}}^2(p^{\otimes k}, \bar{q}^{\otimes k}) \leq n\log\left(1 + \binom{k}{2}\frac{9\varepsilon^2}{4n^2}\right) \leq \frac{9k^2\varepsilon^2}{4n}.
$$

If $k < 0.9\cdot\sqrt{n}/\varepsilon$, then $d_{\mathrm{TV}}^2(p^{\otimes k}, \bar{q}^{\otimes k}) < 49/50$ and thus no algorithm can distinguish between the two with probability $\geq 99/100$. This completes the proof of Theorem 14.

7.3.2 Proof of Theorem 15

This lower bound similarly applies Le Cam's two-point method, as described in the previous section. We proceed with a description of our construction. Assume that n is even. As before, $\mathcal{P}$ consists of the single distribution $p \triangleq \mathcal{U}_n$, the Ising model on n nodes with 0 potentials on every node and edge. Let $\mathcal{M}$ denote the set of all $(n-1)!!$ perfect matchings on the clique on n nodes. Each $M \in \mathcal{M}$ defines a corresponding $q_M \in \mathcal{Q}$, where the potential $\delta = \sqrt{3\varepsilon/n}$ is placed on each edge present in the graph.

The follow proposition follows similarly to Proposition 1.

Proposition 2. *For each $q \in \mathcal{Q}$, $d_{\mathrm{SKL}}(q,\mathcal{U}_n) \geq \varepsilon$.*

The goal is to upper bound $d_{\mathrm{TV}}(p^{\otimes k}, \bar{q}^{\otimes k})$. We again apply Lemma 18 to $2d^2_{\mathrm{TV}}(p^{\otimes k}, \bar{q}^{\otimes k})$ and focus on the quantity inside the logarithm. Let $X^{(i)} \in \{\pm 1\}^n$ represent the realization of the ith sample and $X_u \in \{\pm 1\}^k$ represent the realization of the k samples on node u. Let $H(.,.)$ represent the Hamming distance between two vectors, and for sets S_1 and S_2, let $S = S_1 \uplus S_2$ be the very commonly used multiset addition operation. Let M_0 be the matching with edges $(2i-1, 2i)$ for all $i \in [n/2]$.

$$\begin{aligned}\mathbf{E}_{\bar{q}^{\otimes k}}\left[\frac{\bar{q}^{\otimes k}}{p^{\otimes k}}\right] &= \sum_{X=(X^{(1)},\ldots,X^{(k)})} \frac{(\bar{q}^{\otimes k}(X))^2}{p^{\otimes k}(X)} \\ &= 2^{nk} \sum_{X=(X^{(1)},\ldots,X^{(k)})} (\bar{q}^{\otimes k}(X))^2\end{aligned}$$

We can expand the inner probability as follows. Given a randomly selected matching, we can break the probability of a realization X into a product over the edges. By examining the PMF of the Ising model, if the two endpoints of a given edge agree, the probability is multiplied by a factor of $\left(\frac{e^\delta}{2(e^\delta+e^{-\delta})}\right)$, and if they disagree, a factor of $\left(\frac{e^{-\delta}}{2(e^\delta+e^{-\delta})}\right)$. Since (given a matching) the samples are independent, we take the product of this over all k samples. We average this quantity using a uniformly random choice of matching. Writing these ideas mathematically, the expression above is equal to

$$\begin{aligned}
&2^{nk}\sum_X \left(\frac{1}{(n-1)!!}\sum_{M\in\mathcal{M}}\prod_{(u,v)\in M}\prod_{i=1}^{k}\left(\frac{e^{\delta}}{2(e^{\delta}+e^{-\delta})}\right)^{\mathbb{1}(X_u^{(i)}=X_v^{(i)})}\left(\frac{e^{-\delta}}{2(e^{\delta}+e^{-\delta})}\right)^{\mathbb{1}(X_u^{(i)}\neq X_v^{(i)})}\right)^2\\
&= 2^{nk}\sum_{X=(X^{(1)},\ldots,X^{(k)})}\left(\frac{1}{(n-1)!!}\sum_{M\in\mathcal{M}}\prod_{(u,v)\in M}\left(\frac{1}{2(e^{\delta}+e^{-\delta})}\right)^{k} e^{\delta(k-H(X_u,X_v))}e^{-\delta H(X_u,X_v)}\right)^2\\
&= \left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}}\right)^{nk}\sum_{X=(X^{(1)},\ldots,X^{(k)})}\left(\frac{1}{(n-1)!!}\sum_{M\in\mathcal{M}}\prod_{(u,v)\in M}\exp(-2\delta H(X_u,X_v))\right)^2\\
&= \left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}}\right)^{nk}\frac{1}{(n-1)!!^2}\sum_{X=(X^{(1)},\ldots,X^{(k)})}\left(\sum_{M\in\mathcal{M}}\prod_{(u,v)\in M}\exp(-2\delta H(X_u,X_v))\right)^2\\
&= \left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}}\right)^{nk}\frac{1}{(n-1)!!^2}\sum_{X=(X^{(1)},\ldots,X^{(k)})}\sum_{M_1,M_2\in\mathcal{M}}\prod_{(u,v)\in M_1\uplus M_2}\exp(-2\delta H(X_u,X_v))
\end{aligned}$$

At this point, we note that if we fix matching the matching M_1, summing over all matchings M_2 gives the same value irrespective of the value of M_1. Therefore, we multiply by a factor of $(n-1)!!$ and fix the choice of M_1 to be M_0.

$$\begin{aligned}
&\left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}}\right)^{nk}\frac{1}{(n-1)!!}\sum_{M\in\mathcal{M}}\sum_{X=(X^{(1)},\ldots,X^{(k)})}\prod_{(u,v)\in M_0\uplus M}\exp(-2\delta H(X_u,X_v))\\
&= \left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}}\right)^{nk}\frac{1}{(n-1)!!}\sum_{M\in\mathcal{M}}\left(\sum_{X^{(1)}}\prod_{(u,v)\in M_0\uplus M}\exp\left(-2\delta H\left(X_u^{(1)},X_v^{(1)}\right)\right)\right)^k
\end{aligned}$$

