The Annals of Statistics
2020, Vol. 48, No. 2, 713–737
https://doi.org/10.1214/19-AOS1808

LEARNING A TREE-STRUCTURED ISING MODEL IN ORDER TO MAKE PREDICTIONS

BY GUY BRESLER[1] AND MINA KARZAND[2]

[1]*Center for Statistics and Data Science, Department of EECS, Massachusetts Institute of Technology, guy@mit.edu*
[2]*Wisconsin Institute for Discovery, University of Wisconsin–Madison, karzand@wisc.edu*

We study the problem of learning a tree Ising model from samples such that subsequent predictions made using the model are accurate. The prediction task considered in this paper is that of predicting the values of a subset of variables given values of some other subset of variables. Virtually all previous work on graphical model learning has focused on recovering the true underlying graph. We define a distance ("small set TV" or ssTV) between distributions P and Q by taking the maximum, over all subsets $\mathcal{S}$ of a given size, of the total variation between the marginals of P and Q on $\mathcal{S}$; this distance captures the accuracy of the prediction task of interest. We derive nonasymptotic bounds on the number of samples needed to get a distribution (from the same class) with small ssTV relative to the one generating the samples. One of the main messages of this paper is that far fewer samples are needed than for recovering the underlying tree, which means that accurate predictions are possible using the wrong tree.

1. Introduction. Markov random fields, or undirected graphical models, are a useful way to represent high-dimensional probability distributions [24, 44]. A Markov random field is a probability distribution described by a graph: each node in the graph corresponds to a random variable, and the variables are required to satisfy the Markov property whereby a variable is conditionally independent of all other variables given its neighbors.

The practical utility of Markov random fields is in large part due to (1) edges between variables capture direct interactions, which make the model *interpretable* and (2) the graph structure facilitates efficient approximate *inference* from partial observations, for example, using loopy belief propagation or variational methods. A prediction for X_i based on observed values $X_{\mathcal{S}} = x_{\mathcal{S}}$ for a subset of variables $\mathcal{S}$ can be easily obtained from the conditional probability $P(X_i = x_i | X_{\mathcal{S}} = x_{\mathcal{S}})$. The inference task relevant to this paper is therefore evaluation of conditional probabilities or marginals.

In applications, it is often necessary to learn the model from data, and it makes sense to measure accuracy of the learned model in a manner corresponding to the intended use. While in some applications it is of interest to learn the graph itself, in many machine learning problems the focus is on making predictions. In the literature, learning the graph is called *structure learning*; this problem has been studied extensively in recent years; see, for example, [1, 11, 28, 29, 34, 43]. In contrast, we consider in this paper the problem of learning a good model *for the purpose of performing subsequent prediction from partial observations*. For instance, one might wish to use the learned model to predict the preference of a user for an item in a recommendation system based on ratings obtained for a few items. This objective has been called "inferning" (inference + learning) [21], and has received significantly less attention. This paper contains, to the best of our knowledge, the first results on estimating graphical models with a prediction-centric loss that are applicable to the high-dimensional setting.

Received November 2016; revised April 2018.
MSC2010 subject classifications. 62F12, 62H12.
Key words and phrases. High-dimensional statistics, model selection, Markov random fields, Ising model, prediction, tree model.

Structure learning becomes statistically more challenging, meaning more data is required, when interactions between variables are very weak or very strong [8, 37, 41]. It is intuitively clear that very weak edges are difficult to detect, leading to nonidentifiability of the model. The goal of this paper is to show that learning a model that makes accurate predictions is possible even when structure learning is not.

With the goal of making predictions in mind, we introduce a loss function to evaluate learning algorithms based on the accuracy of low-order marginals. The small-set total variation between true distribution P and learned distribution Q is defined to be

$$\mathcal{L}^{(k)}(P, Q) \triangleq \max_{\mathcal{S}:|\mathcal{S}|=k} d_{\mathrm{TV}}(P_{\mathcal{S}}, Q_{\mathcal{S}}),$$

where $P_{\mathcal{S}}$ denotes the marginal on set $\mathcal{S}$. The small-set total variation is inherently far less stringent than the total variation over the entire joint distribution and this makes a crucial difference in high-dimensional models. This same local total variation metric was used by Rebeschini and van Handel in a somewhat different context [35] and has appeared earlier in Dobrushin's work on Gibbs measures [20]. As discussed in Section 3.2, small loss $\mathcal{L}^{(k)}$ guarantees accurate posterior distributions conditioned on sets of size $k-1$.

