Electronic Colloquium on Computational Complexity, Revision 2 of Report No. 6 (2017)

Testing Ising Models

Constantinos Daskalakis[*]
EECS, MIT
costis@mit.edu

Nishanth Dikkala[†]
EECS, MIT
nishanthd@csail.mit.edu

Gautam Kamath[‡]
EECS, MIT
g@csail.mit.edu

October 30, 2017

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 far from every product distribution? Similarly, is it possible to distinguish whether p equals a given distribution q versus p and q being 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 [BFF$^+$01, Pan08, VV17, ADK15, DK16]. 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 initiate the study of 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. One of the key technical challenges we face along the way is bounding the variance of functions of the Ising model.

[*] Supported by a Microsoft Research Faculty Fellowship, NSF CCF-1551875, CCF-1617730, CCF-1650733, and ONR N00014-12-1-0999.

[†] Supported by NSF CCF-1551875, CCF-1617730, CCF-1650733, and ONR N00014-12-1-0999.

[‡] Supported by NSF CCF-1551875, CCF-1617730, CCF-1650733, and ONR N00014-12-1-0999.

ISSN 1433-8092

Contents

1 Introduction

The two most fundamental problems in Statistics are perhaps testing independence and goodness-of-fit. *Independence testing* is the problem of distinguishing, given samples from a multivariate distribution p, whether or not it is the product of its marginals. The applications of this problem abound: for example, a central problem in genetics is to test, given genomes of several individuals, whether certain single-nucleotide-polymorphisms (SNPs) are independent from each other. In anthropological studies, a question that arises over and over again is testing whether the behaviors of individuals on a social network are independent; see e.g. [CF07]. The related problem of *goodness-of-fit testing* is that of distinguishing, given samples from p, whether or not it equals a specific "model" q. This problem arises whenever one has a hypothesis (model) about the random source generating the samples and needs to verify whether the samples conform to the hypothesis.

Testing independence and goodness-of-fit have a long history in statistics, since the early days; for some old and some more recent references see, e.g., [Pea00, Fis35, RS81, Agr12]. Traditionally, the emphasis has been on the asymptotic analysis of tests, pinning down their error exponents as the number of samples tends to infinity [Agr12, TAW10]. In the two decades or so, distribution testing has also piqued the interest of theoretical computer scientists, where the emphasis has been different [BFF$^+$01, Pan08, LRR13, VV17, ADK15, CDGR16, DK16]. In contrast to much of the statistics literature, the goal has been to minimize the number of samples required for testing. From this vantage point, our testing problems take the following form:

Goodness-of-fit (or Identity) Testing: Given 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 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, and $d(\cdot, \cdot)$ some notion of distance or divergence between distributions, such as the total variation distance or the KL divergence. As usual, $\frac{2}{3}$ 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 multiplicative factor of $O(\log 1/\delta)$ in the sample complexity.

For both testing problems, recent work has identified tight upper and lower bounds on their sample complexity [Pan08, VV17, ADK15, DK16]: 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 [ADK15, DK16], or from a particular distribution q [Pan08]. Worst-case 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 [Jor10] for the basics of MRFs and the references [STW10, KNS07] 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] Recall that the Ising model is a distribution over $\{-1,1\}^n$, defined in terms of a graph $G=(V,E)$ with n nodes. It is parameterized by a scalar parameter $\theta_{u,v}$ for every edge $(u,v)\in E$, and a scalar parameter θ_v for every node $v\in V$, in terms of which it samples a vector $x\in\{\pm1\}^V$ with probability:

$$p(x)=\exp\left(\sum_{v\in V}\theta_v x_v+\sum_{(u,v)\in E}\theta_{u,v}x_u x_v-\Phi(\vec{\theta})\right), \tag{1}$$

where $\vec{\theta}$ is the parameter vector and $\Phi(\vec{\theta})$ 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 [Isi25]. 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 [LPW09, Cha05, Fel04, DMR11, GG86, Ell93, MS10]. 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., [CL68, AKN06, CT06, RWL10, JJR11, SW12, BGS14, Bre15, VMLC16, BK16, Bha16, BM16, MdCCU16, HKM17, KM17]. 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.

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

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 Applications: 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, [Ell93, MS10] 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, define 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). 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.

As another application, we turn to the field of computer vision. In the Bayesian setting, it is assumed that images are generated according to some prior distribution. Often, practitioners take this prior to be an Ising model in the binary case, or, in general, a higher-order MRF [GG86]. As such, a dataset of images can be pictured as random samples from this prior. A natural question to ask is, given some distribution, does a set of images conform to this prior? This problem corresponds to goodness-of-fit testing for Ising models.

A third application comes up in the field of medicine and computational biology. In order to improve diagnosis, symptom prediction and classification, as well as to improve overall healthcare outcomes, graphical models are trained on data, often using heuristic methods [FLNP00] and surgeon intuition, thereby incorporating hard-wired expert knowledge; see, i.e., the pneumonia graphical model identified in [LAFH01]. Our methods give efficient algorithms for testing the accuracy of such models. Furthermore, when the discrepancy is large, we expect that our algorithms could reveal the structural reasons for the discrepancy, i.e., blaming a large portion of the error on a misspecified edge.

Main Results and Techniques: Our main result is the following:

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

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 $O\left(\frac{n^2 d_{\max}^2 \beta^2 + n d_{\max} h^2}{\varepsilon^2}\right)$, 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 $d_{\mathrm{SKL}}(p, q) > \varepsilon$, 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 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 4.

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

B. Anchoring Our Expectations. Our 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 statistically 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 $\Omega\left(\frac{n^6\beta^2 + n^4 h^2}{\varepsilon^2}\right)$ samples and is time inefficient. In particular, our baseline algorithm already beats this sample complexity and is also time-efficient. Alternatively, one could aim to parameter-learn p; see, e.g., [SW12, Bre15, VMLC16] and their references. However, these algorithms require sample complexity that is exponential in the maximum degree [SW12], and they typically use samples exponential in β as well [Bre15, VMLC16]. For instance, if we use [VMLC16], 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.

The problem of learning the structure of Ising models (i.e., determining which edges are present in the graph) has enjoyed much study, especially in information theory – see [Bre15, VMLC16] for some recent results. At a first glance, one may hope that these results have implications for testing Ising models. However, thematic similarities aside, the two problems are qualitatively very different – our problem focuses on statistical estimation, while theirs looks at structural estimation. To point out some qualitative differences for these two problems, the complexity of structure learning is exponential in the maximum degree and β, while only logarithmic in n. On the other hand, for testing Ising models, the complexity has a polynomial dependence in all three parameters, which is both necessary and sufficient.

C. Trees and Ferromagnets. When p is a tree-structured (or forest-structured) Ising model, then independence testing can be performed computationally efficiently without any dependence on β, with an additional quadratic improvement with respect to the other parameters. In particular, without external fields, i.e. $\max\{|\theta^p_u|\} = 0$, independence can be solved from $O(\frac{n}{\varepsilon})$ samples, and this result is tight when $m = O(n)$; see Theorem 3 for an upper bound and Theorem 19 for a lower bound. Interestingly, we show the dependence on β cannot be avoided in the presence of external fields, or if we switch to the problem of identity testing; see Theorem 20. In the latter case, we can at least maintain the linear dependence on n; see Theorem 4. Similar results hold when p is a ferromagnet, i.e. $\theta^p_{u,v} \geq 0$, with no external fields, even if it is not a tree. In particular, the sample complexity becomes $O(\frac{\max\{m,n\}}{\varepsilon})$ (which is again tight when $m = O(n)$), see Theorem 5.

Technical Discussion 2.0: "Testing via Strong Localization." The improvements that we have just discussed are obtained via the same localization approach discussed earlier, which resulted into our baseline tester. That is, we are still going to "blame it on a node or an edge." The removal of the β dependence and the improved running times are due to the proof of a structural lemma, which relates the parameter $\theta_{u,v}$ on some edge (u,v) of the Ising model to the $\mathbf{E}[X_u X_v]$. We show that for forest-structured Ising models with no external fields, $\mathbf{E}[X_u X_v] = \tanh(\theta_{u,v})$, see Lemma 8. A similar statement holds for ferromagnets with no external field, i.e., $\mathbf{E}[X_u X_v] \geq \tanh(\theta_{u,v})$, see Lemma 11. The proof of the structural lemma for trees/forests is straightforward. Intuitively, the only source of correlation between the endpoints u and v of some edge (u,v) of the Ising model is the edge itself, as besides this edge there are no other paths between u and v that would provide alternative avenues for correlation. Significant more work is needed to prove the inequality for ferromagnets on arbitrary graphs. Now, there may be several paths between u and v besides the edge connecting them. Of course, because the model is a ferromagnet, these paths should intuitively only contribute to increase $\mathbf{E}[X_u X_v]$ beyond $\tanh(\theta_{u,v})$. But making this formal is not easy, as calculations involving the Ising model quickly become unwieldy beyond trees.[2] Our argument uses a coupling between (an appropriate generalization of) the Fortuin-Kasteleyn random cluster model and the Ising model. The coupling provides an alternative way to sample the Ising model by first sampling a random clustering of the nodes, and then assigning uniformly random spins to the sampled clusters. Moreover, it turns out that the probability that two nodes u and v land in the same cluster increases as the vector of parameters $\vec{\theta}$ of the Ising model increases. Hence, we can work inductively. If only edge (u,v) were present, then $\mathbf{E}[X_u X_v] = \tanh(\theta_{u,v})$. As we start adding edges, the probability that u, v land in the same cluster increases, hence the probability that they receive the same spin increases, and therefore $\mathbf{E}[X_u X_v]$ increases.

A slightly more detailed discussion of the structural result for ferromagnets is in Section 1.1.1, and full details about our testers for trees and ferromagnets can be found in Sections 4.1 and 4.2, respectively.

D. Dobrushin's Uniqueness Condition and 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 [Geo11]. 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. The Ising model has been studied extensively in such high-temperature regimes [Dob56, Cha05, Hay06, DGJ08], and it is a regime that is often used in practice.

In the high-temperature regime, we show that we can improve our baseline result without making ferromagnetic or tree-structure assumptions, using a non-localization based argument, explained next. In particular, we show in Theorem 7 that under high temperature and with no external fields independence testing can be done computationally efficiently from $\tilde{O}\left(\frac{n^{10/3}}{\varepsilon^2 d_{\max}^2}\right)$ samples, which

[2] We note that the partition function is #P-hard to compute[JS93].

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 6. Similar improvements hold when external fields are present (Theorem 9), as well as for identity testing, without and with external fields (Theorems 10 and 11).

We offer some intuition about the improvements in Figures 1 and 2 (appearing in Section 6), which are plotted for high temperature and no external fields. In Figure 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 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, and Variance Bounds." One way or another all our results up to this point 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 cases. 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 signs that have a non-trivial agreement with the correct signs, from fewer samples. Despite the $X_u X_v$ terms potentially having nasty correlations with each other, a careful analysis using anti-concentration calculations allows us to sidestep this $O(n^2)$ 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 trivially be bounded by $O(n^4)$. However, applying this bound in our algorithm gives a vacuous sample complexity. As the X_u's will experience a complex correlation structure, it is not clear how one might arrive at non-trivial bounds for the variance of such statistics, leading to the following natural question:

Question 1. *How can one bound the variance of statistics over high-dimensional distributions?*

This meta-question is at the heart of many high-dimensional statistical tasks, and we believe it is important to develop general-purpose frameworks for such settings. In the context of the Ising model, in fairly general regimes, we can show the variance to be $\tilde{O}(n^2)$. We consider this to be surprising – 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 pairwise independent.

This question has been studied in-depth in two recent works [DDK17, GLP17], which prove concentration of measure for d-linear statistics over the Ising model. We note that these results are stronger than what we require in this work – we need only variance bounds (which are implied by concentration of measure) for bilinear statistics. Despite these stronger bounds, for completeness, we present a proof of the variance bounds for bilinear statistics which we require[3]. This approach

[3] We thank Yuval Peres for directing us towards the reference [LPW09] and the tools required to prove these bounds.

uses tools from [LPW09]. It requires a bound on the spectral gap of the Markov chain, and an expected Lipschitz property of the statistic when a step is taken at stationarity. The technique is described in Section 7, and the variance bounds are given in Theorems 16 and 17.

E. Our Main Lower Bound. The proof of our linear lower bound applies Le Cam's method [LC73]. Our construction is inspired by Paninski's lower bound for uniformity testing [Pan08], 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. We analyze a construction of a similar nature as a warm-up for our main lower bound, while also proving a lower bound for uniformity testing on product distributions over a binary alphabet (which are a special case of the Ising model where no edges are present), see Theorem 18. To achieve the linear lower bound, we instead consider a *random* matching of the nodes. The analysis of this case turns out to be much more 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. Our analysis of this construction is tight, as uniformity testing on forests can be achieved with $O(n)$ samples. We believe that a super-linear lower bound would be very interesting, but also quite difficult to obtain. Proving our linear lower bound already required a very careful analysis for a relatively simple construction, and an improved lower bound would require analyzing a distribution over dense constructions, for which an improved structural understanding is needed. A further technical discussion of this lower bound is in Section 1.1.2, see Section 8 and Theorem 19 for a formal statement and full analysis of our main lower bound. As mentioned before, we also show that the sample complexity must depend on β and h in certain cases, see Theorem 20 for a formal statement.