We observe that multiset union of two matchings will form a collection of even length cycles, and this can be rewritten as follows.

$$\begin{aligned}
&\left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}}\right)^{nk}\frac{1}{(n-1)!!}\sum_{M\in\mathcal{M}}\left(\sum_{X^{(1)}}\prod_{\substack{\text{cycles} C\\ \in M_0\uplus M}}\prod_{(u,v)\in C}\exp\left(-2\delta H\left(X_u^{(1)},X_v^{(1)}\right)\right)\right)^k\\
&= \left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}}\right)^{nk}\frac{1}{(n-1)!!}\sum_{M\in\mathcal{M}}\left(\prod_{\substack{\text{cycles } C\\ \in M_0\uplus M}}\sum_{X_C^{(1)}}\prod_{(u,v)\in C}\exp\left(-2\delta H\left(X_u^{(1)},X_v^{(1)}\right)\right)\right)^k \qquad (7.1)
\end{aligned}$$

We now simplify this using a counting argument over the possible realizations of $X^{(1)}$

when restricted to edges in cycle C. Start by noting that

$$\sum_{X_C^{(1)}} \prod_{(u,v)\in C} (e^{2\delta})^{-2H\left(X_u^{(1)},X_v^{(1)}\right)} = 2\sum_{i=0}^{n/2} \left(\binom{|C|-1}{2i-1} + \binom{|C|-1}{2i} \right) (e^{2\delta})^{-2i}.$$

This follows by counting the number of possible ways to achieve a particular Hamming distance over the cycle. The $|C|-1$ (rather than $|C|$) and the grouping of consecutive binomial coefficients arises as we lose one "degree of freedom" due to examining a cycle, which fixes the Hamming distance to be even. Now, we apply Pascal's rule and can see

$$2\sum_{i=0}^{n/2} \left(\binom{|C|-1}{2i-1} + \binom{|C|-1}{2i} \right) (e^{2\delta})^{-2i} = 2\sum_{i=0}^{n/2} \binom{|C|}{2i} (e^{2\delta})^{-2i}.$$

This is twice the sum over the even terms in the binomial expansion of $(1+e^{-2\delta})^{|C|}$. The odd terms may be eliminated by adding $(1-e^{-2\delta})^{|C|}$, and thus (7.1) is equal to the following.

$$\begin{aligned}
&\left(\frac{e^\delta}{e^\delta+e^{-\delta}}\right)^{nk} \frac{1}{(n-1)!!} \sum_{M\in\mathcal{M}} \left(\prod_{\substack{\text{cycles } C\\ \in M_0\uplus M}} (1+e^{-2\delta})^{|C|} + (1-e^{-2\delta})^{|C|} \right)^k \\
&= \left(\frac{e^\delta}{e^\delta+e^{-\delta}}\right)^{nk} \frac{1}{(n-1)!!} \sum_{M\in\mathcal{M}} \left(\prod_{\substack{\text{cycles } C\\ \in M_0\uplus M}} \left(\frac{e^\delta+e^{-\delta}}{e^\delta}\right)^{|C|} \left(1+\left(\frac{e^\delta-e^{-\delta}}{e^\delta+e^{-\delta}}\right)^{|C|}\right) \right)^k \\
&= \mathbf{E}\left[\left(\prod_{\substack{\text{cycles } C\\ \in M_0\uplus M}} \left(1+\tanh^{|C|}(\delta)\right) \right)^k \right] \qquad (7.2)
\end{aligned}$$

where the expectation is from choosing a uniformly random matching $M\in\mathcal{M}$. At this point, it remains only to bound Equation (7.2). Noting that for all $x>0$ and $t\geq 1$,

$$1+\tanh^t(\delta) \leq 1+\delta^t \leq \exp\left(\delta^t\right),$$

we can bound (7.2) as

$$\mathbf{E}\left[\left(\prod_{\substack{\text{cycles } C\\ \in M_0\uplus M}} \left(1+\tanh^{|C|}(\delta)\right) \right)^k \right] \leq \mathbf{E}\left[\left(\prod_{\substack{\text{cycles } C\\ \in M_0\uplus M}} \exp\left(\delta^{|C|}\right) \right)^k \right].$$

For our purposes, it turns out that the 2-cycles will be the dominating factor, and we use the following crude upper bound:

$$\mathbf{E}\left[\left(\prod_{\substack{\text{cycles } C\\ \in M_0 \uplus M}} \exp\left(\delta^{|C|}\right)\right)^k\right] \leq \exp\left(\delta^4 nk/4\right)\mathbf{E}\left[\exp\left(\delta^2 \zeta k\right)\right],$$

where ζ is a random variable representing the number of 2-cycles in $M_0 \uplus M$, i.e., the number of edges shared by both matchings. We examine the distribution of ζ. Note that

$$\mathbf{E}[\zeta] = \frac{n}{2} \cdot \frac{1}{n-1} = \frac{n}{2(n-1)}.$$

More generally, for any positive integer $z \leq n/2$,

$$\mathbf{E}[\zeta - (z-1) | \zeta \geq z-1] = \frac{n-2z+2}{2} \cdot \frac{1}{n-2z+1} = \frac{n-2z+2}{2(n-2z+1)}.$$

By Markov's inequality,

$$\Pr[\zeta \geq z | \zeta \geq z-1] = \Pr[\zeta - (z-1) \geq 1 | \zeta \geq z-1] \leq \frac{n-2z+2}{2(n-2z+1)}.$$

Therefore,

$$\Pr[\zeta \geq z] = \prod_{i=1}^{z} \Pr[\zeta \geq i | \zeta \geq i-1] \leq \prod_{i=1}^{z} \frac{n-2i+2}{2(n-2i+1)}.$$

In particular, note that for all $z < n/2$,

$$\Pr[\zeta \geq z] \leq (2/3)^z.$$

We return to considering the expectation above:

$$\begin{aligned}
\mathbf{E}\left[\exp\left(\delta^2\zeta k\right)\right] &= \sum_{z=0}^{n/2} \Pr[\zeta = z] \exp\left(\delta^2 z k\right) \\
&\leq \sum_{z=0}^{n/2} \Pr[\zeta \geq z] \exp\left(\delta^2 z k\right) \\
&\leq \frac{3}{2} \sum_{z=0}^{n/2} (2/3)^z \exp\left(\delta^2 z k\right) \\
&= \frac{3}{2} \sum_{z=0}^{n/2} \exp\left((\delta^2 k - \log(3/2)) z\right) \\
&\leq \frac{3}{2} \cdot \frac{1}{1 - \exp\left(\delta^2 k - \log(3/2)\right)},
\end{aligned}$$

where the last inequality requires that $\exp\left(\delta^2 k - \log(3/2)\right) < 1$. This is true as long as $k < \log(3/2)/\delta^2 = \frac{\log(3/2)}{3} \cdot \frac{n}{\varepsilon}$.

