Two-Sample Testing can be as hard as Structure Learning in Ising Models: Minimax Lower Bounds

Aditya Gangrade, Bobak Nazer, and Venkatesh Saligrama
Boston University
{gangrade,bobak,srv}@bu.edu

***Abstract*—Consider the following *structural* two-sample testing problem: given two sets of sample drawn from Ising models, determine whether the underlying network structure has changed. In [1], we showed that for Ising models over p variables with network structures that have degree bounded by d, under mild conditions on the model parameters, the sample complexity of this problem is very close to that of determining either of the network structures. Therefore, the naïve scheme of learning and then comparing the structures of both sets of samples is near data-optimal. However, the minimax lower bounds in [1] relied on Ising models that differed in only one edge, which leads to the natural follow-up question: are large changes significantly easier to detect? We extend the previously developed framework to consider this problem, and show that, in a certain parameter regime, large changes do not provide any significant improvement in the number of necessary samples for reliable two-sample testing.**

***Keywords*—High-dimensional Inference, Two-Sample Testing, Ising Models, Sample Complexity, Change Detection**

I. Introduction

Markov Random Fields (MRFs) provide a very general yet structured framework for describing dependencies in high-dimensional probabilistic models, and are widely applied in a multitude of areas. Recall that an MRF on p variables has an associated undirected graph $G = ([1 : p], E)$ such that the joint distribution of the random variables is Markov according to the edge set of G. We refer to this graph as the Markov network structure of the MRF. In this paper, we will concentrate on the simple instance of an Ising model, which is a $\{\pm 1\}^p$-valued MRF, and in particular on Ising models with network structures that have maximum degree at most d. We refer to the set of graphs on p nodes with degree bounded by d as $\mathcal{G}_{p,d}$.

We consider the two-sample problem of testing if the network structure of a system modeled by an Ising distribution has changed or not on the basis of two sets of samples drawn from it, and refer to the same as structural change detection. The change detection problem and the related problem of change estimation, have received both practical interest, particularly in biological applications [2]–[4], and theoretical interest, with the latter concentrating on algorithms to detect/estimate changes in Ising models, and the study of the sample complexity and consistency conditions of the same [3], [5]–[7].

A. Gangrade was supported by DHS Contract: HSHQDC-15-C-B0003 and NSF grant CCF-1618800. B. Nazer was supported by NSF grant CCF-1618800. V. Saligrama was supported by NSF grant CCF-1320547.

The related one-sample problem of estimating the network structure of an Ising model given samples drawn from it - here referred to as structure learning - has been well studied over the past decade. A large body of work has concentrated on constructing structure learning schemes for the Ising model, including algorithms, regularised estimators, and with particular focus on determining the sample complexity and consistency conditions of the same (see, e.g., [8]–[11] and references therein). A smaller body of work has studied the information-theoretic hardness of structure learning in terms of simple graph properties, providing necessary and sufficient conditions on sample complexity for reliable structure learning [9], [12], [13].

Note that a naïve approach to change detection is to completely learn the structures of the distributions underlying both sets of samples, and then compare them. Such a scheme is naturally considered profligate in the amount of samples it requires for the obvious reason that it is solving a harder problem than it needs to. This is further compounded in the practically interesting case when the underlying models are dense, but the changes between them are sparse, since structure learning sample complexity scales exponentially in d. A possible workaround is suggested by the compressed sensing literature, which demonstrates models where learning sparse changes can be much easier than learning either of the models themselves. A number of recent papers (for instance, [5], [6]) have considered direct change detection schemes that exploit the sparseness of changes in order to estimate changes with sample costs that are agnostic to the complexity of the underlying distributions themselves. In particular, these papers use 'regularised density ratio estimation' schemes to estimate changes with sample complexities $O(\text{poly}(\Delta)\log p)$, where Δ is the number of changed edges. However, these schemes all impose strong conditions on the possible underlying distributions that implicitly limit their complexity.