Tree-structured graphical models have been particularly well studied. Aside from their theoretical appeal, there are two reasons for the practical utility of tree models: (1) The maximum likelihood tree can be easily computed, and the correct graph can be recovered with smaller sample and time complexity as compared to loopy graphs, and (2) Efficient exact inference (computation of marginals or maximum probability assignments) is possible using belief propagation. Sum-product or max-product algorithms are two well-studied examples [25, 32, 44, 45] of inference algorithms on trees. Hence, we focus on tree-structured models.

In this paper, we (further) restrict attention to tree-structured Ising models with no external field, defined as follows. For tree $\mathsf{T} = (\mathcal{V}, \mathcal{E})$ on p nodes and edge parameters θ_{ij} for each edge $(i, j) \in \mathcal{E}$, each configuration $x \in \{-1, +1\}^p$ is assigned probability

$$P(x) = \exp\Bigg(\sum_{(i,j)\in\mathcal{E}} \theta_{ij} x_i x_j - \Phi(\theta) \Bigg), \tag{1.1}$$

where $\Phi(\theta)$ is the normalizing constant. We assume throughout that $\alpha \le |\theta_{ij}| \le \beta$ for some $\alpha, \beta \ge 0$ for each edge $(i, j) \in \mathcal{E}$. Due to the tree structure, it turns out that the variables $Y_{ij} = X_i X_j$ for $(i, j) \in \mathcal{E}$ are jointly independent (as shown in Lemma 8.6), a fact that is useful in the analysis. As a consequence, the correlation between a pair of variables is equal to the product of the correlations $\mathbb{E}[X_i X_j]$ on the edges (i, j) in the path connecting them.

In general, we could have an external field term $\sum_{i\in\mathcal{V}} \theta_i x_i$ in the exponent of (1.1). The assumption of no external field ($\theta_i = 0$) implies uniform singleton marginals, that is, $P(x_i = +1) = 1/2$ for all i. This assumption helps to make the analysis tractable and at the same time captures the central features of the problem.

Suppose we observe i.i.d. samples, generated from a tree-structured Ising model. The main question we address is how many samples are required in order to learn a model with a guarantee on the accuracy of subsequent predictions computed using the learned model. Since computation of marginals for a given tree model is easy, the crux of the task is in learning a model with marginals that are close to those of the original model. One of the take-home messages is that learning for the purpose of making predictions requires dramatically fewer samples than is necessary for correctly recovering the underlying tree. The central technical challenge is that our analysis must therefore also apply when it is impossible to learn the true tree, and this requires careful control of the sorts of errors that can occur.

Our main result gives lower and upper bounds on the number of samples needed to learn a tree Ising model to ensure small $\mathcal{L}^{(2)}$ loss, which in this setting is equivalent to accurate

pairwise marginals. We emphasize that the task is to learn a model from the same class (tree-structured Ising) with these guarantees; this is sometimes called *proper* learning. The main result concerns the maximum likelihood tree (also called Chow–Liu tree), defined in Section 3.

THEOREM 1.1. *Fix $\eta > 0$. Given $n > C \max\{\eta^{-2}, e^{2\beta}\} \log \frac{p}{\delta}$ samples generated according to a tree Ising model P defined in (1.1) with $|\theta_{ij}| \le \beta$, denote the Chow–Liu tree by T^{CL}. The Ising model Q on T^{CL} obtained by matching correlations on the edges satisfies $\mathcal{L}^{(2)}(P, Q) \le \eta$ with probability at least $1 - \delta$. Conversely, if $\tanh\alpha + 2\eta \le \min\{\tanh\beta, 1/2\}$ and $n \le C'\eta^{-2} \log p$, then no algorithm can find a tree model Q such that $\mathcal{L}^{(2)}(P, Q) \le \eta$ with probability greater than half.*

The result shows that the Chow–Liu tree, which can be found in time $O(p^2 \log p)$, gives small $\mathcal{L}^{(2)}$ error. We remark that the Chow–Liu algorithm uses only pairwise marginals of the empirical distribution and can therefore be implemented with missing data as long as the pairwise marginals can be estimated.