Table 1 summarizes our algorithmic results.

The High-Dimensional Frontier and Related Work: We emphasize that we believe the study of high-dimensional distribution testing to be of significant importance, as real-world applications often involve multivariate data. As univariate distribution testing is now very well understood, with a thorough set of tools and techniques, this is the natural next frontier to attack. However, multivariate distributions pose several new technical challenges, and many of these univariate tools are rendered obsolete – as such, we must extend these methods, or introduce new techniques entirely. It is important to develop approaches which may be applicable in much more general high-dimensional distribution testing settings, when there may be complex correlations between random variables. First, it is important to get a grasp on the concentration and variance of statistics in these settings, and we provide exposition of a technique for bounding the variance of some simple statistics. Additionally, our linear lower bound's construction and analysis give insight into which instances cause intractability to arise, and provide a recipe for the style of combinatorics required to analyze them.

In further works, the authors and another group have investigated more properties of multilinear functions over the Ising model [DDK17, GLP17]. In the present work, we require and prove variance bounds for bilinear functions of the Ising model. These other works prove *concentration* bounds (which are qualitatively stronger than variance bounds) for multilinear functions of arbitrary degree d (rather than just bilinear functions, which are of degree $d = 2$).

High-dimensional distribution testing has recently attracted the interest of the theoretical computer science community, with work concurrent to ours on testing Bayes networks[4] [CDKS17, DP17].

[4] Bayes nets are another type of graphical model, and are in general incomparable to Ising models.

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 under Dobrushin/high-temperature using Learn-Then-Test	$\tilde{O}\left(\frac{n^{10/3} \beta^2}{\varepsilon^2}\right)$	$\tilde{O}\left(\frac{n^{10/3} \beta^2}{\varepsilon^2}\right)$
Identity under Dobrushin/high-temperature using Learn-Then-Test	$\tilde{O}\left(\frac{n^{10/3} \beta^2}{\varepsilon^2}\right)$	$\tilde{O}\left(\frac{n^{11/3} \beta^2}{\varepsilon^2} + \frac{n^{5/3} h^2}{\varepsilon^2}\right)$
Independence on Forests using Improved Localization	$\tilde{O}\left(\frac{n}{\varepsilon}\right)$	$\tilde{O}\left(\frac{n^2 \beta^2}{\varepsilon^2}\right)$
Identity on Forests using Improved Localization	$\tilde{O}\left(\frac{n \cdot c(\beta)}{\varepsilon}\right)$	$\tilde{O}\left(\frac{n^2 \beta^2}{\varepsilon^2} + \frac{n^2 h^2}{\varepsilon^2}\right)$
Independence on Ferromagnets using Improved Localization	$\tilde{O}\left(\frac{n d_{\max}}{\varepsilon}\right)$	$\tilde{O}\left(\frac{n^2 d_{\max}^2 \beta^2}{\varepsilon^2}\right)$

Table 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, h = maximum absolute value of node parameters (when applicable), and c is a function discussed in Theorem 4.

It remains to be seen which other multivariate distribution classes of interest allow us to bypass the curse of dimensionality.

We note that the paradigm of distribution testing under structural assumptions has been explored in the univariate setting, where we may assume the distribution is log-concave or k-modal. This often allows exponential savings in the sample complexity [DDS$^+$13, DKN15b, DKN15a].

1.1 Further Technical Discussion and Highlights

In this section, we give a slightly more in-depth discussion of some of the technical highlights of our work. For full details and more discussion, the interested reader can refer to the corresponding sections in the body of the paper.

1.1.1 Structural Results for Ferromagnetic Ising Models

Our general-purpose testing algorithm is a localization-based algorithm – in particular, it operates based on the structural property that if two Ising models (with no external field) are far from each other, they will have a distant edge marginal. We convert this structural property to an algorithm by estimating each edge marginal and testing whether they match for the two models. However, the underlying structural property is quantitatively weak, and leads to sub-optimal testing bounds. In some cases of interest, we can derive quantitatively stronger versions of this structural result, giving us more efficient algorithms.

For instance, one can consider the ferromagnetic case, where one has all edge parameters $\theta_e \geq 0$. We would like to derive a relationship between an edge marginal (i.e., $\mathbf{E}[X_u X_v]$ for an edge $e = (u, v)$) and the parameter on that edge θ_e. For a tree-structured Ising model with no external field (ferromagnetic or not), it is not hard to show that $\mathbf{E}[X_u X_v] = \tanh(\theta_e)$ – for small edge parameters, this indicates a linear relationship between the edge marginal and the edge parameter. Intuitively, if a model is ferromagnetic and contains cycles, these cycles should only increase the correlation between adjacent nodes, i.e., we would expect that $\mathbf{E}[X_u X_v] \geq \tanh(\theta_e)$. While this is true, it proves surprisingly difficult to prove directly, and we must instead view the Ising model through the Fortuin-Kastelyn random cluster model.

At a high level, the Fortuin-Kastelyn random cluster model is defined for a graph $G = (V, E)$ with a probability parameter $0 < r_e < 1$ on each edge. This parameter indicates the probability of a bond being present on edge e (i.e., the distribution gives a measure over $\{0, 1\}^E$), placing this model into the space of bond percolation models (see Section 4.2.1 and (21) for the formal definition). It turns out that an alternative way to draw a sample from the Ising model is through this random cluster model. Namely, we first draw a sample from the Fortuin-Kastelyn model (defined with appropriate parameters), and for each connected component in the resulting graph, we flip a fair coin to determine whether all the nodes in the component should be -1 or $+1$.

With this correspondence in hand, we can apply results for the Fortuin-Kastelyn model – crucial for our purposes is that the fact that the FK model's measure is stochastically increasing. Roughly, this means that if we increase the values of the r_e's, the probability of an edge having a 1 can only increase. Intuitively, this leads to an increase in $\mathbf{E}[X_u X_v]$ in the Ising model, since it increases the probability that the nodes are connected in the FK model, and thus the expectation of any edge can only increase as we increase the ferromagnetic edge parameters. Careful work is needed to carry through the implications of this correspondence, but it allows us to conclude the nearly-optimal sample complexity of $\tilde{O}(m/\varepsilon)$.

Full details are provided in Section 4.2.

1.1.2 A Linear Lower Bound for Testing Ising Models

As a starting point for our lower bound, we use Le Cam's classical two-point method. This is the textbook method for proving lower bounds in distribution testing. It involves defining two families of distributions $\mathcal{P}$ and $\mathcal{Q}$, such that every distribution $p \in \mathcal{P}$ is ε-far from every distribution $q \in \mathcal{Q}$. We consider selecting a uniformly random pair $(p, q) \in (\mathcal{P}, \mathcal{Q})$ and then drawing k independent samples from each of p and q. If we can show that the resulting two transcripts of k samples are close in total variation distance, then k samples are insufficient to distinguish these two cases.

While this method is fairly well-understood in the univariate setting, it proves more difficult to apply in some multivariate settings. This difficulty arises in the definition of the set $\mathcal{Q}$[5]. In the univariate setting, we often decompose the domain into several disjoint sets, and define $\mathcal{Q}$ by applying perturbations to each of these sets independently. This style of construction allows us to analyze each subset locally and compose the results. In the multivariate setting, constructions of this local nature are still possible and are not too hard to analyze – see Theorem 18. In this construction, we consider an Ising model defined by taking a fixed perfect matching on the graph and selecting a distribution from $\mathcal{Q}$ by applying a random sign vector to the edge potentials of this matching. This allows us to prove an $\Omega(\sqrt{n})$ lower bound on the complexity of uniformity testing.

However, such local constructions prove to be limited in the multivariate setting. In order to prove stronger lower bounds, we instead must consider an Ising model generated by taking a *random*

[5] We note that for simplicity, $\mathcal{P}$ is often chosen to be a singleton.

perfect matching on the graph. This construction is more global in nature, since the presence of an edge gives us information about the presence of other edges in the graph. As a result, the calculations no longer decompose elegantly over the (known) edges in the matching. While at a first glance, the structure of such a construction may seem too complex to analyze, we reduce it to analyzing the structure of a random pair of matchings by exploiting combinatorial symmetries. An important step in the proof requires us to understand the random variable representing the number of edges shared by two random perfect matchings. This analysis allows us to prove a quadratically-better lower bound of $\Omega(n)$. We believe our analysis may be useful in proving lower bounds for such global constructions in other multivariate settings.

Full details are provided in Section 8.

1.2 Organization

In Section 2, we discuss preliminaries and the notation that we use throughout the paper. In Section 3, we give a simple localization-based algorithm for independence testing and its corresponding variant for goodness-of-fit testing. In Section 4, we present improvements to our localization-based algorithms for forest-structured and ferromagnetic Ising models. In Section 5, we describe our main algorithm for the high-temperature regime which uses a global statistic on the Ising model. In Section 6, we compare our algorithms from Sections 3 and 5. In Section 7, we discuss the bounds in [LPW09] and apply them to bounding the variance of bilinear statistics over the Ising model. In Section 8, we describe our lower bounds.

2 Preliminaries

Recall the definition of the Ising model from Eq. (1). 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 paper we will use the convention that $E = \{(u, v) \mid u, v \in V, u \neq v\}$ 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 paper, 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 $\{\pm 1\}^n$.

In this paper, we will consider *Rademacher* random variables, where *Rademacher*(p) takes value 1 with probability p, and -1 otherwise.

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

Definition 2. *In the* ferromagnetic *setting,* $\theta_e \geq 0$ *for all* $e \in E$.

Definition 3 (Dobrushin's Uniqueness Condition)**.** *Consider an Ising model p defined on a graph $G = (V, E)$ with $|V| = n$ and parameter vector $\vec{\theta}$. Suppose* $\max_{v \in V} \sum_{u \neq v} \tanh(|\theta_{uv}|) \leq 1 - \eta$ *for some constant $\eta > 0$. Then p is said to satisfy Dobrushin's uniqueness condition, or be in the high temperature regime. Note that since* $\tanh(|x|) \leq |x|$ *for all x, the above condition follows from more simplified conditions which avoid having to deal with hyperbolic functions. For instance, either of the following two conditions:*

$$\max_{v \in V} \sum_{u \neq v} |\theta_{uv}| \leq 1 - \eta \textit{ or}$$
$$\beta d_{\max} \leq 1 - \eta$$

are sufficient to imply Dobrushin's condition (where $\beta = \max_{u,v} |\theta_{uv}|$ and $d_{\max}$ is the maximum degree of G).

In general, when one refers to the temperature of an Ising model, a high temperature corresponds to small θ_e values, and a low temperature corresponds to large θ_e values. In this paper, we will only use the precise definition as given in Definition 3.

Remark 1. *We note that high-temperature is not strictly needed for our results to hold – we only need Hamming contraction of the "greedy coupling." This condition implies rapid mixing of the Glauber dynamics (in $O(n \log n)$ steps) via path coupling (Theorem 15.1 of [LPW09]). See [DDK17, GLP17] for further discussion of this weaker condition.*

Lipschitz functions of the Ising model have the following variance bound, which is in Chatterjee's thesis [Cha05]:

Lemma 1 (Lipschitz Concentration Lemma)**.** *Suppose that $f(X_1, \ldots, X_n)$ is a function of an Ising model in the high-temperature regime. Suppose the Lipschitz constants of f are $l_1, l_2, \ldots, l_n$ respectively. That is,*

$$\left| f(X_1, \ldots, X_i, \ldots, X_n) - f(X_1, \ldots, X_i', \ldots, X_n) \right| \leq l_i$$

for all values of $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ and for any X_i and X_i'. Then for some absolute constant c_1,

$$\Pr\left[|f(X) - \mathbf{E}[f(X)]| > t\right] \leq 2 \exp\left(-\frac{c_1 t^2}{2 \sum_{i=1}^n l_i^2}\right).$$

In particular, for some absolute constant c_2,

$$\mathbf{Var}(f(X)) \leq c_2 \sum_i l_i^2.$$

We will use the symmetric KL divergence, defined as follows:

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

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,*

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

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

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

Hence the symmetric KL-divergence between two distributions upper bounds both the KL-divergence and total variation (TV) distance between them under appropriate scaling. Therefore, our results which hold for testing with respect to the SKL-divergence also hold for testing with respect to KL-divergence and TV distance.

We will use the following folklore result on estimating the parameter of a Rademacher random variable.

Lemma 3. *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 $1 - \delta$.*

In Section 7 we use the Glauber dynamics on the Ising model. Glauber dynamics is the canonical Markov chain for sampling from an Ising model. We consider the basic variant known as single-site Glauber dynamics. The dynamics are a Markov chain defined on the set Σ^n where $\Sigma = \{\pm 1\}$. 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:

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

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

2.1 Input to Goodness-of-Fit Testing Algorithms

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 paper. 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 $\{\mu_{uv} = \mathbf{E}[X_u X_v]\}, \{\mu_u = \mathbf{E}[X_u]\}$.

2.2 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 [SW12]:

$$d_{\text{SKL}}(p,q) = \sum_{v \in V} \left(\theta_v^p - \theta_v^q\right)\left(\mu_v^p - \mu_v^q\right) + \sum_{e=(u,v) \in E} \left(\theta_e^p - \theta_e^q\right)\left(\mu_e^p - \mu_e^q\right). \tag{3}$$

This expression will form the basis for all our algorithms.