In [1] we took an information-theoretic approach to the change detection problem, and showed that for Ising models on $\mathcal{G}_{p,d}$ with bounded edge parameters, the change detection problem is *almost* as sample intensive as structure

learning. In particular, we provided lower bounds on the sample complexity of change detection that are in general exponential in d, and are separated from the corresponding structure learning sample complexity by a factor that is at most $\mathrm{poly}(d)$. However, our proof technique relied on the difficulty of detecting very small changes in the network structures of an Ising model - that of precisely one edge. A natural follow-up question then is whether larger changes are significantly easier to detect.

In this paper, we extend the framework of [1] to consider the detection of large changes. Informally, we posit that if a change occurs, then the network structures of the two underlying distributions must differ in at least Δ of the edges for a freely chosen parameter Δ. We call this problem Δ-change detection, and, as with the previous work, we obtain minimax sample complexity lower bounds required for the same. Depending on the parameter regime, we are able to demonstrate that either Δ-change detection requires roughly the same sample complexity as structure learning or that Δ must be exponentially large in d in order to obtain any benefits.

Notation: For a natural n, we use $[1:n]$ as shorthand for the set $\{1, 2, \ldots, n\}$. We denote random vectors, usually of dimension d or p, by X or $\widetilde{X}$, and let $\mathcal{X}$ denote the alphabet from which each entry is drawn. For a natural i, X_i is the i^{th} component of X. Similarly, for a set of naturals S, $X_S = (X_i)_{\{i \in S\}}$ is the sub-vector consisting of the entries of X whose indices are in S. We use X^n to denote an n-length sequence of i.i.d. random vectors, frequently referred to as a 'dataset'. For a distribution P, and given $X \sim P$, $P^{\otimes n}$ is the distribution of X^n. Further, given X^n, the t^{th} sample in that dataset is denoted $X^{(t)}$. The two element sets $\{i, j\}$ are interchangeably denoted (i, j) when referring to edges in an undirected graph, and as just ij when they appear in a subscript. Vectors in $\mathbb{R}^{\binom{n}{2}}$ are indexed by cardinality-two subsets of $[1:n]$. For instance, a vector $\theta \in \mathbb{R}^{\binom{3}{2}}$ is represented as $(\theta_{12}, \theta_{13}, \theta_{23})$. The identity matrix of size p is denoted as I_p.

For functions f, g, $f = O(g)$ if there exists a positive constant such that $\lim_{n\to\infty} f(n)/g(n) \leq C$, and $f = o(g)$ if $\lim f/g = 0$. Similarly, $f = \Omega(g)$ if $g = O(f)$, and $f = \omega(g)$ if $g = o(f)$. For a single variable n, we say that $g = \mathrm{poly}(n)$ if there exists a polynomial f such that $g = O(f(n))$.

II. Problem Statement and Definitions

A. Ising Models

Given a graph G and a vector $\theta \in \mathbb{R}^{\binom{p}{2}}$ such that $(i, j) \notin E \Rightarrow \theta_{ij} = 0$, a 0-external field *Ising model* on G with parameter θ is a $\{\pm 1\}$-valued Markov random field with the distribution

$$\mathbb{P}_{(G,\theta)}(X = x) = \frac{1}{Z(\theta)} \exp\left(\sum_{(u,v)\in E} \theta_{uv} x_u x_v\right),$$

where $Z(\theta)$ is a normalising constant commonly known as the partition function. We let $\mathcal{I}_{p,d}(\alpha, \beta)$ be the set of Ising models on graphs with p nodes and maximum degree d such that for every u, v, either $\theta_{uv} = 0$, or $\alpha \leq |\theta_{uv}| \leq \beta$ holds. Note that every θ determines a network structure, which we refer to as $G(\theta)$. We frequently describe Ising models in terms of their Markov network structures. In particular, we say that an Ising model has the edge (i, j) with weight w if $\theta_{ij} = w$. If all the edges of an Ising model have the same weight, we say that the distribution has uniform edge weights.

We say that two Ising models P, Q are Δ-separated if the edge sets of $G(P)$ and $G(Q)$ differ in at least Δ locations.