Combining Lemma 18 with the above derivation, we have that

$$\begin{aligned}
2d_{\mathrm{TV}}^2(p^{\otimes k}, \bar{q}^{\otimes k}) &\leq \log\left(\exp(\delta^4 nk/4) \cdot \frac{3}{2(1 - \exp\left(\delta^2 k - \log(3/2)\right))}\right) \\
&= \delta^4 nk/4 + \log\left(\frac{3}{2(1 - \exp\left(\delta^2 k - \log(3/2)\right))}\right) \\
&= \frac{9\varepsilon^2}{4n} k + \log\left(\frac{3}{2(1 - \exp\left(3k\varepsilon/n - \log(3/2)\right))}\right).
\end{aligned}$$

If $k < \frac{1}{25} \cdot \frac{n}{\varepsilon}$, then $d_{\mathrm{TV}}(p^{\otimes k}, \bar{q}^{\otimes k}) < 49/50$ and thus no algorithm can distinguish between the two cases with probability $\geq 99/100$. This completes the proof of Theorem 15.

7.3.3 Proof of Theorem 16

We provide constructions for our lower bounds of Theorem 16 which show that a dependence on β is necessary in certain cases.

Lemma 19. *There exists a constant $c > 0$ such that, for all $\varepsilon < 1$ and $\beta \geq 0$, any algorithm, given sample access to an Ising model p, which distinguishes between the cases $p \in \mathcal{I}_n$ and $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$ with probability at least $99/100$ requires $k \geq c\beta/\varepsilon$ samples.*

Proof. Consider the following two models, which share some parameter $\tau > 0$:

1. An Ising model p on two nodes u and v, where $\theta_u^p = \theta_v^p = \tau$ and $\theta_{uv} = 0$.

2. An Ising model q on two nodes u and v, where $\theta_u^q = \theta_v^q = \tau$ and $\theta_{uv} = \beta$.

We note that $\mathbf{E}[X_u^p X_v^p] = \frac{\exp(2\tau+\beta)+\exp(-2\tau+\beta)-\exp(-\beta)}{\exp(2\tau+\beta)+\exp(-2\tau+\beta)+\exp(-\beta)}$ and $\mathbf{E}[X_u^q X_v^q] = \tanh^2(\tau)$. By (2.2), these two models have $d_{\mathrm{SKL}}(p,q) = \beta\left(\mathbf{E}[X_u^p X_v^p] - \mathbf{E}[X_u^q X_v^q]\right)$. For any for any fixed β sufficiently large and $\varepsilon > 0$ sufficiently small, τ can be chosen to make $\mathbf{E}[X_u^p X_v^p] - \mathbf{E}[X_u^q X_v^q] = \frac{\varepsilon}{\beta}$. This is because at $\tau = 0$, this is equal to $\tanh(\beta)$ and for $\tau \to \infty$, this approaches 0, so by continuity, there must be a τ which causes the expression to equal this value. Therefore, the SKL distance between these two models is ε. On the other hand, it is not hard to see that $d_{\mathrm{TV}}(p,q) = \Theta\left(\mathbf{E}[X_u^p X_v^p] - \mathbf{E}[X_u^q X_v^q]\right) = \Theta(\varepsilon/\beta)$, and therefore, to distinguish these models, we require $\Omega(\beta/\varepsilon)$ samples. □

Lemma 20. *There exists constants $c_1, c_2 > 0$ such that, for all $\varepsilon < 1$ and $\beta \geq c_1 \log(1/\varepsilon)$, any algorithm, given a description of an Ising model q with no external field (i.e., $h = 0$) and has sample access to an Ising model p, and which distinguishes between the cases $p = q$ and $d_{\mathrm{SKL}}(p,q) \geq \varepsilon$ with probability at least $99/100$ requires $k \geq c_2\beta/\varepsilon$ samples.*

Proof. This construction is very similar to that of Lemma 19. Consider the following two models, which share some parameter $\tau > 0$:

1. An Ising model p on two nodes u and v, where $\theta_{uv}^p = \beta$.

2. An Ising model q on two nodes u and v, where $\theta_{uv}^p = \beta - \tau$.

We note that $\mathbf{E}[X_u^p X_v^p] = \tanh(\beta)$ and $\mathbf{E}[X_u^q X_v^q] = \tanh(\beta - \tau)$. By (2.2), these two models have $d_{\mathrm{SKL}}(p,q) = \tau\left(\mathbf{E}[X_u^p X_v^p] - \mathbf{E}[X_u^q X_v^q]\right)$. Observe that at $\tau = \beta$, $d_{\mathrm{SKL}}(p,q) = \beta\tanh(\beta)$, and at $\tau = \beta/2$, $d_{\mathrm{SKL}}(p,q) = \frac{\beta}{2}(\tanh(\beta) - \tanh(\beta/2)) = \frac{\beta}{2}(\tanh(\beta/2)\,\mathrm{sech}(\beta)) \leq \beta\exp(-\beta) \leq \varepsilon$, where the last inequality is based on our condition that β is sufficiently large. By continuity, there exists some $\tau \in [\beta/2, \beta]$ such that $d_{\mathrm{SKL}}(p,q) = \varepsilon$. On the other hand, it is not hard to see that $d_{\mathrm{TV}}(p,q) = \Theta\left(\mathbf{E}[X_u^p X_v^p] - \mathbf{E}[X_u^q X_v^q]\right) = \Theta(\varepsilon/\beta)$, and therefore, to distinguish these models, we require $\Omega(\beta/\varepsilon)$ samples. □

The lower bound construction and analysis for the h lower bound follow almost identically, with the model q consisting of a single node with parameter h.