3 A General Purpose 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 4. *Given two Ising models p and q, if $d_{\text{SKL}}(p,q) \geq \varepsilon$, then either*

- *There exists an edge $e = (u,v)$ such that $\left(\theta_{uv}^p - \theta_{uv}^q\right)\left(\mu_{uv}^p - \mu_{uv}^q\right) \geq \frac{\varepsilon}{2m}$; or*
- *There exists a node u such that $\left(\theta_u^p - \theta_u^q\right)\left(\mu_u^p - \mu_u^q\right) \geq \frac{\varepsilon}{2n}$.*

Proof of Lemma 4: We have,

$$\begin{aligned} d_{\text{SKL}}(p,q) &= \sum_{e=(u,v) \in E} \left(\theta_e^p - \theta_e^q\right)\left(\mu_e^p - \mu_e^q\right) + \sum_{v \in V} \left(\theta_v^p - \theta_v^q\right)\left(\mu_v^p - \mu_v^q\right) \geq \varepsilon \\ \implies \sum_{e=(u,v) \in E} &\left(\theta_e^p - \theta_e^q\right)\left(\mu_e^p - \mu_e^q\right) \geq \varepsilon/2 \quad \text{or} \quad \sum_{v \in V} \left(\theta_v^p - \theta_v^q\right)\left(\mu_v^p - \mu_v^q\right) \geq \varepsilon/2 \end{aligned}$$

In the first case, there has to exist an edge $e = (u,v)$ such that $\left(\theta_{uv}^p - \theta_{uv}^q\right)\left(\mu_{uv}^p - \mu_{uv}^q\right) \geq \frac{\varepsilon}{2m}$ and in the second case there has to exist a node u such that $\left(\theta_u^p - \theta_u^q\right)\left(\mu_u^p - \mu_u^q\right) \geq \frac{\varepsilon}{2n}$ 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 $d_{\text{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$ 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 $d_{\text{SKL}}(p,q) \geq \varepsilon$ with probability at least $2/3$ where $\beta = \max\{|\theta_{uv}|\}$ and $h = \max\{|\theta_u|\}$. The above algorithms assume that m, an upper bound on the number of edges, is known. If no upper bound is known, we may use the trivial upper bound of $\binom{n}{2}$. 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 $\lambda_{uv} = \mathbf{E}[X_u X_v] - \mathbf{E}[X_u]\mathbf{E}[X_v]$ 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 4 and is described in further detail in the proof of Lemma 6. A precise description of the test is given in in Algorithm 1 and its correctness is proven via Lemmas 5 and 6. We note that this algorithm is phrased as if an upper bound on the number of edges m is known. If we instead know an upper bound on the maximum degree $d_{\max}$, then we can replace m by $nd_{\max}$.

Algorithm 1 Test if an Ising model p is product

```
function LocalizationTest(sample access to Ising model $p$, accuracy parameter $\varepsilon, \beta, m$)
    Draw $k = O\left(\frac{m^2\beta^2 \log n}{\varepsilon^2}\right)$ 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 $(u, v)$, $\left|\hat{\lambda}_{uv}\right| \geq \frac{\varepsilon}{4m\beta}$ return that $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$
    Otherwise, return that $p \in \mathcal{I}_n$
end function
```

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 5. *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 $\lambda_{uv} = \mathbf{E}[X_u X_v] - \mathbf{E}[X_u]\mathbf{E}[X_v]$. Namely, the algorithm computes the empirical value of $|\lambda_{uv}|$ for each pair of nodes and identifies pairs such that this value is sufficiently far from* 0.

Proof. This lemma is a direct consequence of Lemma 3. Note that for any edge $e = (u, v) \in E$, $X_u X_v \sim \mathit{Rademacher}((1+\mu_e)/2)$. Also $X_u \sim \mathit{Rademacher}((1+\mu_u)/2)$ and $X_v \sim \mathit{Rademacher}((1+\mu_v)/2)$. We will use Lemma 3 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 3 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 by triangle inequality 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 $|\lambda_{uv} - \hat{\lambda}_{uv}| \leq \frac{4\varepsilon}{10}$. 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 6. *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. This will involve using the algorithm of Lemma 5 with parameter "ε" as $\varepsilon/2\beta m$. 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 (3), 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

$$\sum_{v \in V} \left(\theta_v^p - \theta_v^q\right)\left(\mu_v^p - \mu_v^q\right) = 0 \tag{4}$$

$$\therefore d_{\mathrm{SKL}}(p, q) \geq \varepsilon$$

$$\implies \exists e^* = (u, v) \text{ s.t } \left(\theta_e^p - \theta_e^q\right)\left(\mu_e^p - \mu_e^q\right) \geq \frac{\varepsilon}{m} \tag{5}$$

$$\implies \exists e^* = (u, v) \text{ s.t } |(\mu_e^p - \mu_e^q)| \geq \frac{\varepsilon}{2\beta m} \tag{6}$$

$$\implies \exists e^* = (u, v) \text{ s.t } |\lambda_{e^*}| \geq \frac{\varepsilon}{2\beta m}.$$

where (4) follows because $\mu_v^p = \mu_v^q$ for all $v \in V$, (5) follows from Lemma 4 and (6) 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 5 as stated in Lemma 7 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 7. *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 7: The proof follows along the same lines as Lemma 5. Let $X \sim p$. Then, for any pair of nodes (u, v), $X_u X_v \sim \mathit{Rademacher}((1 + \mu_e^p)/2)$. Also $X_u \sim \mathit{Rademacher}((1 + \mu_u^p)/2)$ for any node u. For any pair of nodes u, v, with $O(\log n/\varepsilon^2)$ samples, Lemma 3 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 $\varepsilon$,$\beta$,$h$,$m$)
    Draw $k = O\left(\frac{(m^2\beta^2+n^2h^2)\log n}{\varepsilon^2}\right)$ samples from $p$. 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 $(u, v)$, $|\hat{\mu}_{uv}^p - \mu_{uv}^q| \geq \frac{\varepsilon}{8m\beta}$ return that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$
    If for any node $u$, if $|\hat{\mu}_u^p - \mu_u^q| \geq \frac{\varepsilon}{8nh}$ 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 7. We omit the proof here.

4 Improved Tests for Forests and Ferromagnetic Ising Models

In this section we will describe testing algorithms for two commonly studied classes of Ising models, namely forests and ferromagnets. In these cases, the sample complexity improves compared to the baseline result when in the regime of no external field. The testers are still localization based (like those of Section 3), but we can now leverage structural properties to obtain more efficient testers.

First, we consider the class of all forest structured Ising models, where the underlying graph $G = (V, E)$ is a forest. Such models exhibit nice structural properties which can be exploited to obtain more efficient tests. In particular, under no external field, the edge marginals μ_e, which, in general are hard to compute, have a simple closed form expression. This structural information enables us to improve our testing algorithms from Section 3 on forest graphs. We state the improved sample complexities here and defer a detailed description of the algorithms to Section 4.1.

Theorem 3 (Independence testing of Forest-Structured Ising Models)**.** *Algorithm 3 takes in $\tilde{O}\left(\frac{n}{\varepsilon}\right)$ samples from an Ising model $X \sim p$ whose underlying graph is a forest and which is under no external field and outputs whether $p \in \mathcal{I}_n$ or $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$ with probability $\geq 9/10$.*

Remark 2. *Note that Theorem 3 together with our lower bound described in Theorem 19 indicate a tight sample complexity up to logarithmic factors for independence testing on forest-structured Ising models under no external field.*

Theorem 4 (Identity Testing of Forest-Structured Ising Models)**.** *Algorithm 4 takes in the edge parameters of an Ising model q on a forest graph and under no external field as input, and draws $\tilde{O}\left(c(\beta)\frac{n}{\varepsilon}\right)$ samples from an Ising model $X \sim p$ (where $c(\beta)$ is a function of the parameter β) whose underlying graph is a forest and under no external field, and outputs whether $p = q$ or $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$ with probability $\geq 9/10$.*

Note that for identity testing, any algorithm necessarily has to have at least a β dependence due to the lower bound we show in Theorem 20.

The second class of Ising models we consider this section are ferromagnets. For a ferromagnetic Ising model, $\theta_{uv} \geq 0$ for every pair of nodes u, v. Ferromagnets may potentially contain cycles but since all interactions are ferromagnetic, the marginal of every edge is at least what it would have been if it was a solo edge. This intuitive property turns out to be surprisingly difficult to prove in a direct way. We prove this structural property using an alternative view of the Ising model density which comes from the Fortuin-Kasteleyn random cluster model. Using this structural property, we give a quadratic improvement in the dependence on parameter m for testing independence under no external field. We state our main result in this regime here and a full description of the algorithm and the structural lemma are provided in Section 4.2.

Theorem 5 (Independence Testing of Ferromagnetic Ising Models)**.** *Algorithm 5 takes in $\tilde{O}\left(\frac{nd_{\max}}{\varepsilon}\right)$ samples from a ferromagnetic Ising model $X \sim p$ which is under no external field and outputs whether $p \in \mathcal{I}_n$ or $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$ with probability $\geq 9/10$.*

4.1 Improved Algorithms for Independence and Identity Testing on Forests

Before we present the improved algorithms, we will prove the following fact about the edge marginals of an arbitrary Ising model with no external field where the underlying graph is a forest. This result was known prior to this work by the community but we couldn't find a proof of the same, hence we provide our own proof of the lemma.

Lemma 8 (Structural Lemma for Forest-Structured Ising Models)**.** *If p is an Ising model on a forest graph with no external field, and $X \sim p$, then for any $(u, v) \in E$, we have*

$$\mathbf{E}\left[X_u X_v\right] = \tanh(\theta_{uv}).$$

Proof. Consider any edge $e = (u, v) \in E$. Consider the tree (T, E_T) which contains e. Let n_T be the number of nodes in the tree. We partition the vertex set T into U and V as follows. Remove edge e from the graph and let U denote all the vertices which lie in the connected component of node u except u itself. Similarly, let V denote all the vertices which lie in the connected component of node v except node v itself. Hence, $T = U \cup V \cup \{u\} \cup \{v\}$. Let X_U be the vector random variable which denotes the assignment of values in $\{\pm 1\}^{|U|}$ to the nodes in U. X_V is defined similarly. We will also denote a specific value assignment to a set of nodes S by x_S and $-x_S$ denotes the assignment which corresponds to multiplying each coordinate of x_S by -1. Now we state the following claim which follows from the tree structure of the Ising model.

Claim 1. $\Pr\left[X_U = x_U, X_u = 1, X_v = 1, X_V = x_V\right] = \exp(2\theta_{uv}) \Pr\left[X_U = x_U, X_u = 1, X_v = -1, X_V = -x_V\right].$

In particular the above claim implies the following corollary which is obtained by marginalization of the probability to nodes u and v.

Corollary 1. *If X is an Ising model on a forest graph $G = (V, E)$ with no external field, then for any edge $e = (u, v) \in E$, $\Pr\left[X_u = 1, X_v = 1\right] = \exp(2\theta_{uv}) \Pr\left[X_u = 1, X_v = -1\right]$.*

Now,

$$\begin{aligned}
\mathbf{E}\left[X_uX_v\right] &= \Pr\left[X_uX_v = 1\right] - \Pr\left[X_uX_v = -1\right] \qquad &(7)\\
&= 2Pr\left[X_u = 1, X_v = 1\right] - 2\Pr\left[X_u = 1, X_v = -1\right] &(8)\\
&= \frac{2Pr\left[X_u = 1, X_v = 1\right] - 2\Pr\left[X_u = 1, X_v = -1\right]}{2Pr\left[X_u = 1, X_v = 1\right] + 2\Pr\left[X_u = 1, X_v = -1\right]} &(9)\\
&= \frac{Pr\left[X_u = 1, X_v = 1\right] - \Pr\left[X_u = 1, X_v = -1\right]}{Pr\left[X_u = 1, X_v = 1\right] + \Pr\left[X_u = 1, X_v = -1\right]} &(10)\\
&= \left(\frac{\exp(2\theta_{uv}) - 1}{\exp(2\theta_{uv}) + 1}\right)\frac{\Pr\left[X_u = 1, X_v = -1\right]}{\Pr\left[X_u = 1, X_v = -1\right]} &(11)\\
&= \tanh(\theta_{uv}) &(12)
\end{aligned}$$

where (8) follows because $\Pr\left[X_u = 1, X_v = 1\right] = \Pr\left[X_u = -1, X_v = -1\right]$ and $\Pr\left[X_u = -1, X_v = 1\right] = \Pr\left[X_u = 1, X_v = -1\right]$ by symmetry. Line (9) divides the expression by the total probability which is 1 and (11) follows from Corollary 1.

□

Given the above structural lemma, we give the following simple algorithm for testing independence on forest Ising models under no external field.