In Section 3.3, we discuss the assumption $\tanh\alpha + 2\eta \le \min\{\tanh\beta, 1/2\}$ made in the theorem, which captures the fact that the learner does not know the magnitude of the edge parameters a priori.

It turns out that for trees, accuracy of pairwise marginals translates to accuracy of higher order marginals, and a bound to this effect is proved in Appendix H. We believe that the dependence on k can be improved.

COROLLARY 1.2. *Let T and T' be two (possibly distinct) trees. Let P and Q be probability distributions represented according to T and T' using (1.1) such that $\mathcal{L}^{(2)}(P, Q) < \eta$. Then for all k, we have $\mathcal{L}^{(k)}(P, Q) < k2^k\eta$.*

Numerical simulations in Section 9 show the dependence of the loss on number of samples for different values of α and β, supporting Theorem 1.1. There are a few important issues that are not addressed by Theorem 1.1, which we also investigate via simulations in Appendix I. These include robustness of the results to model misspecification, that is, samples are not from a tree-structured Ising model; external field in the Ising model (1.1) generating the data; accurate marginals of size $k \ge 3$.

We next place the result in the context of related work.

1.1. *Related work.* Tree-structured graphical models have applications in image processing and computer vision [18, 33, 36, 47], artificial intelligence [32], coding theory [19] and statistical physics [7].

Structure learning in general graphical models has been studied extensively. Information-theoretic bounds on the number of samples have been derived [8, 9, 27, 37, 41, 48]. Structure learning of trees has been studied by [12, 39]. Learning of generalizations of tree-structured models has been studied, including: forest approximations [26, 40], polytrees [15], bounded treewidth graphs [31, 38], loopy graphs with correlation decay [5, 6] and mixtures of trees [4, 28].

Loopy belief propagation yields accurate marginals in high girth (locally tree-like) graphs with correlation decay. This fact was used by Heinemann and Globerson [21] to justify an algorithm that recovers all the edges of a model from this family, given sufficiently many samples. The output of their algorithm can have extra edges, which are proved to be weak. Given the number of samples at least linear in p, they show that the learned distribution is at most constant Kullback–Leibler divergence from the true one.

Narasimhan and Bilmes [31] found an algorithm with polynomial runtime that uses a polynomial number of samples to learn bounded tree-width graphical models with respect to KL-divergence. They use ideas from submodular optimization and the specific factorization of the distribution over bounded tree-width graphs.

Structure learning of latent tree models has been well studied in the phylogenetic reconstruction literature. Erdős et al. [17] studied sample and time complexity of tree metric based algorithms to reconstruct phylogenetic trees. Daskalakis et al. [16] and Mossel [30] use distorted tree metrics to get approximations of phylogenetic trees when exact reconstruction of the tree is impossible. In [30], a forest approximation of the latent tree is recovered. The maximum number of connected components in this forest is a function of the edge strengths, maximum distance between the leaves and the number of leaves. Daskalakis et al.[16] removed the prior assumptions on the phylogenetic tree and instead a forest structure is recovered that contains all edges that are sufficiently strong and sufficiently close to the leaves.