Lemma 21. *There exists constants $c_1, c_2 > 0$ such that, for all $\varepsilon < 1$ and $h \geq c_1 \log(1/\varepsilon)$, any algorithm, given a description of an Ising model q with no edge potentials(i.e., $\beta = 0$) and has sample access to an Ising model p, and which distinguishes between the cases $p = q$ and $d_{\mathrm{SKL}}(p,q) \geq \varepsilon$ with probability at least $99/100$ requires $k \geq c_2 h/\varepsilon$ samples.*

Together, Lemmas 19, 20, and 21 imply Theorem 16.

Chapter 8

Conclusion

Distributional Property Testing questions in the finite sample regime have been studied primarily for low-dimensional distributions where tight bounds are known in many cases. Little was known when it came to multi-dimensional distributions except that we run into lower bounds which are exponential in the dimension when we consider general multi-dimensional distributions. This thesis explores property testing on Ising models which are multi-dimensional distributions with a rich structure. Ising models have been studied extensively by physicists, statisticians, mathematicians and computer scientists and are known to exhibit complex behavior depending on the parameters of the model. The results presented in this thesis show that property testing on Ising models can, in general, be done with a number of samples which is polynomial in the dimension of the distribution. Upper and lower bounds for testing independence and identity were presented. A number of challenges lie in the way of attaining tight upper and lower bounds. For instance, to attain the linear lower bound of Theorem 15, we used matchings as the underlying graphs. Intuitively, denser graphs would have more power in thwarting detection by a tester but they end up being significantly harder to analyze for the purposes of the lower bound. In our upper bounds, we noted that since we do not know the signs of the pairwise correlations in the model, we need to expend some samples to perform the weak learning process described in Chapter 4. If we can circumvent the necessity for weak-learning then we could save on the sample complexity by a polynomial factor. Next, some potential future directions of the work presented in this thesis are listed.

8.1 Future Directions

There are many interesting directions to pursue building on the results in this thesis. One interesting question is what happens to the sample complexity if we impose structural restrictions on the underlying graph such as, for instance, requiring that the graph is a forest. Similarly, we can also study the testing question on special families of Ising models such as ferromagnetic Ising models.

A natural direction to pursue is to study testing problems on other structured high-dimensional distributions. Even within the class of graphical models, ones with a larger alphabet and k-way edges instead of the 2-way edges in Ising models, are an interesting open question. If we allow n-way edges, then we end up dealing with the class of all n-dimensional distributions for which we know from previous work that we need $\Theta\left(2^{n/2}\right)$ samples. It would be interesting to see what the sample complexity is if k-way edges are allowed and whether there is smooth interpolation of the sample complexity between the two extremes.

A central contribution of this thesis is the application and extension of Chatterjee's technique of exchangeable pairs for bounding variance of functions on the Ising model. The technique with slight modifications can be used to also prove exponential concentration bounds for functions on the Ising model. Attaining such a concentration bound for multilinear functions on the Ising model is an open problem and it would be interesting to see if the techniques in this thesis can be applied to attain the same.

Application of the algorithms developed in this thesis to real-world data is another direction. An application domain where the Ising model assumption could make sense would be predicting votes of people in a social network. Applying the theoretical framework developed here to such areas in practice would provide valuable insights into how practical the assumptions made in theory are and how efficient theory is in practice.

Bibliography

[1] Oded Goldreich. Combinatorial property testing (a survey). In *In: Randomization Methods in Algorithm Design*, pages 45–60. American Mathematical Society, 1998.

[2] Eldar Fischer. The art of uninformed decisions: A primer to property testing. *Science*, 75:97–126, 2001.

[3] Dana Ron. Property testing. In *Handbook of Randomized Computing, Vol. II*, pages 597–649. Kluwer Academic Publishers, 2000.

[4] Ronitt Rubinfeld and Asaf Shapira. Sublinear time algorithms. *SIAM J. Discret. Math.*, 25(4):1562–1588, November 2011.

[5] Clément L. Canonne, Dana Ron, and Rocco A. Servedio. Testing equivalence between distributions using conditional samples. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 1174–1192, Philadelphia, PA, USA, 2014. SIAM.

[6] Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the power of conditional samples in distribution testing. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, pages 561–580, New York, NY, USA, 2013. ACM.

[7] Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapathi, and Ananda Theertha Suresh. Faster algorithms for testing under conditional sampling. *CoRR*, abs/1504.04103, 2015.

[8] Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *J. Comput. Syst. Sci.*, 72(6):1012–1042, September 2006.

[9] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, 1900.

[10] Ronald A. Fisher. *The Design of Experiments.* Macmillan, 1935.

[11] Jon N.K. Rao and Alastair J. Scott. The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the Americal Statistical Association*, 76(374):221–230, 1981.

[12] Alan Agresti and Maria Kateri. *Categorical Data Analysis.* Springer, 2011.

[13] Vincent Y.F. Tan, Animashree Anandkumar, and Alan S. Willsky. Error exponents for composite hypothesis testing of Markov forest distributions. In *Proceedings of the 2010 IEEE International Symposium on Information Theory*, ISIT '10, pages 1613–1617, Washington, DC, USA, 2010. IEEE Computer Society.

[14] Tugkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, FOCS '01, pages 442–451, Washington, DC, USA, 2001. IEEE Computer Society.

[15] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[16] Reut Levi, Dana Ron, and Ronitt Rubinfeld. Testing properties of collections of distributions. *Theory of Computing*, 9(8):295–347, 2013.

[17] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pages 51–60, Washington, DC, USA, 2014. IEEE Computer Society.

[18] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 3577–3598. Curran Associates, Inc., 2015.

[19] Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. In *Proceedings of the 33rd Symposium on Theoretical Aspects of Computer Science*, STACS '16, pages 25:1–25:14, 2016.

[20] Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 685–694, Washington, DC, USA, 2016. IEEE Computer Society.

[21] Michael Jordan. Lecture notes for bayesian modeling and inference, 2010.