Algorithm 3 Test if a forest Ising model p under no external field is product

```
function TestForestIsing-Product(sample access to Ising model $p$)
    Run the algorithm of Lemma 5 to identify all edges $e = (u, v)$ such that $|\mathbf{E}[X_uX_v]| \geq \sqrt{\frac{\varepsilon}{n}}$.
    using $\tilde{O}\left(\frac{n}{\varepsilon}\right)$ samples. If it identifies any edges, return that $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$
    Otherwise, return that $p$ is product.
end function
```

Algorithm 3, at a high level, works as follows. If there is an edge parameter whose absolute value is larger than a certain threshold, it will be easy to detect due to the structural information about the edge marginals. In case all edges have parameters smaller in absolute value than this threshold, the expression for $d_{\mathrm{SKL}}(.,.)$ between two Ising models tells us that there still has to be at least one edge with a significantly large value of μ_e in case the model is far from uniform, and hence will still be detectable by the algorithm of Lemma 5. The proof of Theorem 3 shows this formally.

Proof of Theorem 3: Firstly, note that under no external field, the only product Ising model is the uniform distribution $\mathcal{U}_n$. Therefore the problem reduces to testing whether p is uniform or not. Consider the case when p is indeed uniform. That is, there are no edges in the underlying graph of the Ising model. In this case with probability at least 9/10 the localization algorithm of Lemma 5 will output no edges. Hence Algorithm 3 will output that p is uniform.
In case $d_{\mathrm{SKL}}(p, \mathcal{U}_n) \geq \varepsilon$, we split the analysis into two cases.

- *Case 1:* There exists an edge $e = (u, v)$ such that $|\theta_{uv}| \geq \sqrt{\frac{\varepsilon}{n}}$. In this case, $\mathbf{E}[X_uX_v] = \tanh(\theta_{uv})$ and in the regime where $|\theta| = o(1)$, $|\tanh(\theta)| \geq |\theta/2|$. Hence implying that $|\mathbf{E}[X_uX_v]| \geq |\theta_{uv}/2| \geq \left|\sqrt{\frac{\varepsilon}{n}}/2\right|$. Therefore the localization algorithm of Lemma 5 would identify such an edge with probability at least 9/10. Note that the regime where the inequality $|\tanh(\theta)| \geq |\theta/2|$ isn't valid is easily detectable using $\tilde{O}(\frac{n}{\varepsilon})$ samples, as this would imply that $|\theta| \geq 1.9$ and $|\mathbf{E}[X_uX_v]| \geq 0.95$.

- *Case 2:* All edges $e = (u, v)$ are such that $|\theta_{uv}| \le \left|\sqrt{\frac{\varepsilon}{n}}\right|$. In this case we have,

$$\begin{aligned} d_{\mathrm{SKL}}(p, \mathcal{U}_n) &\ge \varepsilon & (13)\\ \implies \exists \text{ edge } e = (u, v) \text{ s.t } \theta_{uv}\mathbf{E}[X_u X_v] &\ge \frac{\varepsilon}{n} & (14)\\ \implies \exists \text{ edge } e = (u, v) \text{ s.t } |\mathbf{E}[X_u X_v]| &\ge \left|\frac{\varepsilon}{n} \times \sqrt{\frac{n}{\varepsilon}}\right| & (15)\\ &= \sqrt{\frac{\varepsilon}{n}} & (16) \end{aligned}$$

  Hence, the localization algorithm of Lemma 5 would identify such an edge with probability at least 9/10.

□

Next, we will present an algorithm for identity testing on forest Ising models under no external field.

Algorithm 4 Test if a forest Ising model p under no external field is identical to a given Ising model q

```
function TestForestIsing-Identity(Ising model $q$,sample access to Ising model $p$)
    If the Ising model $q$ is not a forest, or has a non-zero external field on some node, return. $d_{\mathrm{SKL}}(p, q) \ge \varepsilon$
    Run the algorithm of Lemma 5 to identify all edges $e = (u, v)$ such that. $|\mathbf{E}[X_u X_v] - \tanh(\theta^q_{uv})| \ge \sqrt{\frac{\varepsilon}{n}}$ using $\tilde{O}\left(\frac{n}{\varepsilon}\right)$ samples. If it identifies any edges, return that $d_{\mathrm{SKL}}(p, q) \ge \varepsilon$
    Otherwise, return that $p = q$.
end function
```

Proof of Theorem 4: Consider the case when p is indeed q. In this case with probability at least 9/10 the localization algorithm of Lemma 5 will output no edges. Hence Algorithm 4 will output that p is uniform.
In case $d_{\mathrm{SKL}}(p, q) \ge \varepsilon$, we split the analysis into two cases.

- *Case 1:* There exists an edge $e = (u, v)$ such that $|\theta^p_{uv} - \theta^q_{uv}| \ge \sqrt{\frac{\varepsilon}{n}}$. In this case, $\mathbf{E}[X_u X_v] - \mu^q_{uv} = \tanh(\theta^p_{uv}) - \tanh(\theta^q_{uv})$ and hence has the same sign as $\theta^p_{uv} - \theta^q_{uv}$. Assume that $\theta^p_{uv} \ge \theta^q_{uv}$. The argument for the case $\theta^q_{uv} > \theta^p_{uv}$ will follow similarly. If $\theta^p_{uv} - \theta^q_{uv} \le 1/2 \tanh(\beta)$, then the following inequality holds from Taylor's theorem.

$$\tanh(\theta^p_{uv}) - \tanh(\theta^q_{uv}) \ge \frac{\mathrm{sech}^2(\beta)\,(\theta^p_{uv} - \theta^q_{uv})}{2}$$

  which would imply $\tanh(\theta^p_{uv}) - \tanh(\theta^q_{uv}) \ge \frac{\mathrm{sech}^2(\beta)}{2}\sqrt{\frac{\varepsilon}{n}}$ and hence the localization algorithm of Lemma 5 would identify edge e with probability at least 9/10 using $\tilde{O}\left(\frac{c_1(\beta)n}{\varepsilon}\right)$ samples (where $c_1(\beta) = \cosh^4(\beta)$). If $\theta^p_{uv} - \theta^q_{uv} > 1/2\tanh(\beta)$, then $\tanh(\theta^p_{uv}) - \tanh(\theta^q_{uv}) \ge \tanh(\beta) - \tanh\left(\beta - \frac{1}{2\tanh(\beta)}\right)$ and hence the localization algorithm of Lemma 5 would identify edge e with probability at least 9/10 using $\tilde{O}\left(c_2(\beta)\right)$ samples where $c_2(\beta) = \frac{1}{(\tanh(\beta) - \tanh(\beta - 1/2\tanh(\beta)))^2}$.

Note that as β grows small, $c_2(\beta)$ gets worse. However it cannot grow unbounded as we also have to satisfy the constraint that $\theta^p_{uv} - \theta^q_{uv} \leq 2\beta$. This implies that

$$c_2(\beta) = \min\left\{\beta^2, \frac{1}{(\tanh(\beta) - \tanh(\beta - 1/2\tanh(\beta)))^2}\right\}$$

samples suffice in this case. Therefore the algorithm will give the correct output with probability $> 9/10$ using $\tilde{O}\left(c(\beta)\frac{n}{\varepsilon}\right)$ samples where $c(\beta) = \max\{c_1(\beta), c_2(\beta)\}$.

- *Case 2:* All edges $e = (u, v)$ are such that $|\theta^q_{uv} - \theta^q_{uv}| \leq \sqrt{\frac{\varepsilon}{n}}$. In this case we have,

$$\begin{aligned} d_{\mathrm{SKL}}(p,q) &\geq \varepsilon & (17)\\ \implies \exists \text{ edge } e=(u,v) \text{ s.t } \left(\theta^p_{uv} - \theta^q_{uv}\right)\left(\mathbf{E}[X_u X_v] - \mu^q_{uv}\right) &\geq \frac{\varepsilon}{n} & (18)\\ \implies \exists \text{ edge } e=(u,v) \text{ s.t } \left|\mathbf{E}[X_u X_v] - \mu^q_{uv}\right| &\geq \left|\frac{\varepsilon}{n} \times \sqrt{\frac{n}{\varepsilon}}\right| & (19)\\ &= \sqrt{\frac{\varepsilon}{n}} & (20) \end{aligned}$$

Hence, the localization algorithm of Lemma 5 would identify such an edge with probability at least 9/10.

□

4.2 Ferromagnetic Ising Models: A Structural Understanding and an Improved Independence Test

In this section we will describe an algorithm for testing independence of ferromagnetic Ising models under no external field. The tester follows the localization based recipe of Section 3 but leverages additional structural information about ferromagnets to obtain an improved sample complexity.

At a high level, the algorithm is as follows: if there exists an edge with a large edge parameter, then we lower bound its marginal by $\tanh(\theta_{uv})$ where uv is the edge under consideration. This implies that its marginal sticks out and is easy to catch via performing local tests on all edges. If all the edge parameters were small, then Algorithm 1 is already efficient.

We first prove a structural lemma about ferromagnetic Ising models. We will use the Fortuin-Kasteleyn random cluster model and its coupling with the Ising model (described in Chapter 10 of [RAS15]) to argue that in any ferromagnetic Ising model $\mu_{uv} \geq \tanh(\theta_{uv})$ for all pairs u, v.

4.2.1 Random Cluster Model

Let $G = (V, E)$ be a finite graph. The random cluster measure is a probability distribution on the space $\Omega = \{0,1\}^E$ of bond configurations denoted by $\eta = (\eta(e))_{e \in E} \in \{0,1\}^E$. Each edge has an associated bond $\eta(e)$. $\eta(e) = 1$ denotes that bond e is open or present and $\eta(e) = 0$ implies that bond e is closed or unavailable. A random cluster measure is parameterized by an edge probability $0 < r < 1$ and by a second parameter $0 < s < \infty$. Let $k(\eta)$ denote the number of connected components in the graph (V, η). The random cluster measure is defined by

$$\rho_{r,s}(\eta) = \frac{1}{Z_{r,s}}\left(\prod_{e \in E} r^{\eta(e)}(1-r)^{1-\eta(e)}\right)s^{k(\eta)}$$

where $Z_{r,s}$ is a normalizing factor to make ρ a probability density. We consider a generalization of the random cluster model where each edge is allowed to have its own parameter $0 < r_e < 1$. Under this generalization, the measure becomes

$$\rho_{\vec{r},s}(\eta) = \frac{1}{Z_{\vec{r},s}} \left(\prod_{e \in E} r_e^{\eta(e)} (1 - r_e)^{1-\eta(e)} \right) s^{k(\eta)}. \tag{21}$$

The random cluster measure is stochastically increasing in $\vec{r}$ when $s \geq 1$. This property is formally stated in Lemma 10.3 of [RAS15]. We state a generalized version of the Lemma here which holds when each edge is allowed its own probability parameter r_e.

Lemma 9. *[Lemma 10.3 from [RAS15]] For $s \geq 1$, and $\vec{r_1} \leq \vec{r_2}$ coordinate-wise, $\rho_{\vec{r_1},s} \leq \rho_{\vec{r_2},s}$ where given two bond configurations η_1 and η_2, $\eta_1 \geq \eta_2$ iff $\eta_1(e) = 1$ for all e such that $\eta_2(e) = 1$.*

4.2.2 Coupling between the Random Cluster Model and the Ising model

We will now describe a coupling between the random cluster measure and the probability density function for a ferromagnetic Ising model. In particular, the edge probability r_e under the random cluster measure and the edge parameters θ_e of the Ising model are related by

$$r_e = 1 - \exp(-2\theta_e)$$

and the parameter $s = 2$ because the Ising model has two spins ± 1. The coupling Q will be a joint distribution on the spin variables $X = (X_1 \ldots X_n)$ of the Ising model and the bond variables $\eta = (\eta(e))_{e \in E}$. The measure Q is defined as

$$Q(X, \eta) = \frac{1}{Z} \prod_{e=(u,v) \in E} r_e^{\eta(e)} (1 - r_e)^{1-\eta(e)} \left(\mathbb{1}_{X_u = X_v} + (1 - \eta(e)) \mathbb{1}_{X_u \neq X_v} \right)$$

where Z is a normalizing constant so as to make Q a probability measure. Under the relation stated above between r_e and θ_e, the following properties regarding the marginal distributions of Q hold.

$$\begin{aligned} \sum_{\eta \in \{0,1\}^E} Q(X, \eta) &= \frac{1}{Z'} \exp\left(\sum_{u \neq v} \theta_{uv} X_u X_v \right) \\ \sum_{X \in \{\pm 1\}^n} Q(X, \eta) &= \frac{1}{Z''} \left(\prod_{e \in E} r_e^{\eta(e)} (1 - r_e)^{1-\eta(e)} \right) 2^{k(\eta)} = \rho_{\vec{r},2}(\eta) \end{aligned} \tag{22}$$

where Z', Z'' are normalizing constants to make the marginals probability densities. The above equations imply that the measure Q is a valid coupling and more importantly they yield an alternative way to sample from the Ising model as follows:

First sample a bond configuration η according to $\rho_{\vec{r},2}(\eta)$. For each connected component in the bond graph, flip a fair coin to determine if the variables in that component will be all $+1$ or all -1.

In addition to the above information about the marginals of Q, we will need the following simple observations.

1. $Q(X, \eta) = 0$ if $\eta(e) = 1$ for any $e \notin E$.
2. $Q(X, \eta) = 0$ if for any $e = (u, v) \in E$, $\eta(e) = 1$ and $X_u \neq X_v$.