B. Change Detection

For Ising models $P, Q \in \mathcal{I}_{p,d}(\alpha, \beta)$, let $X^{n_1} \sim P^{\otimes n_1}$, and $\widetilde{X}^{n_2} \sim Q^{\otimes n_2}$ be finite sets of samples, also referred to as datasets, drawn independently and identically from P and Q, respectively. An (n_1, n_2)-sample Δ-change detector for $\mathcal{I}_{p,d}(\alpha, \beta)$ is a map $\phi : \mathcal{X}^{n_1} \times \mathcal{X}^{n_2} \to \{0, 1\}$. Let $\mathbf{\Phi}_{n_1,n_2}$ be the set of all (n_1, n_2)-sample Δ-change detectors. Let the risk of a detector ϕ be

$$\begin{aligned} &R(\phi; n_1, n_2) \\ &:= \sup_{P,Q \in \mathcal{I}_{p,d}(\alpha,\beta)} \mathbb{P}\left\{\phi(X^{n_1}, \widetilde{X}^{n_2}) = 1 \mid G(P) = G(Q)\right\} \\ &\qquad + \mathbb{P}\left\{\phi(X^{n_1}, \widetilde{X}^{n_2}) = 0 \mid |G(P) \triangle G(Q)| \geq \Delta\right\}, \end{aligned}$$

where $A \triangle B$ denotes the symmetric difference between sets A and B. Also, let the minimax change detection risk with (n_1, n_2) samples over $\mathcal{I}_{p,d}(\alpha, \beta)$ be

$$R_{\text{cd}}(n_1, n_2) := \inf_{\phi \in \mathbf{\Phi}_{n_1,n_2}} R(\phi; n_1, n_2).$$

We say that an (n_1, n_2)-sample Δ-change detector is δ-reliable over $\mathcal{I}_{p,d}(\alpha, \beta)$ if $R_{\text{cd}}(\phi; n_1, n_2) < \delta$, and say that the change detection problem can be solved over $\mathcal{I}_{p,d}(\alpha, \beta)$ δ-reliably with (n_1, n_2) samples if there exists an (n_1, n_2)-sample Δ-change detector over $\mathcal{I}_{p,d}(\alpha, \beta)$ that is δ-reliable or, equivalently, if $R_{\text{cd}}(n_1, n_2) < \delta$. The parameter δ is occasionally referred to as the reliability level.[1]

The main aim of this paper is to study the trade-off between the reliability level δ of change detection and the sample size (n_1, n_2).

Remark: Note that the above definition makes the restriction that if a change occurs, then it must occur in at least Δ edges. A more natural framework for detecting large changes might be to require the detector to differentiate between changes in less than Δ edges or more than Δ edges. However, the minimax risks of such a procedure are dominated by distributions on the boundary of such changes, and are the same as those in [1]. The obvious modification - to distinguish

[1] In this paper, we concentrate on achieving the reliability level $1/2$, but retain this definition for the sake of continuity with the definitions of [1].

between changes of less than, say, $\Delta/2$ edges and more than Δ edges is of interest, but since the above problem is reducible to such a detector, our lower bounds continue to hold for the same, and are already large enough to make our main point that change detection is nearly as data hungry as structure learning.

III. Main Results

As in [1], our main results are sample complexity lower bounds for reliable detection of large changes on $\mathcal{I}_{p,d}(\alpha,\beta)$. We begin by stating these results, and provide a proof sketch for the same. The complete proofs are omitted for lack of space. This is followed by a few remarks comparing these results with known upper bounds on the sample complexity of structure learning, and contextualising the regimes of Δ where the lower bounds deviate from them appreciably.

Theorem 1. *Let $\Delta \leq p^{1/3}/2$. If the number of samples is such that*

$$\min(n_1,n_2) < \frac{1}{\log(1+\tanh^2\alpha)} \log \frac{p}{4\Delta^3},$$

then the Δ-change detection risk over $\mathcal{I}_{p,d}(\alpha,\beta)$ exceeds $1/2$.

Theorem 2. *Let $\Delta \leq p^{1/3}/2$. If $d \geq 11$, and $\beta(d-1-\sqrt{8d}) \geq \log(d-1-\sqrt{8d})$, and if the number of samples is such that*

$$\min(n_1,n_2) < \frac{e^{2\beta(d-1-\sqrt{8d})}}{3\max(\Delta,d+1)} \log 1 + \left\lfloor \frac{p}{\max(\Delta,d+1)} \right\rfloor,$$

then the Δ-change detection risk over $\mathcal{I}_{p,d}(\alpha,\beta)$ exceeds $1/2$.