[22] Sujay Sanghavi, Vincent Tan, and Alan Willsky. Learning graphical models for hypothesis testing and classification. *IEEE Transactions on Signal Processing*, 58(11):5481–5495, 2010.

[23] Jeongwoo Ko, Eric Nyberg, and Luo Si. A probabilistic graphical model for joint answer ranking in question answering. In *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 343–350, New York, NY, USA, 2007. ACM.

[24] Guy Bresler. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the 47th Annual ACM Symposium on the Theory of Computing*, STOC '15, pages 771–782, New York, NY, USA, 2015. ACM.

[25] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

[26] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

[27] Sourav Chatterjee. *Concentration Inequalities with Exchangeable Pairs*. PhD thesis, Stanford University, June 2005.

[28] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates Sunderland, 2004.

[29] Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Evolutionary trees and the Ising model on the Bethe lattice: A proof of Steel's conjecture. *Probability Theory and Related Fields*, 149(1):149–189, 2011.

[30] Stuart Geman and Christine Graffigne. Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Mathematicians*, pages 1496–1517. American Mathematical Society, 1986.

[31] Glenn Ellison. Learning, local interaction, and coordination. *Econometrica*, 61(5):1047–1071, 1993.

[32] Andrea Montanari and Amin Saberi. The spread of innovations in social networks. *Proceedings of the National Academy of Sciences*, 107(47):20196–20201, 2010.

[33] C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

[34] Pieter Abbeel, Daphne Koller, and Andrew Y. Ng. Learning factor graphs in polynomial time and sample complexity. *Journal of Machine Learning Research*, 7(Aug):1743–1788, 2006.

[35] Imre Csiszár and Zsolt Talata. Consistent estimation of the basic neighborhood of Markov random fields. *The Annals of Statistics*, 34(1):123–145, 2006.

[36] Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[37] Ali Jalali, Christopher C. Johnson, and Pradeep K. Ravikumar. On learning discrete graphical models using greedy methods. In *Advances in Neural Information Processing Systems 24*, NIPS '11, pages 1935–1943. Curran Associates, Inc., 2011.

[38] Narayana P. Santhanam and Martin J. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012.

[39] Guy Bresler, David Gamarnik, and Devavrat Shah. Structure learning of antiferromagnetic Ising models. In *Advances in Neural Information Processing Systems 27*, NIPS '14, pages 2852–2860. Curran Associates, Inc., 2014.

[40] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of Ising models. In *Advances in Neural Information Processing Systems 29*, NIPS '16, pages 2595–2603. Curran Associates, Inc., 2016.

[41] Guy Bresler and Mina Karzand. Learning a tree-structured Ising model in order to make predictions. *arXiv preprint arXiv:1604.06749*, 2016.

[42] Bhaswar B. Bhattacharya. Power of graph-based two-sample tests. *arXiv preprint arXiv:1508.07530*, 2016.

[43] Bhaswar B. Bhattacharya and Sumit Mukherjee. Inference in Ising models. *Bernoulli*, 2016.

[44] Hans-Otto Georgii. *Gibbs Measures and Phase Transitions*. Walter de Gruyter, 2011.

[45] Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

[46] Jayadev Acharya and Constantinos Daskalakis. Testing Poisson binomial distributions. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 1829–1840, Philadelphia, PA, USA, 2015. SIAM.

[47] José A. Adell and Pedro Jodrá. Exact Kolmogorov and total variation distances between some familiar discrete distributions. *Journal of Inequalities and Applications*, 2006(1):1–8, 2006.

[48] Jon Feldman, Ryan O'Donnell, and Rocco A. Servedio. Learning mixtures of product distributions over discrete domains. *SIAM Journal on Computing*, 37(5):1536–1564, 2008.

Appendix A

Weakly Learning Rademacher Random Variables

In this section, we examine the concept of "weakly learning" Rademacher random variables. This problem we study is classical, but our regime of study and goals are slightly different. Suppose we have k samples from a random variable, promised to either be *Rademacher*$(1/2+\lambda)$ or *Rademacher*$(1/2-\lambda)$, for some $0 < \lambda \leq 1/2$. How many samples do we need to tell which case we are in? If we wish to be correct with probability (say) $\geq 2/3$, it is folklore that $k = \Theta(1/\lambda^2)$ samples are both necessary and sufficient. In our weak learning setting, we focus on the regime where we are sample limited (say, when λ is very small), and we are unable to gain a constant benefit over randomly guessing. More precisely, we have a budget of k samples from some *Rademacher*(p) random variable, and we want to guess whether $p > 1/2$ or $p < 1/2$. The "margin" $\lambda = |p - 1/2|$ may not be precisely known, but we still wish to obtain the maximum possible advantage over randomly guessing, which gives us probability of success equal to 1/2. We show that with any $k \leq 1/4\lambda^2$ samples, we can obtain success probability $1/2 + \Omega(\lambda\sqrt{k})$. This smoothly interpolates within the "low sample" regime, up to the point where $k = \Theta(1/\lambda^2)$ and folklore results also guarantee a constant probability of success. We note that in this low sample regime, standard concentration bounds like Chebyshev and Chernoff give trivial guarantees, and our techniques require a more careful examination of the Binomial PMF.

We go on to examine the same problem under alternate centerings – where we are trying to determine whether $p > \mu$ or $p < \mu$, generalizing the previous case where $\mu = 1/2$. We provide a simple "recentering" based reduction to the previous case, showing that the same upper bound holds for all values of μ. We note that our reduction holds even when the centering μ is not explicitly known, and we only have limited sample access to *Rademacher*(μ).

We start by proving the following lemma, where we wish to determine the direction of bias with respect to a zero-mean Rademacher random variable.

Lemma 22. *Let $X_1, \ldots, X_k$ be iid random variables, distributed as Rademacher(p) for any $p \in [0,1]$. There exists an algorithm which takes $X_1, \ldots, X_k$ as input and outputs a value $b \in \{\pm 1\}$, with the following guarantees: there exists constants $c_1, c_2 > 0$ such that for any $p \neq \frac{1}{2}$,*

$$\Pr\left(b = \mathbf{sign}\left(\lambda\right)\right) \geq \begin{cases} \frac{1}{2} + c_1|\lambda|\sqrt{k} & \textit{if } k \leq \frac{1}{4\lambda^2} \\ \frac{1}{2} + c_2 & \textit{otherwise,} \end{cases}$$

where $\lambda = p - \frac{1}{2}$. *If* $p = \frac{1}{2}$, *then* $b \sim$ *Rademacher* $\left(\frac{1}{2}\right)$.