Next we state another property of the coupling $Q(.,.)$ which says that if two nodes u and v are in different connected components in the bond graph specified by η, then the probability that $X_u = X_v$ is the same as the probability that $X_u \neq X_v$.

Claim 2. *Let $C_\eta(u, v)$ denote the predicate that under the bond configuration η, u and v are connected with a path of open bonds. Then,*

$$\sum_{\substack{\eta \ s.t \\ C_\eta(u,v)=0}} \sum_{\substack{X \ s.t. \\ X_u=X_v}} Q(X, \eta) = \sum_{\substack{\eta \ s.t \\ C_\eta(u,v)=0}} \sum_{\substack{X \ s.t. \\ X_u \neq X_v}} Q(X, \eta)$$

The proof of the above claim is quite simple and follows by matching the appropriate terms in the probability density Q when u and v lie in different connected components. The proof is omitted here.

Armed with the coupling Q and its properties stated above, we are now ready to state the main structural lemma we show for ferromagnetic Ising models.

Lemma 10. *Consider two ferromagnetic Ising models p and q under no external field defined on $G_p = (V, E_p)$ and $G_q = (V, E_q)$. Denote the parameter vector of p model by $\vec{\theta}^p$ and that of q model by $\vec{\theta}^q$. If $\vec{\theta}^p \geq \vec{\theta}^q$ coordinate-wise, then for any two nodes $u, v \in V$, $\mu^p_{uv} \geq \mu^q_{uv}$.*

Proof. Since

$$\mu^p_{uv} = \Pr_p [X_u = X_v] - \Pr_p [X_u \neq X_v]$$
$$\implies \mu^p_{uv} = 2 \Pr_p [X_u = X_v] - 1$$

to show that $\mu^p_{uv} \geq \mu^q_{uv}$ it suffices to show that $\Pr_p [X_u = X_v] \geq \Pr_q [X_u = X_v]$. Consider the coupling $Q(X, \eta)$ described above between the random cluster measure and the Ising model probability. $\Pr_p [X_u = X_v]$ can be expressed in terms of $Q_p(X, \eta)$ as follows:

$$\Pr_p [X_u = X_v] = \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} \sum_{\eta} Q_p(X, \eta)$$

Denote the sum on the right in the above equation by S_p. It suffices to show that $S_p \geq S_q$.

Lemma 10.3 of [RAS15] gives that for any bond configuration η_0,

$$\sum_{\eta \geq \eta_0} \rho_p^{E_b}(\eta) \geq \sum_{\eta \geq \eta_0} \rho_q^{E_b}(\eta).$$

This follows because the parameter vectors of p and q satisfy the condition of the lemma that $\vec{\theta}^p \geq \vec{\theta}^q$. Again, let $C_\eta(u, v)$ denote the predicate that under the bond configuration η, u and v are connected. Let H be the set of all bond configurations such that u and v are connected by a single distinct path. Therefore $C_{\eta_0}(u, v) = 1$ for all $\eta_0 \in H$. Then the set

$$C = \{\eta | \eta \geq \eta_0 \text{ for some } \eta_0 \in H\}$$

represents precisely the bond configurations in which u and v are connected. Applying Lemma 10.3 of [RAS15] on each $\eta_0 \in H$ and summing up the inequalities obtained, we get

$$\begin{aligned}
&\sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \rho_p^{E_b}(\eta) \geq \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \rho_q^{E_b}(\eta) \\
\implies &\sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_X Q_p(X,\eta) \geq \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_X Q_q(X,\eta) \\
\implies &\sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q_p(X,\eta) \geq \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q_q(X,\eta) \qquad (23)
\end{aligned}$$

where the last inequality follows because $Q(X,\eta) = 0$ if for any pair u, v, $\eta(uv) = 1$ but $X_u \neq X_v$.

Also, from Claim 2, we have that for any Ising model,

$$\sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=0}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q(X,\eta) = \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=0}} \sum_{\substack{X \text{ s.t.} \\ X_u\neq X_v}} Q(X,\eta) \qquad (24)$$

And since $Q(.,.)$ is a probability measure we have that for any Ising model,

$$\begin{aligned}
&\sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q(X,\eta) + \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=0}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q(X,\eta) + \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=0}} \sum_{\substack{X \text{ s.t.} \\ X_u\neq X_v}} Q(X,\eta) + \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_{\substack{X \text{ s.t.} \\ X_u\neq X_v}} Q(X,\eta) = 1 \quad (25) \\
\implies &\sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q(X,\eta) + \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=0}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q(X,\eta) + \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=0}} \sum_{\substack{X \text{ s.t.} \\ X_u\neq X_v}} Q(X,\eta) = 1 \quad (26) \\
\implies &\sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q(X,\eta) + 2 \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=0}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q(X,\eta) = 1 \quad (27)
\end{aligned}$$

where (26) follows because the last term in (25) is 0 and (27) follows from (24).

Equation (27) implies that

$$\begin{aligned}
S_p &= \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q_p(X,\eta) + \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=0}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q_p(X,\eta) \\
&= \frac{1}{2} \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q_p(X,\eta) + \frac{1}{2}
\end{aligned}$$

Therefore from (23), we get

$$S_p \geq S_q$$

□

Using the above lemma, we now prove the main structural lemma for ferromagnets which will be crucial to our algorithm for testing ferromagnetic Ising models.

Lemma 11 (Structural Lemma about Ferromagnetic Ising Models)**.** *If $X \sim p$ is a ferromagnetic Ising model on a graph $G = (V,E)$ under zero external field, then $\mu_{uv} \geq \tanh(\theta_{uv})$ for all edges $(u,v) \in E$.*

Proof. Fix the edge of concern $e = (u, v)$. If the graph doesn't contain cycles, then from Lemma 8 $\mu_{uv} = \tanh(\theta_{uv})$ and the statement is true. To show that the statement holds for general graphs we will use induction on the structure of the graph. Graph G can be constructed as follows. Start with the single edge $e = (u, v)$ and then add the remaining edges in $E \backslash \{e\}$ one by one in some order. Denote the intermediate graphs obtained during this process as $G_0, G_1, \ldots, G_m = G$ where G_0 is the graph consisting of just a single edge. For each graph G_i we can associate the corresponding Ising model p_i to be the model which has $\theta_e^{p_i} = \theta_e$ for $e \in E_{G_i}$ and $\theta_e^{p_i} = 0$ otherwise. For each graph G_i in the sequence, we will use $\mu_{uv}^{p_i}$ to denote $\mathbf{E}\left[X_u X_v\right]$ for the Ising model corresponding to graph G_i. We will prove that $\mu_{uv}^{p} \geq \tanh(\theta_{uv})$ by induction on this sequence of graphs. The statement can be easily verified to be true for G_0. In fact, $\mu_{uv}^{p_0} = \tanh(\theta_{uv})$. Suppose the statement was true for some G_i in the sequence. By Lemma 10, we have that $\mu_{uv}^{p_{i+1}} \geq \mu_{uv}^{p_i}$. This implies that $\mu_{uv}^{Gpi+1} \geq \tanh(\theta_{uv})$ hence showing the statement to be true for all graphs G_i in the sequence. □

Given the above structural lemma about ferromagnetic Ising models under no external field, we present the following algorithm for testing whether a ferromagnetic Ising model is product or not.

Algorithm 5 Test if a ferromagnetic Ising model p under no external field is product

```
function TestFerroIsing-Independence(sample access to an Ising model $p$)
    Run the algorithm of Lemma 5 to identify if all edges $e = (u, v)$ such that $\mathbf{E}[X_u X_v] \geq \sqrt{\varepsilon}/n$.
    using $\tilde{O}\left(\frac{n^2}{\varepsilon}\right)$ samples. If it identifies any edges, return that $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$
    Otherwise, return that $p$ is product.
end function
```

Proof of Theorem 5: Firstly, note that under no external field, the only product Ising model is the uniform distribution $\mathcal{U}_n$. To the problem reduces to testing whether p is uniform or not. Consider the case when p is indeed uniform. That is, there are no edges in the underlying graph of the Ising model. In this case with probability at least 9/10 the localization algorithm of Lemma 5 with output no edges. Hence Algorithm 5 will output that p is product.
In case $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$, we split the analysis into two cases.

- *Case 1:* There exists an edge $e = (u, v)$ such that $|\theta_{uv}| \geq \sqrt{\frac{\varepsilon}{n^2}}$. In this case, $|\mathbf{E}[X_u X_v]| \geq |\tanh(\theta_{uv})|$ and in the regime where ε is a fixed constant, $|\tanh(\theta)| \geq |\theta/2|$. Hence implying that $|\mathbf{E}[X_u X_v]| \geq |\theta_{uv}/2| \geq \sqrt{\frac{\varepsilon}{n^2}}/2$. Therefore the localization algorithm of Lemma 5 would identify such an edge with probability at least 9/10. (The regime where the inequality $|\tanh(\theta)| \geq |\theta/2|$ isn't valid would be easily detectable using $\tilde{O}(\frac{n^2}{\varepsilon})$ samples.)
- *Case 2:* All edges $e = (u, v)$ are such that $\theta_{uv} \leq \sqrt{\frac{\varepsilon}{n^2}}$. In this case we have,

$$\begin{aligned} d_{\mathrm{SKL}}(X, \mathcal{I}_n) &\geq \varepsilon && (28) \\ \implies \exists \text{ edge } e = (u, v) \text{ s.t } \theta_{uv}\mathbf{E}[X_u X_v] &\geq \frac{\varepsilon}{n^2} && (29) \\ \implies \exists \text{ edge } e = (u, v) \text{ s.t } \mathbf{E}[X_u X_v] &\geq \frac{\varepsilon}{n^2} \times \sqrt{\frac{n^2}{\varepsilon}} && (30) \\ &= \sqrt{\frac{\varepsilon}{n^2}} && (31) \end{aligned}$$

Hence, the localization algorithm of Lemma 5 would identify such an edge with probability at least 9/10.

□

5 An Improved Test for High-Temperature Ising Models: A Learn-then-Test Algorithm

In this section, 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 Section 3 for dense graphs. This is the more technically involved part of our paper and we modularize the description and analysis into different parts. We will give a high level overview of our approach here.

The main approach we take in this section is to consider a global test statistic over all the variables on the Ising model in contrast to the localized statistics of Section 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 5.5.

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 (3), we have that under no external field ($\theta_u = 0$ for all $u \in V$),

$$\begin{aligned} d_{\mathrm{SKL}}(p, \mathcal{U}_n) &= \sum_{e=(u,v)\in E} \theta_{uv}\mu_{uv} \\ \implies d_{\mathrm{SKL}}(p, \mathcal{U}_n) &\leq \sum_{u\neq v} \beta \left|\mu_{uv}\right| \qquad (32) \\ \implies \frac{d_{\mathrm{SKL}}(p, \mathcal{U}_n)}{\beta} &\leq \sum_{u\neq v} \left|\mu_{uv}\right|. \qquad (33) \end{aligned}$$

where (32) 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 (33), 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

$d_{\mathrm{SKL}}(p,\mathcal{U}_n) \geq \varepsilon$, $\mathbf{E}[\sum_{e=(u,v)\in E}\Gamma_e\,(X_uX_v)] \geq \varepsilon/\zeta\beta$, where $\zeta = \operatorname{poly}(n)$. 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_uX_v). Naively, we could get an appropriate Γ_e for (X_uX_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:

$$Z' = \sum_{u\neq v} c_{uv}X_uX_v \tag{34}$$

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 16).

$$\mathbf{Var}(Z') = \tilde{O}(n^2).$$

In other words, despite the potentially complex structure of the Ising model and potential correlations, the variables X_uX_v contribute to the variance of Z' roughly as if they were all independent! We describe the result and techniques involved in the analysis of the variance bound in Section 7. 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 5.1-5.4.

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 5.5.

Theorem 6 (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 $\tilde{O}\left(\frac{n^{10/3}\beta^2}{\varepsilon^2}\right)$ 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 6 with sample complexities we obtain when β is close to the high temperature threshold.

Theorem 7 (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 $\tilde{O}\left(\frac{n^{10/3}}{\varepsilon^2 d_{\max}^2}\right)$ 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 1 shows the dependence of sample complexity of testing as $d_{\max}$ is varied in the regime of Theorem 7 for the case of no external field.

The description of our algorithm is presented in Algorithm 6. It contains a parameter τ, which we choose to be the value achieving the minimum in the sample complexity of Theorem 8. The algorithm follows a learn-then-test framework, which we outline here.

Algorithm 6 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 7 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 14 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 6 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 8. When optimized for the parameter τ, this yields Theorem 6.

Theorem 8. *Given $\tilde{O}\left(\min_{\tau>0}\left(n^{2+\tau}+n^{6-2\tau}\right)\frac{\beta^2}{\varepsilon^2}\right)$ 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 section is as follows. We describe and analyze our weak learning procedure in Section 5.1. Given a vector with the appropriate weak learning guarantees, we describe and analyze the testing procedure in Section 5.2. In Section 5.3, 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 8. Finally, in Section 5.4, we optimize the sample complexities from Theorem 8 for the parameter τ and filter out cleaner statement of Theorem 6. 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.