Comment: Note that Theorem 2 continues to hold even if the $\Delta \leq p^{1/3}/2$ condition is weakened to $\Delta \leq p/2$. We believe that the same would be true of Theorem 1, and that the $p^{1/3}/2$ condition is just an artifact of our proof technique. However, for the sake of exactness, we hereafter only consider the case $\Delta \leq p^{1/3}/2$.

Proof Sketch: We closely follow the proof strategy laid out in [1], which is loosely described here. Let $n = \min(n_1,n_2)$, P be an Ising model, and $\mathcal{Q}$ be a set of Ising models such that every $Q \in \mathcal{Q}$ has a network structure that is Δ-separated from $G(P)$. Suppose that we are told the distribution of the larger set of samples is P, and are also told that if a change occurs, the latter set of samples will be drawn from some $Q \in \mathcal{Q}$. Given this extra information, solving the change detection problem is reduced to solving a n sample goodness of fit test

$$H_0 : X^n \sim P^{\otimes n}$$
$$H_1 : X^n \sim Q^{\otimes n} \text{ for some } Q \in \mathcal{Q}$$

Let R' be the risk of the above test. By [1, Lemma 3],

$$R_{\mathrm{cd}} \geq R' \geq 1 - \frac{1}{2}\sqrt{\mathbb{E}_{P^{\otimes n}}[L_n^2] - 1},$$

where

$$L_n(X^n) := \frac{1}{|\mathcal{Q}|} \sum_{Q \in \mathcal{Q}} \frac{Q^{\otimes n}(X^n)}{P^{\otimes n}(X^n)}.$$

The above statement follows from noting that the risks of the above hypothesis test $1 - d_{\mathrm{TV}}(P^{\otimes n}, \widetilde{Q}^n)$, where $\widetilde{Q}^n$ is the average of the alternate densities $\{Q^{\otimes n}\}$, and observing that the χ^2-divergence is an upper bound on the total variation.

Now note that picking appropriate $(P,\mathcal{Q})$ from $\mathcal{I}_{p,d}(\alpha,\beta)$ yields lower bounds on the change detection risk. We call such pairs change detection ensembles. The ensembles used to prove the above bounds are as follows:

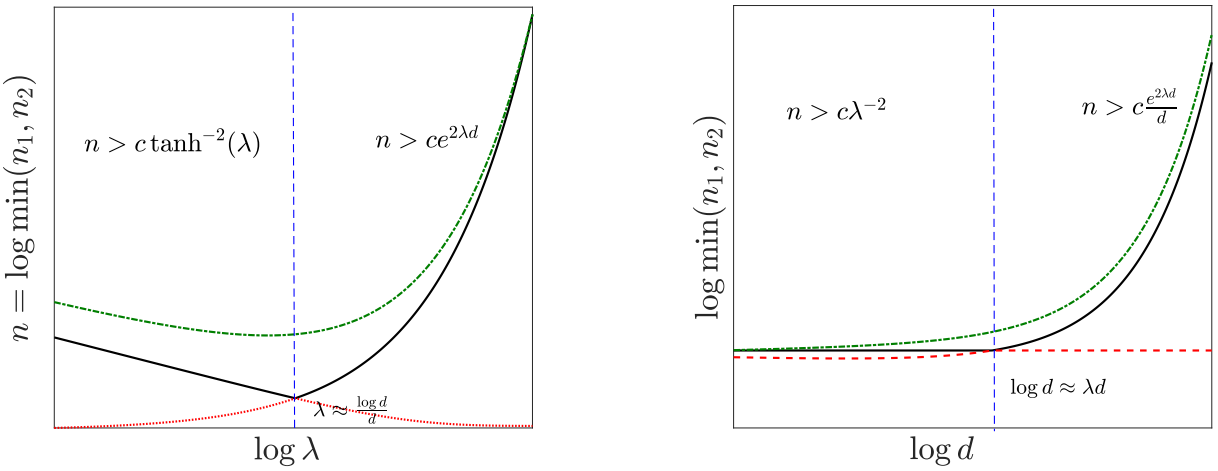
- Theorem 1: P is the Ising model on the graph with no edges on p nodes, while $\mathcal{Q}$ is the set of $\binom{p/2}{\Delta}$ Ising models formed by selecting Δ edges out of $\{(1,2),(3,4),\ldots,(p-1,p)\}$ and setting each edge weight to α.
- Theorem 2: P is the Ising model with uniform edge weight β on $\lfloor p/(d+1) \rfloor$ separate cliques of size $d+1$ each, while $\mathcal{Q} = \{Q_i : 1 \leq i \leq p/\Delta\}$, where each Q_i is formed by the following procedure:
  - Label the cliques in the structure of P as $\{1,2,\ldots,\lfloor p/(d+1) \rfloor\}$. Within each clique, fix a labelling of the nodes.
  - Let K' be the graph formed by taking the complete graph on $d+1$ labelled nodes, and deleting all edges in the subclique formed by the first $\sqrt{2d}$ nodes.
  - Let G_i be the graph formed by taking the structure of P, and replacing the cliques numbered $\frac{\Delta(i-1)}{d}+1$ through $\frac{\Delta i}{d}$ by K'
  - Q_i is the Ising model on G_i with uniform edge weight β.