Proof. The algorithm is as follows: let $S = \sum_{i=1}^{k} X_i$. If $S \neq 0$, then output $b = \mathbf{sign}(S)$, otherwise output $b \sim$ *Rademacher* $\left(\frac{1}{2}\right)$.

The $p = 1/2$ case is trivial, as the sum S is symmetric about 0. We consider the case where $\lambda > 0$ (the negative case follows by symmetry) and when k is even (odd k can be handled similarly). As the case where $k > \frac{1}{4\lambda^2}$ is well known (see Lemma 1), we focus on the former case, where $\lambda \leq \frac{1}{2\sqrt{k}}$. By rescaling and shifting the variables, this is equivalent to lower bounding $\Pr\left(\mathit{Binomial}\left(k, \frac{1}{2} + \lambda\right) \geq \frac{k}{2}\right)$. By a symmetry argument, this is equal to

$$\frac{1}{2} + d_{\mathrm{TV}}\left(\mathit{Binomial}\left(k, \frac{1}{2} - \lambda\right), \mathit{Binomial}\left(k, \frac{1}{2} + \lambda\right)\right).$$

It remains to show this total variation distance is $\Omega(\lambda\sqrt{k})$.

$$\begin{align*}
& d_{\mathrm{TV}}\left(\mathit{Binomial}\left(k, \frac{1}{2} - \lambda\right), \mathit{Binomial}\left(k, \frac{1}{2} + \lambda\right)\right) \\
\geq \quad & d_{\mathrm{TV}}\left(\mathit{Binomial}\left(k, \frac{1}{2}\right), \mathit{Binomial}\left(k, \frac{1}{2} + \lambda\right)\right) \\
\geq \quad & k \min_{\ell \in \{\lceil k/2 \rceil, \ldots, \lceil k/2 + k\lambda \rceil\}} \int_{1/2}^{1/2+\lambda} \Pr\left(\mathit{Binomial}\left(k-1, u\right) = l - 1\right) du \tag{A.1} \\
\geq \quad & \lambda k \cdot \Pr\left(\mathit{Binomial}\left(k-1, 1/2 + \lambda\right) = k/2\right) \\
= \quad & \lambda k \cdot \binom{k-1}{k/2} \left(\frac{1}{2} + \lambda\right)^{k/2} \left(\frac{1}{2} - \lambda\right)^{k/2-1} \\
\geq \quad & \Omega(\lambda k) \cdot \sqrt{\frac{1}{2k}} \left(1 + \frac{1}{\sqrt{k}}\right)^{k/2} \left(1 - \frac{1}{\sqrt{k}}\right)^{k/2} \tag{A.2} \\
= \quad & \Omega(\lambda\sqrt{k}) \cdot \left(1 - \frac{1}{k}\right)^{k/2} \\
\geq \quad & \Omega(\lambda\sqrt{k}) \cdot \exp\left(-1/2\right) \left(1 - \frac{1}{k}\right)^{1/2} \tag{A.3} \\
= \quad & \Omega(\lambda\sqrt{k}),
\end{align*}$$

as desired.

(A.1) applies Proposition 2.3 of [47]. (A.2) is by an application of Stirling's approximation and since $\lambda \leq \frac{1}{2\sqrt{k}}$. (A.3) is by the inequality $\left(1 - \frac{c}{k}\right)^k \geq \left(1 - \frac{c}{k}\right)^c \exp(-c)$. □

We now develop a corollary allowing us to instead consider comparisons with respect to different centerings.

Corollary 1. *Let* $X_1, \ldots, X_k$ *be iid random variables, distributed as Rademacher*(p) *for any* $p \in [0, 1]$. *There exists an algorithm which takes* $X_1, \ldots, X_k$ *and* $q \in [0, 1]$ *as input and outputs a value* $b \in \{\pm 1\}$, *with the following guarantees: there exists constants* $c_1, c_2 > 0$ *such that for any* $p \neq q$,

$$\Pr\left(b = \mathbf{sign}\left(\lambda\right)\right) \geq \begin{cases} \frac{1}{2} + c_1|\lambda|\sqrt{k} & \text{if } k \leq \frac{1}{4\lambda^2} \\ \frac{1}{2} + c_2 & \text{otherwise,} \end{cases}$$

where $\lambda = \frac{p-q}{2}$*. If* $p = q$*, then* $b \sim$ *Rademacher* $\left(\frac{1}{2}\right)$*.*

This algorithm works even if only given k *iid samples* $Y_1, \ldots, Y_k \sim$ *Rademacher*(q)*, rather than the value of* q*.*

Proof. Let $X \sim$ *Rademacher*(p) and $Y \sim$ *Rademacher*(q). Consider the random variable Z defined as follows. First, sample X and Y. If $X \neq Y$, output $\frac{1}{2}(X - Y)$. Otherwise, output a random variable sampled as *Rademacher* $\left(\frac{1}{2}\right)$. One can see that $Z \sim$ *Rademacher* $\left(\frac{1}{2} + \frac{p-q}{2}\right)$.

Our algorithm can generate k iid samples $Z_i \sim$ *Rademacher* $\left(\frac{1}{2} + \frac{p-q}{2}\right)$ in this method using X_i's and Y_i's, where Y_i's are either provided as input to the algorithm or generated according to *Rademacher*(q). At this point, we provide the Z_i's as input to the algorithm of Lemma 22. By examining the guarantees of Lemma 22, this implies the desired result. □

Appendix B

An Attempt towards Testing by Learning in KL-divergence

One approach to testing problems is by learning the distribution which we wish to test. If the distance of interest is the total variation distance, then a common approach to learning is a cover-based method. One first creates a set of hypothesis distributions H which $O(\varepsilon)$-covers the space. Then by drawing $k = \tilde{O}(\log |H|/\varepsilon^2)$ samples from p, we can output a distribution from H with the guarantee that it is at most $O(\varepsilon)$-far from p. The algorithm works by computing a score based on the samples for each of the distributions in the hypothesis class and then choosing the one with the maximum score.