5.1 Weak Learning

Our overall goal of this section is "weakly learn" the sign of $\mu_e = \mathbf{E}[X_u X_v]$ 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 3), 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 20 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 7.

Algorithm 7 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 5.1.1. We invoke the weak-learning procedure of Algorithm 7 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 12). 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 13). 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.

5.1.1 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

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

We start by proving that the weak learning algorithm 7 yields a $\vec{\Gamma}$ for which such a bound holds in expectation. The following is fairly straightforward from Lemma 20 and linearity of expectations.

Lemma 12. *Given $k = O\left(\frac{n^{2\tau_2}\beta^2}{\varepsilon^2}\right)$ samples from an Ising model $X \sim p$ such that $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$ and $|\mu_e| \leq \frac{\varepsilon}{\beta n^{\tau_2}}$ for all $e \in E$, Algorithm 7 outputs $\vec{\Gamma} = \{\Gamma_e\} \in \{\pm 1\}^{|E|}$ such that*

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

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 20 (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 13. *Given $k = O\left(\frac{n^{2\tau_2}\beta^2}{\varepsilon^2}\right)$ i.i.d. samples from an Ising model p such that $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$ and $|\mu_e| \leq \frac{\varepsilon}{\beta n^{\tau_2}}$ for all $e \in E$, Algorithm 7 outputs $\vec{\Gamma} = \{\Gamma_e\} \in \{\pm 1\}^{|E|}$ where: Define χ_{τ_2} to be the event that*

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

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 $r = \mathbf{Pr}_{\Gamma}\left[\chi_{\tau_2}\right]$. 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

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

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

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

Then we have

$$\begin{aligned} rU + (1-r)L &\geq 2T \text{ (From Lemma 12)} \\ \implies r &\geq \frac{2T - L}{U - L} \end{aligned}$$

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

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

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

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

Since $L \leq T$, we get

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

Substituting in the value for T we get

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

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

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

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

as desired. □

5.2 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}, which says that the expectation the statistic with this vector is sufficiently large). With such a $\vec{\Gamma}$, we show that we can distinguish between $p \in \mathcal{I}_n$ and $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$.

Lemma 14. *Let p be an Ising model, let $X \sim p$, 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 $k = O\left(\sigma^2 \cdot \frac{n^{4-2\tau_2}\beta^2 \log(1/\delta)}{\varepsilon^2}\right)$ i.i.d samples from p, which satisfies either $p \in \mathcal{I}_n$ or $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$, and $\vec{\Gamma} = \{\Gamma_e\} \in \{\pm 1\}^{|E|}$ which satisfies χ_{τ_2} (as defined in Lemma 13) in the case that $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$, 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 $p \in \mathcal{I}_n$. Otherwise, it will output that $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$.

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

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

If $p \in \mathcal{I}_n$, 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 $d_{\mathrm{SKL}}(p, \mathcal{I}_n) \geq \varepsilon$ follows similarly, but additionally using the fact that χ_{τ_2} implies that

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

□

5.3 Combining Learning and Testing

In this section, we combine lemmas from the previous sections to complete the proof of Theorem 8. Lemma 13 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 14 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 5, 13, and 14, 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.

$$O\left(\frac{n^{2+\tau_2}\beta^2}{\varepsilon^2}\right) + O\left(\sigma^2 \frac{n^{4-2\tau_2}\beta^2}{\varepsilon^2} \log n\right). \tag{35}$$

Plugging in the variance bounds from Section 7, Theorems 16 and 17 gives Theorem 8.

5.4 Balancing Weak Learning and Testing

The sample complexities in the statement of Theorem 8 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 3 which can be easily verified and arrive at Theorem 6.

Claim 3. *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 $\frac{2+s}{3}$.*

Claim 3 together with the variance bound (Theorem 16) implies Theorem 6.

Theorem 6 (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 $\tilde{O}\left(\frac{n^{10/3}\beta^2}{\varepsilon^2}\right)$ 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.

5.5 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 5.1-5.4 to obtain testers for independence under an arbitrary external field (Section 5.5.1), identity without an external field (Section 5.5.2), and identity under an external field (Section 5.5.3).

5.5.1 Independence Testing under an External Field

Under an external field, the statistic we considered in Section 5 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_u^q is such that $\mu_u^q = \mu_u^p$ for all $u \in V$. Then,

$$\begin{aligned} d_{\mathrm{SKL}}(p, \mathcal{I}_n) &\leq d_{\mathrm{SKL}}(p, q) \\ &= \sum_{e=(u,v)\in E} \theta_{uv}^p \left(\mu_{uv}^p - \mu_{uv}^q\right) \\ &= \sum_{e=(u,v)\in E} \theta_{uv}^p \left(\mu_{uv}^p - \mu_u^p \mu_v^p\right) \\ &\leq \sum_{e=(u,v)\in E} \beta \left|\mu_{uv}^p - \mu_u^p \mu_v^p\right| \\ \implies \frac{d_{\mathrm{SKL}}(p, \mathcal{I}_n)}{\beta} &\leq \sum_{e=(u,v)\in E} \left|\mu_{uv}^p - \mu_u^p \mu_v^p\right| . \end{aligned} \tag{36}$$

The above inequality suggests a statistic Z such that $\mathbf{E}[Z] = \sum_{e=(u,v)\in E} |\lambda_{uv}^p|$ where $\lambda_{uv}^p = \mu_{uv}^p - \mu_u^p \mu_v^p$. We consider $Z = \sum_{u \neq v} \mathbf{sign}(\lambda_{uv}) \left(X_u^{(1)} - X_u^{(2)}\right)\left(X_v^{(1)} - X_v^{(2)}\right)$ 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: $Z_{uv} = \frac{1}{4}\left(X_u^{(1)} - X_u^{(2)}\right)\left(X_v^{(1)} - X_v^{(2)}\right)$. 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 $Z'_{uv} \sim$ *Rademacher* $\left(\frac{1}{2} + \frac{\lambda_{uv}}{4}\right)$ and by Lemma 20 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 12 and 13 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:

$$Z'_{cen} = \sum_{u \neq v} c_{uv} \left(X_u^{(1)} - X_u^{(2)}\right)\left(X_v^{(1)} - X_v^{(2)}\right) \tag{37}$$

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 techniques described in Section 7 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 17 and the details are in Section 7.3. Combining the weak learning part and the variance bound gives us the following sample complexity for independence testing under an external field:

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

Balancing for the optimal value of the τ parameter gives Theorem 9.

Theorem 9 (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 $\tilde{O}\left(\frac{n^{10/3}\beta^2}{\varepsilon^2}\right)$ 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 8.

Algorithm 8 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 7 on $S = \{(X_u^{(1)} - X_u^{(2)})(X_v^{(1)} - X_v^{(2)})\}_{u \neq v}$ with param-. eters $\tau_2 = \tau$ 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 14 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
```

5.5.2 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,

$$d_{\mathrm{SKL}}(p, q) = \sum_{(u,v) \in E} \left(\theta_{uv}^p - \theta_{uv}^q\right)\left(\mu_{uv}^p - \mu_{uv}^q\right)$$

$$\implies \frac{d_{\mathrm{SKL}}(p, q)}{2\beta} \leq \sum_{u \neq v} \left|\left(\mu_{uv}^p - \mu_{uv}^q\right)\right|$$

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

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

If $p = q$, $\mathbf{E}[Z] = 0$ and if $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$, $\mathbf{E}[Z] \geq \varepsilon/2\beta$. 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

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

where $c \in \{\pm 1\}^{\binom{V}{2}}$. Hence, the variance bound of Theorem 16 holds for $\mathbf{Var}(Z)$.

As for the weakly learning the signs, using Corollary 2 of Lemma 20 we get that for each pair u, v, with k samples, we can achieve a success probability $1/2 + c_1\sqrt{k}\,|\mu_{uv}^p - \mu_{uv}^q|$ of correctly estimating $\mathbf{sign}(\mu_{uv}^p - \mu_{uv}^q)$. Following this up with analogous proofs of Lemmas 12 and 13 where μ_e is replaced by $\mu_e^p - \mu_e^q$, 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 10 (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 $\tilde{O}\left(\frac{n^{10/3}\beta^2}{\varepsilon^2}\right)$ 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 9.

Algorithm 9 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 7 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 14 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
```

5.5.3 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, it is unclear how to define an appropriately centered statistic which has a variance of $O(n^2)$ in this setting, and we consider this an interesting open question. Instead, we use the uncentered statistic which has variance $\Theta(n^3)$.

We now 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 $\sum_{v \in V} |(\mu^p_v - \mu^q_v)| = 0$ 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 $Z = \sum_{v \in V} \mathbf{sign}(\mu_v^p) \left(X_v - \mu_v^q\right)$. 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 5.1-5.4 to weakly learn a sign vector correlated with the true sign vector. In particular, since $X_v \sim \textit{Rademacher}(1/2{+}\mu_v/2)$, from Corollary 2, 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 1, 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 5.3 and 5.4, 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 5.5.2:

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

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 Lemma 1 to get a variance bound of $O(n^3)$ which yields a sample complexity of $\tilde{O}\left(\frac{n^{11/3}\beta^2}{\varepsilon^2}\right)$ for this setting.
Theorem 11 captures the total sample complexity of our identity tester under the presence of external fields.

Theorem 11 (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 10.

6 Comparing Localization and Learn-then-Test Algorithms

At this point, we now have two algorithms: the localization algorithm of Section 3 and the learn-then-test algorithm of Section 5. Both algorithms are applicable in all temperature regimes