Note: The ensemble used for the proof of Theorem 1 is canonical in that any reasonable graph class that allows for Δ changes should contain all graphs used. Thus, Theorem 1 applies in nearly every context of interest. In particular, it applies to forest-structured Ising models, and shows that (not too) large changes have an entirely negligible effect on the sample complexity when $p \gg 1, \Delta \leq p^{1/3-\epsilon}$.

A. Remarks

In the high-dimensional setting, one considers how the above bounds vary as p grows large, allowing α,β,d,Δ to vary with p arbitrarily. In this setting, three distinct properties emerge (below c is some arbitrary constant):

1) If α,β,d are held constant, then reliable Δ-change detection requires at least $c \log \frac{p}{\Delta^3}$ samples.
2) If, on the other hand, α and β are held constant while d is allowed to grow with p, then for modetly large d, the bound of Theorem 2 dominates, and Δ-change detection requires at least $c\frac{e^{2\beta d(1-O(d^{-1/2}))}}{\Delta} \log \frac{p}{\Delta}$ samples. Note that in order for the exponential factor in d to be ameliorated, one requires that $\Delta > e^{2\beta d(1-(8d)^{-1/2})}$, i.e., exponentially large in d.
3) Lastly, if α and β are allowed to vary as well, then unless $\beta(d-1-\sqrt{8d}) < \log(d-1-\sqrt{8d})$, we are forced into the exponential in d growth regime. However, since $\alpha \leq \beta$, reducing β strongly leads to an increased



(a) (b) Fig2

Fig. 1: A pastiche of log-log plots of sample complexity lower bounds when $\alpha = \beta = \lambda$, and p, Δ are held constant, and terms involving them are hidden in the factor of c above. Note that these plots are cartoons, and the plots above may very up to constant factors. In either figure, the black solid line denotes our lower bound, the dotted red line denotes inactive lower bounds, the alternating dashed green line is the upper bound from [12, Thm. 3a)], and the dashed blue vertical line indicates the crossover point at which the bound of Theorem 2 begins to dominate that of 1. Figure a) considers the case when d is held constant and λ is varied. Figure b) considers the case when $\lambda < 1$ is held constant and d is allowed to vary. The constant branch disappears when $\lambda > 1$. Note the transition of behaviour around the point $\lambda = \log d/d$ in either plot.

cost due to the bound in Theorem 1. Optimising α, β to *minimise* this scaling cost leads to the fact that no matter how the parameters are varied, if $\Delta = \text{poly}(d)$, then one needs at least $c\frac{d^2}{\log^2 d} \log \frac{p}{\Delta}$ samples for reliable Δ-change detection. This is in stark contrast to results such as those in [5], [6], which suggest that direct change detection can be done with sample complexity $O(\text{poly}(\Delta) \log p)$ with no dependence on d, and hints at the restrictivenes of the technical conditions imposed in these papers.

B. Comparison with Structure Learning Bounds

Our main point of comparison is the following result of Santhanam Wainwright [12], which is obtained by analysing the maximum likelihood structure learning scheme. Note that both the phrasing and notation have been altered to suit the needs of this paper.