However, it is not clear if this approach would work for testing in KL-divergence (an easier problem than testing in SKL-divergence) because KL-divergence does not satisfy the triangle inequality. In particular, if p and q are far, and we learn a distribution $\hat{p}$ which is close to p, we no longer have the guarantee that $\hat{p}$ and q are still far. Even if this issue were somehow resolved, the best known sample complexity for learning follows from the maximum likelihood algorithm. We state the guarantees provided by Theorem 17 of [48].

Theorem 17 (Theorem 17 from [48])**.** *Let $b, a, \varepsilon > 0$ such that $a < b$. Let $\mathcal{Q}$ be a set of hypothesis distributions for some distribution p over the space X such that at least one $q^* \in \mathcal{Q}$ is such that $d_{\mathrm{KL}}(p||q^*) \le \varepsilon$. Suppose also that $a \le q(x) \le b$ for all $q \in \mathcal{Q}$ and for all x such that $p(x) > 0$. Then running the maximum likelihood algorithm on $\mathcal{Q}$ using a set S of i.i.d. samples from p, where $|S| = k$, outputs a $q^{ML} \in \mathcal{Q}$ such that $d_{\mathrm{KL}}(p||q^{ML}) \le 4\varepsilon$ with probability $1 - \delta$ where*

$$\delta \le (|\mathcal{Q}| + 1) \exp\left(\frac{-2k\varepsilon^2}{\log^2\left(\frac{b}{a}\right)}\right).$$

To succeed with probability at least 2/3, we need that

$$k \ge \frac{\log\left(3(|\mathcal{Q}| + 1)\right) \log^2\left(\frac{b}{a}\right)}{2\varepsilon^2}$$

For the Ising model, a KL-cover $\mathcal{Q}$ would consist of creating a poly(n/ε) mesh for each parameter. Since there are $O(n^2)$ parameters, the cover will have a size of poly$(n/\varepsilon)^{n^2}$. Letting β and h denote the maximum edge and node parameter (respectively), then the ratio b/a in the above theorem is such that

$$\frac{b}{a} \geq \exp\left(O(n^2\beta + nh)\right).$$

Therefore, the number of samples required by this approach would be

$$\begin{aligned} k &= O\left(\frac{n^2 \log\left(\frac{n}{\varepsilon}\right) \dot{(}n^2\beta + nh)^2}{\varepsilon^2}\right) \\ &= \tilde{O}\left(\frac{n^6\beta^2 + n^4h^2}{\varepsilon^2}\right) \end{aligned}$$

which is more expensive than our baseline, the localization algorithm of Theorem 2. Additionally, this algorithm is computationally inefficient, as it involves iterating over all hypotheses in the exponentially large set $\mathcal{Q}$. To summarize, there are a number of issues preventing a learning-based approach from giving an efficient tester.

Appendix C

High-Temperature Mixing Times and Concentration of Lipschitz Functions

We show several useful properties of the Ising model in the high temperature regime of Definition 2. In fact, we will show these properties for an even more permissive regime, captured by the following definition.

Definition 5. *For all $(u, v) \in E$, suppose $\theta_{uv} \leq \frac{\eta}{4\max\{d_u,d_v\}}$, where d_u and d_v are the degrees of u, v in G, and $\eta < 1$ is any constant.*

Lemma 23. *Consider the $V \times V$ matrix $A = (a_{uv})_{uv}$ where, for all $u \neq v$, $a_{u,v} = 4\theta_{uv}$ and, for all u, $a_{uu} = 0$. Suppose also that, for all $u \neq v$, θ_{uv} satisfies the conditions of Definition 5. Then $|A|_2 \leq \eta < 1$, where η is as in Definition 5.*

Proof of Lemma 23: Take any vector x such that $|x|_2 = 1$. Then

$$\begin{aligned}|A \cdot x|_2^2 &= \sum_u \left(\sum_{v \in N(u)} 4\theta_{uv} x_v \right)^2 \leq \sum_u \frac{\eta^2}{d_u d_v} \left(\sum_{v \in N(u)} |x_v| \right)^2 \\ &\leq \sum_u \frac{\eta^2}{d_u d_v} \left[\left(\sum_{v \in N(u)} x_v^2 \right) d_u \right] = \sum_v \frac{\eta^2 \cdot x_v^2}{d_v} \left(\sum_{u \in N(v)} 1 \right) = \eta^2 \cdot |x|_2^2 \leq \eta^2.\end{aligned}$$

where the second inequality is by Cauchy-Schwarz. □

Lemma 24. *The mixing time of the Glauber dynamics in an Ising model satisfying the high temperature conditions of Definition 5 is $O(n \log n)$.*

Proof of Lemma 24: This is quite standard and related to Dobrushin's uniqueness criterion. As we have not seen it stated in the full spectrum of Ising models we consider here, we provide a proof for completeness. Our proof follows the line of argumentation in the proof of Theorem 4.3 in [27], where a concentration bound is proven.

We use a coupling argument, considering two coupled executions $(X_t)_t$ and $(X'_t)_t$ of the Glauber dynamics starting at arbitrary states $X_0 = x$ and $X'_0 = x'$. We couple these executions using the greedy coupling explained in Section 5.1.1. Namely, at each step $t > 0$

of the coupled executions, we choose to update the same (uniformly randomly chosen) vertex v in both chains and we set $X_{t,v}$ and $X'_{t,v}$ so as to maximize the probability that they are equal. In particular, if we choose to update node v in the 2 Chainz, then the probability that $X_{t,v}$ and $X'_{t,v}$ are different is:

$$\Pr[X_{t,v} \neq X'_{t,v} | v \text{ is chosen}, X_{t-1}, X'_{t-1}] = d_{\mathrm{TV}}(\mu_v(\cdot|X_{t-1,N(v)}), \mu_v(\cdot|X'_{t-1,N(v)})),$$

where d_{TV} denotes total variation distance and $\mu_v(\cdot|X_{t-1,N(v)})$, $\mu_v(\cdot|X'_{t-1,N(v)})$ represent the conditional measures at node v conditioning respectively on the states $X_{t-1,N(v)}, X'_{t-1,N(v)}$ of v's neighborhood. Defining matrix A as in the statement of Lemma 23, it follows from Lemma 4.4 of [27] that

$$d_{\mathrm{TV}}(\mu_v(\cdot|X_{t-1,N(v)}), \mu_v(\cdot|X'_{t-1,N(v)})) \leq \sum_{u \in N(v)} a_{vu} \mathbb{1}_{X_{t-1,u} \neq X'_{t-1,u}} \equiv \sum_u a_{vu} \mathbb{1}_{X_{t-1,u} \neq X'_{t-1,u}}$$