A tree metric over p nodes is associated with a weighted spanning tree such that the distance between every pair of nodes is the sum of weights of the edges along the path between the nodes in the tree. Agarwala et al. [2], approximate a pairwise distance matrix D over p nodes by a tree metric with induced distance matrix T. Let $\epsilon = \min_T\{\|T - D\|_\infty\}$ where T is a tree metric. They propose an $O(p^2)$ algorithm which produces $\widehat{T}$ with $\|\widehat{T} - D\|_\infty \le 3\epsilon$. They prove that finding a T with $\|T - D\|_\infty \le 9/8\epsilon$ is NP-hard. Ambainis et al. in [3] studied the *leaf variational distance* between the original distribution on a latent tree under the Cavender–Faris (CF) model and the learned latent tree. Let tree T with p leaves $\mathcal{V}$ be a CF-tree with the property that all its edges are of length at least $1/\sqrt{n}$ (this is translated as $\log\tanh\alpha < 1/\sqrt{n}$ in our model). Then, given n observations, their proposed algorithm produces a distribution Q on a tree T' with leaves $\mathcal{V}$ such that $\mathcal{L}^{(2)}(P_{\mathcal{V}}, Q_{\mathcal{V}}) = O(\sqrt{pe^{2\beta}/n})$ ($P_{\mathcal{V}}$ and $Q_{\mathcal{V}}$ are marginals of P and Q on leaves $\mathcal{V}$). We will further discuss these results and compare them with our setup in detail in Appendix J.

Wainwright [46] was motivated by the same general problem of learning a graphical model to be subsequently used for making predictions, but his focus was on computational rather than statistical limits. For loopy graphs, both estimation of parameters and prediction based on partial observations are computationally challenging tasks. Hence, for both, approximate heuristic methods are often used. For given model parameters, one such heuristic for prediction is reweighted sum-product (a convex relaxation). Intriguingly, when using such approximate prediction algorithms, an *inconsistent* procedure for estimation of parameters can give better predictions. The results elucidate asymptotic performance, but the analysis does not apply to the high-dimensional setting of interest, with dimension p larger than number of samples n.

1.2. *Outline of paper.* The next section contains background on the Ising model, tree models and graphical model learning. Section 3.1 introduces the problem of learning tree-structured Ising models and records the sample complexity of exact recovery. Then, in Section 3.2 we define the small-set TV loss function motivated by prediction computations and state our main result in Section 3.3. Section 4.1 analyzes an illustrative example that gives intuition for the main result. Section 4.2 introduces a natural forest approximation algorithm and analyzes its performance in terms of ssTV. Section 5 sketches the proof of the main result and Section 6 fills in the details. Sections 7 and 8 contain further proofs. Numerical simulations addressing the theorems in the paper are in Section 9. Appendices contain additional proofs, numerical simulations and discussions on related work, available in the Supplementary Material [10].

2. Preliminaries.

2.1. *Notation.* For a given tree $\mathsf{T} = (\mathcal{V}, \mathcal{E})$ and positive numbers α and β, let $\mathcal{P}_{\mathsf{T}}(\alpha, \beta)$ be the set of Ising models (1.1) with the restriction $\alpha \le |\theta_{ij}| \le \beta$ for each edge $(i, j) \in \mathcal{E}$ and $\theta_{ij} = 0$ for $(i, j) \notin \mathcal{E}$. Denote by $\mathcal{P}_{\mathsf{T}} = \mathcal{P}_{\mathsf{T}}(0, \infty)$ the set of Ising models on T with no restrictions on parameter strength.

Denote by $\mu_{ij} = \mathbb{E}_P X_i X_j$ the correlation between the variables corresponding to $i, j \in \mathcal{V}$. For an edge $e = (i, j)$, we write $\mu_e = \mu_{ij}$ and for a set of edges $\mathcal{A} \subseteq \mathcal{E}$, $\mu_{\mathcal{A}} = \prod_{e \in \mathcal{A}} \mu_e$. Given n i.i.d. samples $X^{(1:n)} = X^{(1)}, \dots, X^{(n)}$, the empirical distribution is denoted by $\widehat{P}(x) = \frac{1}{n} \sum_{l=1}^{n} \mathbf{1}_{\{X^{(l)} = x\}}$ and $\widehat{\mu}_{ij} = \mathbb{E}_{\widehat{P}} X_i X_j$ is the empirical correlation between nodes i and j.