Algorithm 10 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 7 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 14 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 7 on $S = \{(X_u X_v - Y_{uv}\}_{u \neq v}$, 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 $\mathbf{sign}\left(\mathbf{E}\left[X_u X_v - \mu_{uv}^q\right]\right)$
    end for
    Using the same set of samples for all $\ell$, run the testing algorithm of Lemma 14 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
```

but learn-then-test beats localization's sample complexity in high temperature under some degree regimes. We note that their sample complexities differ in their dependence on β and $d_{\max}$. In this section, 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, we note that if the algorithm is agnostic of the maximum degree $d_{\max}$, then learn-then-test always outperforms localization in the high temperature regime. This leads to Theorem 12.

Theorem 12. *Suppose p is an Ising model in the high temperature regime. To test either independence or identity agnostic of the maximum degree of the graph $d_{\max}$, localization requires $\tilde{O}\left(\frac{n^4\beta^2}{\varepsilon^2}\right)$ samples from p for a success probability $> 2/3$. Learn-then-test, on the other hand, requires $\tilde{O}\left(\frac{n^{10/3}\beta^2}{\varepsilon^2}\right)$ for independence testing and identity testing under no external field. It requires $\tilde{O}\left(\frac{n^{11/3}\beta^2}{\varepsilon^2}\right)$ for identity testing under an external field.*

When knowledge of $d_{\max}$ is available to the tester, he can improve his sample complexities of localization approach. Now the sample complexity of localization gets worse as $d_{\max}$ increases. As noted in Section 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 Section 7 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 13 (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 $d_{\max} \leq n^{2/3}$, 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 2 for the regime when $\beta = O(n^{-2/3})$.

The comparison for independence testing under the presence of an external field is similar and is presented in Theorem 14.

Theorem 14 (Best Sample Complexity Achieved for Independence Testing, Arbitrary External Field)**.** *Suppose p is an Ising model under **an arbitrary external field**.*

- *if $\beta = O(n^{-2/3})$, then for the range $d_{\max} \leq n^{2/3}$, localization performs better, for independence testing. For the range $n^{2/3} \leq d_{\max} \leq \frac{1}{4\beta}$, learn-then-test performs better than localization for independence 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 independence testing.*

Finally, we note in Theorem 15, the parameter regimes when learn-then-test performs better for identity testing under an external field. Here our learn-then-test approach suffers worse bounds due to a weaker bound on the variance of our statistic.

Theorem 15 (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.*

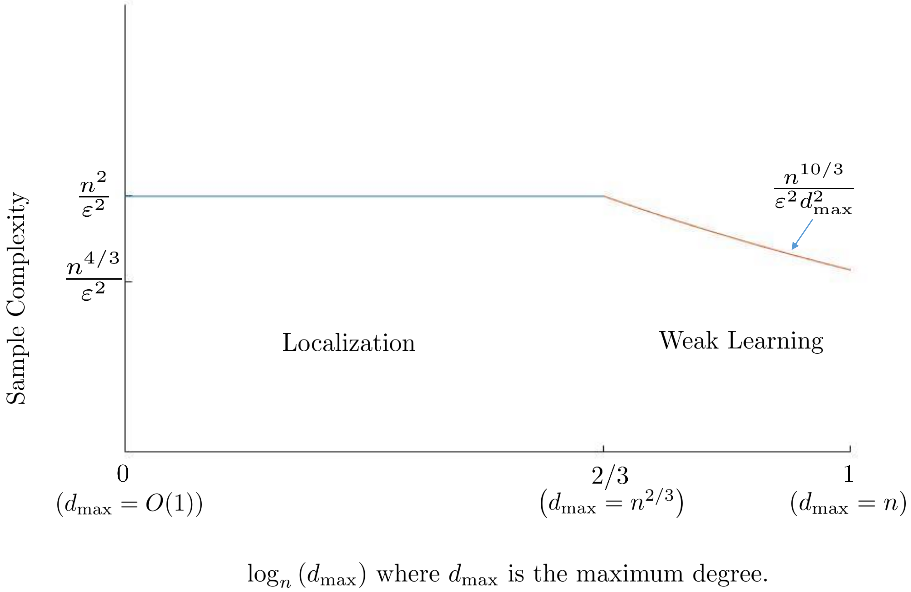

Figure 1: Localization vs Learn-Then-Test: A plot of the sample complexity of testing identity under no external field when $\beta = \frac{1}{4d_{\max}}$ is close to the threshold of high temperature. Note that throughout the range of values of $d_{\max}$ we are in high temperature regime in this plot.

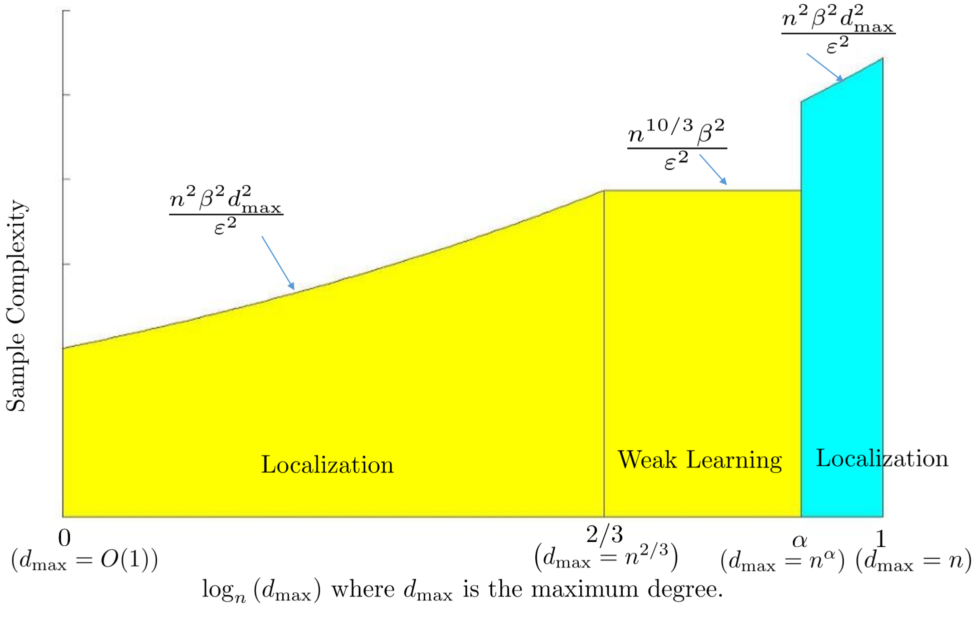

Figure 2: Localization vs Learn-Then-Test: A plot of the sample complexity of testing identity under no external field when $\beta \le n^{-2/3}$. The regions shaded yellow denote the high temperature regime while the region shaded blue denotes the low temperature regime. The algorithm which achieves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

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

In this section, we describe a 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 Section 5, we must bound the variance of statistics of the form $Z' = \sum_{u \neq v} c_{uv} X_u X_v$ (under no external field, see (34)) and $Z'_{cen} = \sum_{u \neq v} c_{uv} \left(X_u^{(1)} - X_u^{(2)}\right)\left(X_v^{(1)} - X_v^{(2)}\right)$ (under an external field, see (37)). While the variance for both the statistics is easily seen to be $O(n^2)$ if the graph has no edges, to prove variance bounds better than the trivial $O(n^4)$ for general graphs requires some work. We show the following two theorems in this section.

The first result, Theorem 16, 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 Section 5 in the absence of an external field.

Theorem 16 (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 3. Then*

$$\mathbf{Var}\left(f_c(X)\right) = \tilde{O}(n^2).$$

The second result of this section, Theorem 17, bounds the variance of functions of the form $\sum_{u \neq v} c_{uv}(X_u^{(1)} - X_u^{(2)})(X_v^{(1)} - X_v^{(2)})$ which captures the statistic of interest for independence testing using the learn-then-test framework of Section 5 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 17 (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 3 and define $g_c : \{\pm 1\}^{V \cup V'} \to \mathbb{R}$ as follows: $g_c(x) = \sum_{\substack{u,v \in V \\ s.t.\ u \neq v}} c_{uv}(x_{u^{(1)}} - x_{u^{(2)}})(x_{v^{(1)}} - x_{v^{(2)}})$. Then*

$$\mathbf{Var}(g_c(X)) = \tilde{O}\left(n^2\right).$$

7.1 Overview of the Technique

We will use tools from Chapter 13 of [LPW09] to obtain the variance bounds of this section. At a high level the technique to bound the variance of a function f on a distribution μ involves first defining a reversible Markov chain with μ as its stationary distribution. By studying the mixing time properties (via the spectral gap) of this Markov chain along with the second moment of the variation of f when a single step is taken under this Markov chain we obtain bounds on the second moment of f which consequently yield the desired variance bounds.

The Markov chain in consideration here will be the Glauber dynamics chain on the Ising model p. As stated in Section 2, the Glauber dynamics are reversible and ergodic for Ising models. Let M be the reversible transition matrix for the Glauber dynamics on some Ising model p. Let γ_* be the absolute spectral gap for this Markov chain. The first step is to obtain a lower bound on γ_*.

Claim 4. *In the high-temperature regime/under Dobrushin conditions, $\gamma_* \geq \Omega\left(\frac{1}{n \log n}\right)$ under an arbitrary external field.*

Proof. From Theorem 15.1 of [LPW09], we have that the mixing time of the Glauber dynamics is $O(n \log n)$. Since the Glauber dynamics on an Ising model is ergodic and reversible, using the relation between mixing and relaxation times (Theorem 12.4 of [LPW09]) we get that

$$\begin{aligned} t_{mix} &\geq \left(\tfrac{1}{\gamma_*} - 1\right)\log(2) && (38) \\ \implies \frac{1}{\gamma_*} &\leq \tfrac{n \log n}{\log(2)} + 1 && (39) \\ \implies \gamma_* &\geq \Omega\left(\tfrac{1}{n \log n}\right). && (40) \end{aligned}$$

□

For a function f, define

$$\mathcal{E}(f) = \frac{1}{2} \sum_{x,y \in \{\pm 1\}^n} [f(x) - f(y)]^2 \pi(x) M(x,y).$$

This can be interpreted as the expected square of the difference in the function, when a step is taken at stationarity. That is,

$$\mathcal{E}(f) = \frac{1}{2}\mathbf{E}\left[(f(x) - f(y))^2\right] \tag{41}$$

where x is drawn from the Ising distribution and y is obtained by taking a step in the Glauber dynamics starting from x. We now state a slight variant of Remark 13.13 which we will use as Lemma 15.

Lemma 15. *For a reversible transition matrix P on state space Ω with stationary distribution π, let*

$$\mathcal{E}(f) := \frac{1}{2} \sum_{x,y \in \Omega} (f(x) - f(y))^2 \pi(x) P(x,y),$$

where f is a function on Ω such that $\mathbf{Var}_\pi(f) > 0$. Also let γ_* be the absolute spectral gap of P. Then

$$\gamma_* \leq \frac{\mathcal{E}(f)}{\mathbf{Var}_\pi(f)}.$$

Note: Remark 13.13 in [LPW09] states a bound on the spectral gap as opposed to the absolute spectral gap bound which we use here. However, the proof of Remark 13.13 also works for obtaining a bound on the absolute spectral gap γ_*.

7.2 Bounding Variance of Z' Under No External Field

We prove Theorem 16 now. Consider the function $f_c(x) = \sum_{u,v} c_{uv} x_u x_v$ where $c \in [-1,1]^{\binom{|V|}{2}}$.

Claim 5. *For an Ising model under no external field, $\mathcal{E}(f_c) = \widetilde{O}(n)$.*

Proof. Since y is obtained by taking a single step on the Glauber dynamics from x, $f_c(x)-f_c(y)$ is a function of the form $\sum_v c_v x_v$ where $c_v \in [-1,1]$ for all $v \in V$. The coefficients $\{c_v\}_v$ depend on which node $v_0 \in V$ was updated by the Glauber dynamics. Since there are n choices for nodes to update, and since the update might also leave x unchanged, i.e. $y = x$, $f_c(x)-f_c(y)$ is one of at most $n+1$ linear functions of the form $\sum_v c_v x_v$. Denote, by $E(x,y)$, the event that $|f_c(x)-f_c(y)| \geq c\sqrt{n}\log n$. We have, from the concentration of linear functions on the Ising model around their expected value (Lemma 1) and a union bound over the $n+1$ possible linear functions, that for a sufficiently large c, under no external field, $\Pr\left[E(x,y)\right] \leq \frac{1}{10n^2}$. Now,

$$\begin{aligned}\mathbf{E}\left[(f_c(x)-f_c(y))^2\right] &= \mathbf{E}\left[(f_c(x)-f_c(y))^2 | E(x,y)\right]\Pr[E(x,y)] \\ &\quad + \mathbf{E}\left[(f_c(x)-f_c(y))^2 | \neg E(x,y)\right]\Pr[\neg E(x,y)] \\ &\leq n^2 \times \tfrac{1}{10n^2} + c^2 n \log^2 n\left(1-\tfrac{1}{10n^2}\right) \\ &= \widetilde{O}(n)\end{aligned}$$

where we used the fact that the absolute maximum value of $(f_c(x)-f_c(y))^2$ is n^2. □

Claim 4 together with Claim 5 are sufficient to conclude an upper bound on the variance of f_c, by using Lemma 15, thus giving us Theorem 16.

7.3 Bounding Variance of Z'_{cen} Under Arbitrary External Field

Under the presence of an external field, we saw that we need to appropriately center our statistics to achieve low variance. The function $g_c(x)$ of interest now is defined over the 2-sample Ising model $p^{\otimes 2}$ and is of the form

$$g_c(x) = \sum_{u,v} c_{uv}(x_u^{(1)} - x_u^{(2)})(x_v^{(1)} - x_v^{(2)})$$

where now $x, y \in \{\pm 1\}^{2|V|}$. First, note that the absolute spectral gap for $p^{\otimes 2}$ is also at least $\widetilde{\Omega}(1/n)$. Now we bound $\mathcal{E}(g_c)$.

Claim 6. *For an Ising model under an arbitrary external field, $\mathcal{E}(g_c) = \widetilde{O}(n)$.*

Proof. Since y is obtained by taking a single step on the Glauber dynamics from x, it can be seen that $g_c(x)-g_c(y)$ is a function of the form $\sum_v c_v\left(x_v^{(1)} - x_v^{(2)}\right)$ where $c_v \in [-1,1]$ for all $v \in V$. The coefficients $\{c_v\}_v$ depend on which node $v_0 \in V$ was updated by the Glauber dynamics. Since there are n choices for nodes to update, and since the update might also leave x unchanged, i.e. $y = x$, $g_c(x) - g_c(y)$ is one of at most $n+1$ linear functions of the form $\sum_v c_v\left(x_v^{(1)} - x_v^{(2)}\right)$. Denote, by $E(x,y)$, the event that $|g_c(x)-g_c(y)| \geq c\sqrt{n}\log n$. We have, from Lemma 1 and a union bound, that for a sufficiently large c, $\Pr\left[E(x,y)\right] \leq \frac{1}{10n^2}$. Now,

$$\begin{aligned}\mathbf{E}\left[(g_c(x)-g_c(y))^2\right] &= \mathbf{E}\left[(g_c(x)-g_c(y))^2 | E(x,y)\right]\Pr[E(x,y)] && (42)\\ &\quad + \mathbf{E}\left[(g_c(x)-g_c(y))^2 | E(x,y)^c\right]\Pr[E(x,y)^c] && (43)\\ &\leq 4n^2 \times \tfrac{1}{10n^2} + c^2 n \log^2 n\left(1-\tfrac{1}{10n^2}\right) && (44)\\ &= \widetilde{O}(n) && (45)\end{aligned}$$

where we used the fact that the absolute maximum value of $(g_c(x)-g_c(y))^2$ is $4n^2$. □

Similar to before, Claim 4 together with Claim 6 are sufficient to conclude an upper bound on the variance of f_c, by using Lemma 15, thus giving us Theorem 17.

8 Lower Bounds

In this section we describe our lower bound constructions and state the main results.

8.1 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 18. *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 $d_{\text{SKL}}(p, \mathcal{U}_n) \geq \varepsilon$ 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 19. *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 $d_{\text{SKL}}(p, \mathcal{U}_n) \geq \varepsilon$ 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 [Pan08]. 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 8.3.1.

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* perfect 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 8.3.2.

Remark 3. *Similar lower bound constructions to those of Theorems 18 and 19 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.*

8.2 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 20.

Theorem 20. *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 8.3.3.

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. Notably, we show that testing independence of a forest-structured Ising model under no external field can be done using $\tilde{O}\left(\frac{n}{\varepsilon}\right)$ samples (Theorem 3).

8.3 Lower Bound Proofs

8.3.1 Proof of Theorem 18

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 $M = (M_1, \ldots, M_n) \in \mathcal{M}$, 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 [AD15], which follows from Pinsker's and Jensen's inequalities:

Lemma 16. *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.

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

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

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

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 18.

8.3.2 Proof of Theorem 19

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 following 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 16 to $2d_{\mathrm{TV}}^2(p^{\otimes k}, \bar{q}^{\otimes k})$ 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 (i.e., combine all the elements from S_1 and S_2, keeping duplicates). Let M_0 be the perfect 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 perfect 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 perfect matching. Writing these ideas mathematically, the expression above is equal to