Theorem [12, Thm. 3a)]. *Suppose that the possible edge-weights are known to the structure learner. It is possible to correctly identify the network structure of a distribution in $\mathcal{I}_{p,d}(\alpha, \beta)$ with probability greater than δ if*

$$n \geq \frac{3(3e^{2\beta d} + 1)}{\sinh^2(\alpha/2)} d \left(3 \log p + \log 2d + \log \frac{1}{\delta} \right).$$

Ignoring the δ terms, this bound is separated from our lower bounds by a factor of $O\left(\frac{d\Delta e^{4\beta\sqrt{2d}}}{\sinh^2 \alpha}\right)$, which for $\alpha^{-1}, \Delta \in \text{poly}(d)$ is subexponential in d. This implies that under such a condition on the parameters, our bounds correctly identify the sample complexity up to exponential order. Further, since our bounds are based on analysing goodness-of-fit tests with known parameter space, the closeness of these bounds also suggests that our techniques cannot yield significantly stronger lower bounds in such a regime.

IV. Discussion

The article determines parameter regimes where large changes have little to no effect on the hardness of change detection. However, it is not clear to us if the $1/\Delta$ factor in Theorem 2 is tight or not. It is an open problem to either determine schemes that enjoy this advantage, or to improve the lower bounds enough to eliminate this factor. On a broader note, we reiterate the importance of change detection in practical contexts described in [1], and mention the interesting problem of determining natural graph classes in which the minimax sample complexity of change detection is exponentially separated from that of structure learning.

References

[1] A. Gangrade, B. Nazer, and V. Saligrama, "Lower bounds for two-sample structural change detection in Ising and Gaussian models," in *Communication, Control, and Computing, 2017. Allerton 2017. 55th Annual Allerton Conference on*. IEEE, 2017.

[2] E. Belilovsky, G. Varoquaux, and M. B. Blaschko, "Testing for differences in Gaussian graphical models: applications to brain connectivity," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 595–603.

[3] S. D. Zhao, T. T. Cai, and H. Li, "Direct estimation of differential networks," *Biometrika*, vol. 101, no. 2, pp. 253–268, 2014.

[4] Y. Xia, T. Cai, T. T. Cai *et al.*, "Testing differential networks with applications to the detection of gene-gene interactions," *Biometrika*, vol. 102, no. 2, pp. 247–266, 2015.

[5] F. Fazayeli and A. Banerjee, "Generalized direct change estimation in Ising model structure," in *Proceedings of The 33rd International Conference on Machine Learning (ICML 2016)*, vol. 48, 2016, pp. 2281–2290.

[6] S. Liu, T. Suzuki, R. Relator, J. Sese, M. Sugiyama, and K. Fukumizu, "Support consistency of direct sparse-change learning in Markov networks," *The Annals of Statistics*, vol. 45, no. 3, pp. 959–990, 2017. [Online]. Available: http://dx.doi.org/10.1214/16-AOS1470

[7] S. Liu, K. Fukumizu, and T. Suzuki, "Learning sparse structural changes in high-dimensional Markov networks," *Behaviormetrika*, vol. 44, no. 1, pp. 265–286, 2017.

[8] G. Bresler, "Efficiently learning Ising models on arbitrary graphs," in *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing (STOC 2015)*. ACM, 2015, pp. 771–782. [Online]. Available: http://doi.acm.org/10.1145/2746539.2746631

[9] A. Anandkumar, V. Y. Tan, F. Huang, and A. S. Willsky, "High-dimensional structure estimation in Ising models: Local separation criterion," *The Annals of Statistics*, vol. 40, no. 3, pp. 1346–1375, 2012.

[10] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, "High-dimensional Ising model selection using ℓ1-regularized logistic regression," *The Annals of Statistics*, vol. 38, no. 3, pp. 1287–1319, 2010.

[11] J. A. Bento and A. Montanari, "Which graphical models are difficult to learn?" in *Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 1303–1311.

[12] N. P. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4117–4134, 2012.

[13] K. Shanmugam, R. Tandon, P. K. Ravikumar, and A. G. Dimakis, "On the information-theoretic limits of learning Ising models," in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2303–2311.