2.2. *Tree models.* A probability measure P on $\mathcal{X}^{\mathcal{V}}$ is Markov with respect to a graph $G = (\mathcal{V}, \mathcal{E})$ if for all $i \in \mathcal{V}$, we have $P(x_i | x_{\mathcal{V} \setminus \{i\}}) = P(x_i | x_{\partial i})$, where ∂i is the neighborhood of i in G. In this paper, we are interested in distributions P that are Markov with respect to a tree $\mathsf{T} = (\mathcal{V}, \mathcal{E})$, and a consequence (see [25]) is that $P(x)$ factorizes as

$$P(x) = \prod_{i \in \mathcal{V}} P(x_i) \prod_{(i,j) \in \mathcal{E}} \frac{P(x_i, x_j)}{P(x_i) P(x_j)}. \tag{2.1}$$

2.3. *Information projection.* Denote by $D(Q \parallel P)$ the Kullback–Leiber divergence between probability measures Q and P defined as $D(Q \parallel P) = \sum_{x \in \mathcal{X}} Q(x) \log \frac{Q(x)}{P(x)}$. For an arbitrary distribution P and tree T, the distribution

$$\widetilde{P}(x) = \operatorname*{arg\,min}_{\substack{Q \text{ is factorized as (2.1)} \\ \text{according to} \mathsf{T}}} D(P \parallel Q)$$

is the best approximation to P within the set of distributions Markov with respect to the tree T. It was observed by Chow and Liu in [12] that $\widetilde{P}$ is obtained by matching the first and second order marginals to those of P, that is, for all $(i, j) \in \mathcal{E}$, and all $x_i, x_j \in \mathcal{X}$, $\widetilde{P}(x_i, x_j) = P(x_i, x_j)$.

Let

$$\Pi_{\mathsf{T}}(P) = \operatorname*{arg\,min}_{Q \in \mathcal{P}_{\mathsf{T}}} D(P \parallel Q) \tag{2.2}$$

be the reverse information projection of P onto the class of Ising models on T with no external field. It follows from the definition of $\mathcal{P}_{\mathsf{T}}$ that $\widetilde{P} = \Pi_{\mathsf{T}}(P)$ can be represented as equation (1.1) for some $\widetilde{\theta}$ supported on T. It is shown in Appendix A in the Supplementary Material that $\widetilde{P} = \Pi_{\mathsf{T}}(P)$ has edge weights $\widetilde{\theta}_{ij}$ for each $(i, j) \in \mathcal{E}_{\mathsf{T}}$ satisfying $\tanh \widetilde{\theta}_{ij} = \mu_{ij} \triangleq \mathbb{E}_P X_i X_j$ (and $\widetilde{\theta}_{ij} = 0$ if $(i, j) \notin \mathcal{E}_{\mathsf{T}}$).

2.4. *Tree structure learning.* Denote the set of all trees on p nodes by $\mathcal{T}$. For some tree T and distribution $P \in \mathcal{P}_{\mathsf{T}}$, one observes n independent samples (configurations) $X^{(1)}, \dots, X^{(n)} \in \{-, +\}^p$ from the Ising model (1.1). In this context, a *structure learning algorithm* is a (possibly randomized) map $\phi : \{-1, +1\}^{p \times n} \to \mathcal{T}$ taking n samples $X^{(1:n)} = X^{(1)}, \dots, X^{(n)}$ to a tree $\phi(X^{(1:n)})$.

The maximum likelihood tree or Chow–Liu tree plays a central role in tree structure learning. Chow and Liu [12] observed that the maximum likelihood tree is the max-weight spanning tree in the complete graph, where each edge has weight equal to the empirical mutual information between the variables at its endpoints. The tree can thus be found greedily via Kruskal's algorithm [12, 14], and the run-time is dominated by computing empirical mutual information between all pairs of nodes.

In order to support the following definition, we analyzed zero-field Ising models on trees in Lemma 2 in Appendix A. This analysis is similar to [12].

DEFINITION 2.1 (Chow–Liu tree). Given n i.i.d. samples $X^{(1:n)}$ from distribution $P \in \mathcal{P}_{\mathsf{T}}$, we define the Chow–Liu tree to be the maximum likelihood tree:

$$\mathsf{T}^{\mathsf{CL}} = \operatorname*{argmax}_{\mathsf{T}\in\mathcal{T}} \max_{P\in\mathcal{P}_{\mathsf{T}}} P\big(X^{(1:n)}\big).$$

This definition is slightly abusing the conventional terminology, as the Chow–Liu tree is classically the maximum likelihood tree assuming that the generative distribution is tree-structured [12], whereas in our definition we assume that the original distribution $P \in \mathcal{P}_{\mathsf{T}}$ can be described by (1.1). Thus, it is not only tree-structured, but also has uniform singleton marginals.