$$2^{nk} \sum_{X=(X^{(1)},\dots,X^{(k)})} \left(\frac{1}{(n-1)!!} \sum_{M\in\mathcal{M}} \prod_{(u,v)\in M} \prod_{i=1}^{k} \left(\frac{e^\delta}{2(e^\delta+e^{-\delta})}\right)^{\mathbb{1}(X_u^{(i)}=X_v^{(i)})} \left(\frac{e^{-\delta}}{2(e^\delta+e^{-\delta})}\right)^{\mathbb{1}(X_u^{(i)}\neq X_v^{(i)})} \right)^2$$

$$= 2^{nk} \sum_{X=(X^{(1)},\dots,X^{(k)})} \left(\frac{1}{(n-1)!!} \sum_{M\in\mathcal{M}} \prod_{(u,v)\in M} \left(\frac{1}{2(e^\delta+e^{-\delta})}\right)^{k} e^{\delta(k-H(X_u,X_v))} e^{-\delta H(X_u,X_v)} \right)^2$$

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

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

$$= \left(\frac{e^\delta}{e^\delta+e^{-\delta}}\right)^{nk} \frac{1}{(n-1)!!^2} \sum_{X=(X^{(1)},\dots,X^{(k)})} \sum_{M_1,M_2\in\mathcal{M}} \prod_{(u,v)\in M_1\uplus M_2} \exp(-2\delta H(X_u,X_v))$$

At this point, we note that if we fix the matching M_1, summing over all perfect 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.

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

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

We observe that multiset union of two perfect matchings will form a collection of even length cycles (if they contain the same edge, this forms a 2-cycle), and this can be rewritten as follows.

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

$$= \left(\frac{e^\delta}{e^\delta+e^{-\delta}}\right)^{nk} \frac{1}{(n-1)!!} \sum_{M\in\mathcal{M}} \left(\prod_{\substack{\text{cycles } C \\ \in M_0\uplus M}} \sum_{X_C^{(1)}} \prod_{(u,v)\in C} \exp\left(-2\delta H\left(X_u^{(1)},X_v^{(1)}\right)\right) \right)^k \qquad (46)$$

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 (46) is equal to the following.

$$\begin{aligned}
&\left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}}\right)^{nk} \frac{1}{(n-1)!!} \sum_{M\in\mathcal{M}} \left(\prod_{\substack{\text{cycles } C\\ \in M_0 \uplus M}} (1+e^{-2\delta})^{|C|} + (1-e^{-2\delta})^{|C|} \right)^{k} \\
&= \left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}}\right)^{nk} \frac{1}{(n-1)!!} \sum_{M\in\mathcal{M}} \left(\prod_{\substack{\text{cycles } C\\ \in M_0 \uplus M}} \left(\frac{e^{\delta}+e^{-\delta}}{e^{\delta}}\right)^{|C|} \left(1 + \left(\frac{e^{\delta}-e^{-\delta}}{e^{\delta}+e^{-\delta}}\right)^{|C|}\right) \right)^{k} \\
&= \mathbf{E}\left[\left(\prod_{\substack{\text{cycles } C\\ \in M_0 \uplus M}} \left(1+\tanh^{|C|}(\delta)\right) \right)^{k} \right] \qquad (47)
\end{aligned}$$

where the expectation is from choosing a uniformly random perfect matching $M \in \mathcal{M}$. At this point, it remains only to bound Equation (47). 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 (47) 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. Let ζ be a random variable representing the number of 2-cycles in $M_0 \uplus M$, i.e., the number of edges shared by both perfect matchings.

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

where in the last inequality, we used the facts that $\delta^{|C|}$ is maximized for $|C| \geq 4$ when $|C| = 4$, and that there are at most $n/4$ cycles of length at least 4.

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 16 with the above derivation, we have that

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

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 19.

8.3.3 Proof of Theorem 20

We provide constructions for our lower bounds of Theorem 20 which show that a dependence on β is necessary in certain cases.

Lemma 17. *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 (3), 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 18. *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 17. 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 (3), 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 $\tau = \beta/2$, $d_{\mathrm{SKL}}(p, q) = \frac{\beta}{2}(\tanh(\beta) - \tanh(\beta/2)) = \frac{\beta}{2}(\tanh(\beta/2)\operatorname{sech}(\beta)) \leq \beta\exp(-\beta) \leq \varepsilon$, 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 19. *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 17, 18, and 19 imply Theorem 20.

Acknowledgements

The authors would like to thank Yuval Peres for directing them towards the reference [LPW09] and the tools used to prove the variance bounds in this paper, and Pedro Felzenszwalb for discussing applications of Ising models in computer vision.

References

[AD15] 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.

[ADK15] 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.

[Agr12] Alan Agresti. *Categorical Data Analysis.* Wiley, 2012.

[AJ06] 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.

[AKN06] 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.

[BFF$^+$01] 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.

[BGS14] 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.

[Bha16] Bhaswar B. Bhattacharya. Power of graph-based two-sample tests. *arXiv preprint arXiv:1508.07530*, 2016.

[BK16] Guy Bresler and Mina Karzand. Learning a tree-structured Ising model in order to make predictions. *arXiv preprint arXiv:1604.06749*, 2016.

[BM16] Bhaswar B. Bhattacharya and Sumit Mukherjee. Inference in Ising models. *Bernoulli*, 2016.

[Bre15] 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.

[CDGR16] 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, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[CDKS17] Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing Bayesian networks. In *Proceedings of the 30th Annual Conference on Learning Theory*, COLT '17, pages 370–448, 2017.

[CF07] Nicholas A. Christakis and James H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379, 2007.

[Cha05] Sourav Chatterjee. *Concentration Inequalities with Exchangeable Pairs*. PhD thesis, Stanford University, June 2005.

[CL68] C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

[CT06] Imre Csiszár and Zsolt Talata. Consistent estimation of the basic neighborhood of Markov random fields. *The Annals of Statistics*, 34(1):123–145, 2006.

[DDK17] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Concentration of multilinear functions of the Ising model with applications to network data. In *Advances in Neural Information Processing Systems 30*, NIPS '17. Curran Associates, Inc., 2017.

[DDS$^+$13] Constantinos Daskalakis, Ilias Diakonikolas, Rocco A. Servedio, Gregory Valiant, and Paul Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '13, pages 1833–1852, Philadelphia, PA, USA, 2013. SIAM.

[DGJ08] Martin Dyer, Leslie Ann Goldberg, and Mark Jerrum. Dobrushin conditions and systematic scan. *Combinatorics, Probability and Computing*, 17(6):761–779, 2008.

[DK16] 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.

[DKN15a] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 1183–1202, Washington, DC, USA, 2015. IEEE Computer Society.

[DKN15b] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 1841–1854, Philadelphia, PA, USA, 2015. SIAM.

[DMR11] 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.

[Dob56] Roland L. Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory of Probability & Its Applications*, 1(1):65–80, 1956.

[DP17] Constantinos Daskalakis and Qinxuan Pan. Square Hellinger subadditivity for Bayesian networks and its applications to identity testing. In *Proceedings of the 30th Annual Conference on Learning Theory*, COLT '17, pages 697–703, 2017.

[Ell93] Glenn Ellison. Learning, local interaction, and coordination. *Econometrica*, 61(5):1047–1071, 1993.

[Fel04] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates Sunderland, 2004.

[Fis35] Ronald A. Fisher. *The Design of Experiments*. Macmillan, 1935.

[FLNP00] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620, 2000.

[FOS08] 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.

[Geo11] Hans-Otto Georgii. *Gibbs Measures and Phase Transitions*. Walter de Gruyter, 2011.

[GG86] 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.

[GLP17] Reza Gheissari, Eyal Lubetzky, and Yuval Peres. Concentration inequalities for polynomials of contracting Ising models. *arXiv preprint arXiv:1706.00121*, 2017.

[Hay06] Thomas P. Hayes. A simple condition implying rapid mixing of single-site dynamics on spin systems. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '06, pages 39–46, Washington, DC, USA, 2006. IEEE Computer Society.

[HKM17] Linus Hamilton, Frederic Koehler, and Ankur Moitra. Information theoretic properties of Markov random fields, and their algorithmic applications. In *Advances in Neural Information Processing Systems 30*, NIPS '17. Curran Associates, Inc., 2017.

[Isi25] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

[JJR11] 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.

[Jor10] Michael Jordan. Lecture notes for Bayesian modeling and inference, 2010.

[JS93] Mark Jerrum and Alistair Sinclair. Polynomial-time approximation algorithms for the Ising model. *SIAM Journal on Computing*, 22(5):1087–1116, 1993.

[KM17] Adam Klivans and Raghu Meka. Learning graphical models using multiplicative weights. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '17, Washington, DC, USA, 2017. IEEE Computer Society.

[KNS07] 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.

[LAFH01] Charles Lagor, Dominik Aronsky, Marcelo Fiszman, and Peter J. Haug. Automatic identification of patients eligible for a pneumonia guideline: comparing the diagnostic accuracy of two decision support models. *Studies in Health Technology and Informatics*, 84(1):493–497, 2001.

[LC73] Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

[LPW09] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

[LRR13] Reut Levi, Dana Ron, and Ronitt Rubinfeld. Testing properties of collections of distributions. *Theory of Computing*, 9(8):295–347, 2013.

[MdCCU16] Abraham Martín del Campo, Sarah Cepeda, and Caroline Uhler. Exact goodness-of-fit testing for the Ising model. *Scandinavian Journal of Statistics*, 2016.

[MS10] Andrea Montanari and Amin Saberi. The spread of innovations in social networks. *Proceedings of the National Academy of Sciences*, 107(47):20196–20201, 2010.

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[Pea00] 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.

[RAS15] Firas Rassoul-Agha and Timo Seppäläinen. *A Course on Large Deviations with an Introduction to Gibbs Measures*. American Mathematical Society, 2015.

[RS81] 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.

[RWL10] 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.

[STW10] 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.

[SW12] 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.

[TAW10] 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.

[VMLC16] 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.

[VV17] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

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 $0<\lambda\le 1/2$. How many samples do we need to tell which case we are in? If we wish to be correct with probability (say) $\ge 2/3$, 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 $k\le 1/4\lambda^2$ 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 20. *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 $p\ne\frac{1}{2}$,*

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

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 $S\ne 0$, 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 3), 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 $\Pr\left(Binomial\left(k, \frac{1}{2}+\lambda\right) \geq \frac{k}{2}\right)$. By a symmetry argument, this is equal to

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

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

$$\begin{aligned}
& d_{\mathrm{TV}}\left(Binomial\left(k, \frac{1}{2}-\lambda\right), Binomial\left(k, \frac{1}{2}+\lambda\right)\right) \\
\geq \quad & d_{\mathrm{TV}}\left(Binomial\left(k, \frac{1}{2}\right), Binomial\left(k, \frac{1}{2}+\lambda\right)\right) \\
\geq \quad & k \min_{\ell \in \{\lceil k/2 \rceil, \ldots, \lceil k/2 + k\lambda \rceil\}} \int_{1/2}^{1/2+\lambda} \Pr\left(Binomial\left(k-1, u\right) = l-1\right) du && (48) \\
\geq \quad & \lambda k \cdot \Pr\left(Binomial\left(k-1, 1/2+\lambda\right) = k/2\right) \\
= \quad & \lambda k \cdot \binom{k-1}{k/2} \left(\frac{1}{2}+\lambda\right)^{k/2} \left(\frac{1}{2}-\lambda\right)^{k/2-1} \\
\geq \quad & \Omega(\lambda k) \cdot \sqrt{\frac{1}{2k}} \left(1 + \frac{1}{\sqrt{k}}\right)^{k/2} \left(1 - \frac{1}{\sqrt{k}}\right)^{k/2} && (49) \\
= \quad & \Omega(\lambda\sqrt{k}) \cdot \left(1 - \frac{1}{k}\right)^{k/2} \\
\geq \quad & \Omega(\lambda\sqrt{k}) \cdot \exp\left(-1/2\right) \left(1 - \frac{1}{k}\right)^{1/2} && (50) \\
= \quad & \Omega(\lambda\sqrt{k}),
\end{aligned}$$

as desired.

(48) applies Proposition 2.3 of [AJ06]. (49) is by an application of Stirling's approximation and since $\lambda \leq \frac{1}{2\sqrt{k}}$. (50) 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 2. *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 $p \neq q$,*

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

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 $Z_i \sim \textit{Rademacher}\left(\frac{1}{2} + \frac{p-q}{2}\right)$ 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 20. By examining the guarantees of Lemma 20, this implies the desired result. □

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 [FOS08].

Theorem 21 (Theorem 17 from [FOS08])**.** *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 $d_{\mathrm{KL}}(p||q^*) \leq \varepsilon$. Suppose also that $a \leq q(x) \leq b$ 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 $d_{\mathrm{KL}}(p||q^{ML}) \leq 4\varepsilon$ with probability $1 - \delta$ where*

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

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

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

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.