So it follows from the above that:

$$\Pr[X_{t,v} \neq X'_{t,v} \ \text{ and } v \text{ is chosen} | X_{t-1}, X'_{t-1}] \leq \frac{1}{n} \sum_u a_{vu} \mathbb{1}_{X_{t-1,u} \neq X'_{t-1,u}}.$$

On the other hand:

$$\Pr[X_{t,v} \neq X'_{t,v} \ \text{ and } v \text{ not chosen} | X_{t-1}, X'_{t-1}] = \left(1 - \frac{1}{n}\right) \mathbb{1}_{X_{t-1,v} \neq X'_{t-1,v}}.$$

Hence, overall:

$$\Pr[X_{t,v} \neq X'_{t,v}] \leq \left(1 - \frac{1}{n}\right) \Pr[X_{t-1,v} \neq X'_{t-1,v}] + \frac{1}{n} \sum_u a_{vu} \Pr[X_{t-1,u} \neq X'_{t-1,u}].$$

So suppose that ℓ_t is a non-negative vector such that $\ell_{t,v} = \Pr[X_{t,v} \neq X'_{t,v}]$. For all $t > 0$, we have:

$$\ell_t \leq \left(\left(1 - \frac{1}{n}\right) I + \frac{1}{n} A\right) \ell_{t-1} =: B\ell_{t-1},$$

where the inequality holds coordinate-wise and we have set $B = \left(1 - \frac{1}{n}\right) I + \frac{1}{n} A$. Note that $|B|_2 \leq \left(1 - \frac{1}{n}\right) + \frac{1}{n}|A|_2 \leq \left(1 - \frac{1-\eta}{n}\right)$, where for the last inequality we used Lemma 23. Setting $t^* = cn \log n$, we have

$$\begin{aligned} |\ell_{t^*}|_2 \leq |B|_2^{t^*} |\ell_0|_2 &\leq \left(1 - \frac{1-\eta}{n}\right)^{t^*} |\ell_0|_2 \\ &\leq \left(1 - \frac{1-\eta}{n}\right)^{t^*} \sqrt{n} \qquad \text{(using that for any vector of probabilities } |\ell_0|_2 <= \sqrt{n}) \\ &\leq e^{-(1-\eta)c \log n} \sqrt{n} \leq 1/(4\sqrt{n}), \end{aligned}$$

for sufficiently large constant c. This means that $|\ell_{t^*}|_1 \leq 1/4$. Hence, $\Pr[X_{t^*} \neq X'_{t^*}] \leq |\ell_{t^*}|_1 \leq 1/4$. So the mixing time of the chain is $O(n \log n)$. □

Lemma 25. *Take any linear function $f(x) = \sum_v s_v x_v$, where $s \in \mathbb{R}^V$. Suppose that X is drawn from an Ising model satisfying the high temperature conditions of Definition 5. Then*

1. $\mathbf{Var}[f(x)] \leq \frac{2\sum_v s_v^2}{1-\eta}$.

2. For all $t \geq 0$,

$$\Pr[|f(X) - \mathbf{E}[f(X)]| \geq t] \leq 2e^{-\frac{(1-\eta)t^2}{4\sum_v s_v^2}}.$$

Proof of Lemma 25: The second claim follows directly from the statement of Theorem 4.3 of [27]. Indeed, the matrix A defined as in the statement of Lemma 23 satisfies, using Lemma 4.4 of [27]:

$$d_{\mathrm{TV}}(\mu_v(\cdot|X_{N(v)}), \mu_v(\cdot|X'_{N(v)})) \leq \sum_{u \in N(v)} a_{vu} \mathbb{1}_{X_u \neq X'_u} \equiv \sum_u a_{vu} \mathbb{1}_{X_u \neq X'_u}.$$

At the same time, $|A|_2 \leq \eta$ by Lemma 23, and function f satisfies the generalized Lipschitz condition: $|f(x) - f(x')| \leq \sum_v 2|s_v| \mathbb{1}_{x_i \neq x'_i}$. So we can directly apply Theorem 4.3 of [27].

To bound the variance of $f(X)$ we appeal to the proof of Theorem 4.3 of [27]. The proof defines an exchangeable pair (X, X'), where X is distributed according to the Ising model, and an antisymmetric function $F(X, X')$ such that

$$f(X) - \mathbf{E}[f(X)] = \mathbf{E}\left[F(X, X')|X\right].$$

In terms of the exchangeable pair and F, we can express the variance of $f(X)$ as follows:

$$\begin{aligned}
\mathbf{Var}(f(X)) &= \frac{1}{2} \cdot \mathbf{E}\left[(f(X) - f(X')) \cdot F(X, X')\right] \\
&= \frac{1}{2} \cdot \mathbf{E}\left[\mathbf{E}\left[(f(X) - f(X')) \cdot F(X, X')|X\right]\right] \\
&\leq \mathbf{E}\left[\frac{1}{2} \cdot \mathbf{E}\left[|(f(X) - f(X')) \cdot F(X, X')||X\right]\right]
\end{aligned}$$

The proof of Theorem of [27] shows that point-wise:

$$\frac{1}{2} \cdot \mathbf{E}\left[|(f(X) - f(X')) \cdot F(X, X')||X\right] \leq \frac{4\sum_v s_v^2}{2(1 - |A|_2)} \leq \frac{2\sum_v s_v^2}{1 - \eta},$$

concluding our proof. □