Note that maximizing the likelihood of i.i.d. samples corresponds to minimizing the KL divergence. Given the samples with empirical distribution $\widehat{P}$,

$$\mathsf{T}^{\mathsf{CL}} = \arg\min_{\mathsf{T}\in\mathcal{T}} \min_{P\in\mathcal{P}_{\mathsf{T}}} D(\widehat{P} \parallel P).$$

It is shown in Lemma 1 in Appendix A that

$$\mathsf{T}^{\mathsf{CL}} = \operatorname*{argmax}_{\{\text{spanning trees } \mathsf{T}'\}} \sum_{e\in\mathcal{E}_{\mathsf{T}'}} |\widehat{\mu}_e|, \tag{2.3}$$

where for $e = (i, j)$, $\widehat{\mu}_e = \mathbb{E}_{\widehat{P}} X_i X_j$ is the empirical correlation between variables X_i and X_j.

Chow and Wagner [13] showed that the maximum likelihood tree is consistent for structure learning of general discrete tree models, that is, in the limit of large sample size the correct graph structure is found. More recently, detailed analysis of error exponents was carried out by Tan et al. [39, 40]. A variety of other results and generalizations have appeared, including for example Liu et al.'s work on forest estimation with nonparametric potentials [26] (we will not address general potentials in this paper).

3. Learning trees to make predictions. In order to place the learning for predictions problem into context, we first discuss the problem of exact structure learning and give tight (up to a constant factor) sample complexity for that problem. Then, in Section 3.2 we define the ssTV distance $\mathcal{L}^{(k)}$, explain how it relates to prediction and in Section 3.3 we state our results.

3.1. *Exact recovery of trees.* The statistical performance of a structure learning algorithm is often measured using the zero–one loss,

$$\mathcal{L}^{0\text{–}1}(\mathsf{T}, \mathsf{T}') = \mathbf{1}_{\{\mathsf{T}\neq\mathsf{T}'\}}, \tag{3.1}$$

meaning that the exact underlying graph must be learned (see, e.g., [11, 26, 37, 40]). The risk, or expected loss, of algorithm ϕ under some distribution $P \in \mathcal{P}_{\mathsf{T}}(\alpha, \beta)$ is then given by the probability of reconstruction error, $\mathbb{E}_P \mathcal{L}^{0\text{–}1}(\mathsf{T}, \phi(X^{(1:n)})) = \mathsf{P}(\phi(X^{(1:n)}) \neq \mathsf{T})$, and the maximum risk is $\sup\{\mathsf{P}(\phi(X^{(1:n)}) \neq \mathsf{T}) : \mathsf{T} \in \mathcal{T}, P \in \mathcal{P}_{\mathsf{T}}(\alpha, \beta)\}$ for given α, β, p and n.

The sample complexity of learning the correct tree underlying the distribution increases as edges become weaker, that is, as $\alpha \to 0$, because weak edges are harder to detect. As the bound on maximum edge parameter β increases, there is a similar increase in sample complexity (as shown by [37, 41] for Ising models on general bounded degree graphs). In the context of tree-structured Ising models, we have the following theorem.

THEOREM 3.1 (Samples necessary for structure learning). *Given* $n < \frac{1}{8}e^{2\beta}/(\alpha \tanh \alpha) \times \log p$ *samples, the worst-case probability of error over trees* $\mathsf{T} \in \mathcal{T}$ *and distributions* $P \in \mathcal{P}_{\mathsf{T}}(\alpha, \beta)$ *is at least half for any algorithm, that is,*

$$\inf_{\phi} \sup_{\substack{\mathsf{T}\in\mathcal{T} \\ P\in\mathcal{P}_{\mathsf{T}}(\alpha,\beta)}} P\big[\phi\big(X^{(1:n)}\big) \neq \mathsf{T}\big] > 1/2.$$

The proof, given in Section 7.1, applies Fano's inequality (Lemma 6.2) to a large set of trees that are difficult to distinguish. The next theorem gives an essentially matching sufficient condition.

THEOREM 3.2 (Samples sufficient for structure learning). *Fix an arbitrary tree* T *and Ising model* $P \in \mathcal{P}_{\mathsf{T}}(\alpha, \beta)$. *If the number of samples is* $n > Ce^{2\beta} \tanh^{-2}(\alpha) \log(p/\delta)$, *then with probability at least* $1 - \delta$ *the Chow–Liu algorithm recovers the true tree, that is,* $\mathsf{T}^{\mathsf{CL}} = \mathsf{T}$.

The proof is presented in Section 7.2. Assuming that α is bounded above by a constant (which is the interesting regime), Theorems 3.1 and 3.2 give matching bounds (up to numerical constant) for the sample complexity of learning the tree structure of an Ising model with zero external field. The necessary number of samples increases as the minimum edge weight α decreases, so if edges can be arbitrarily weak, it is impossible to learn the tree given any bounded number of samples. Figures 3(a) and 3(c) in Section 9 present numerical simulation results supporting this observation.

If the goal is merely to make accurate predictions, it is natural to seek a less stringent, approximate notion of learning. Several papers consider learning a model that is close in KL-divergence, for example, [1, 21, 23, 26, 40]. The sample complexity of learning a model to within constant KL-divergence ϵ scales at least *linearly* with the number of variables p, an unrealistic requirement in the high-dimensional setting of interest. Using a number of samples scaling logarithmically in dimension requires relaxing the KL-divergence to scale linearly in p, but this does not imply a nontrivial guarantee on the quality of approximation for marginals of few variables (as done in this paper). The same observation is true for the total variation as the measure of distance. The sample complexity of learning a model to within constant TV distance between the learned model and the original joint distribution over p variables scales at least linearly with p.

In the next section, we study estimation with respect to the small-set TV loss, which captures accuracy of prediction based on few observations. We will see that the associated sample complexity is independent of the edge strength lower bound α in the original model.

3.2. *Small set total variation.* For a subset of nodes $\mathcal{S} \subseteq [p]$, we denote by $P_{\mathcal{S}}$ the marginal distribution $P_{\mathcal{S}}(x_{\mathcal{S}}) = \sum_{x_{\mathcal{V}\setminus\mathcal{S}}} P(x)$.

Given two distributions P and Q on the same space, for each $k \geq 1$ the small-set total variation distance is the maximum total variation over all size k marginals, and is denoted by

$$\mathcal{L}^{(k)}(P, Q) \triangleq \max_{\mathcal{S}:|\mathcal{S}|=k} d_{\mathrm{TV}}(P_{\mathcal{S}}, Q_{\mathcal{S}}).$$

Note that $\mathcal{L}^{(k)}$ is nondecreasing in k. One can check that $\mathcal{L}^{(k)}$ satisfies the triangle inequality: for any three distributions P, R, Q,

$$\mathcal{L}^{(k)}(P, R) + \mathcal{L}^{(k)}(R, Q) \geq \mathcal{L}^{(k)}(P, Q). \tag{3.2}$$

Closeness of P and Q in $\mathcal{L}^{(k)}$ implies that the respective posteriors conditioned on subsets of variables of size $k - 1$ are close on average. To see this, suppose that we wish to

compute $P(X_i = +|X_{\mathcal{S}})$. We measure performance of the approximation Q by the expected magnitude of error $|P(x_i = +|X_{\mathcal{S}}) - Q(x_i = +|X_{\mathcal{S}})|$ averaged over $X_{\mathcal{S}}$:

$$\begin{aligned}
&\mathbb{E}_{X_{\mathcal{S}}}\big|P(X_i = +|X_{\mathcal{S}}) - Q(X_i = +|X_{\mathcal{S}})\big| \\
&\quad = \sum_{x_{\mathcal{S}}}\big|P(X_i = +, X_{\mathcal{S}} = x_{\mathcal{S}}) - Q(X_i = +|X_{\mathcal{S}} = x_{\mathcal{S}})P(X_{\mathcal{S}} = x_{\mathcal{S}})\big| \\
&\quad \le \sum_{x_{\mathcal{S}}}\big|P(X_i = +, X_{\mathcal{S}} = x_{\mathcal{S}}) - Q(X_i = +, X_{\mathcal{S}} = x_{\mathcal{S}})\big| + \sum_{x_{\mathcal{S}}}\big|Q(x_{\mathcal{S}}) - P(x_{\mathcal{S}})\big| \\
&\quad \le 2\mathcal{L}^{(|\mathcal{S}|+1)}(P, Q).
\end{aligned}$$

The last inequality is a consequence of monotonicity of $\mathcal{L}^{(k)}$ in k.

In this paper, we focus mostly on $\mathcal{L}^{(2)}$. Implications for $k \ge 3$ are stated in Corollary 1.2, and discussed in Appendix H. For trees $\mathsf{T}, \widetilde{\mathsf{T}} \in \mathcal{T}$ and distributions $P \in \mathcal{P}_{\mathsf{T}}$ and $\widetilde{P} \in \mathcal{P}_{\widetilde{\mathsf{T}}}$, let $e = (i, j) \in \mathcal{E}_{\mathsf{T}}$, $\mu_e = \mathbb{E}_P X_i X_j$ and for $e' = (i, j) \in \mathcal{E}_{\widetilde{\mathsf{T}}}$, $\widetilde{\mu}_{e'} = \mathbb{E}_{\widetilde{P}} X_i X_j$. It will be useful to express $\mathcal{L}^{(2)}$ as

$$\begin{aligned}
\mathcal{L}^{(2)}(P, \widetilde{P}) &= \max_{w,\widetilde{w}\in\mathcal{V}} \frac{1}{2} \sum_{x_w, x_{\widetilde{w}} \in \{-,+\}^2} \big|P(x_w, x_{\widetilde{w}}) - \widetilde{P}(x_w, x_{\widetilde{w}})\big| \\
&= \max_{w,\widetilde{w}\in\mathcal{V}} \frac{1}{2}\Big| \prod_{e\in \mathsf{path}_{\mathsf{T}}(w,\widetilde{w})} \mu_e - \prod_{e'\in \mathsf{path}_{\widetilde{\mathsf{T}}}(w,\widetilde{w})} \widetilde{\mu}_{e'}\Big|.
\end{aligned} \tag{3.3}$$

The second equality is derived by noting that $P(x_i = +) = 1/2$, $P(x_w, x_{\widetilde{w}}) = [1 + x_w x_{\widetilde{w}} \mathbb{E}_P[X_w X_{\widetilde{w}}]]/4$ and analogously for $\widetilde{P}$. Also, as noted above after (1.1), it is immediate from Lemma 8.6 that $\mathbb{E}_P X_w X_{\widetilde{w}} = \prod_{e\in \mathsf{path}_{\mathsf{T}}(w,\widetilde{w})} \mu_e$. The same holds for $\widetilde{P} \in \mathcal{P}_{\widetilde{\mathsf{T}}}$, which gives (3.3).

3.3. *Main result.* Our main contribution is to prove upper and lower bounds on the number of samples required to estimate a tree close in $\mathcal{L}^{(2)}$ to the true one. An upper bound on the number of samples is obtained for the Chow–Liu algorithm by bounding the expression in (3.3). The Chow–Liu algorithm produces the maximum likelihood tree, which minimizes the expected zero–one loss in (3.1). As shown in Theorem 3.3, the maximum likelihood tree also performs well in terms of accuracy of pairwise marginals.

Recall from (2.2) that $\Pi_{\mathsf{T}}(P)$ is the reverse information projection of the distribution P onto the set of zero-field Ising models on tree T.

THEOREM 3.3 (Learning for predictions using Chow–Liu algorithm). *For $\mathsf{T} \in \mathcal{T}$, let the distribution $P \in \mathcal{P}_{\mathsf{T}}(0, \beta)$. Given $n > C \max\{e^{2\beta}, \eta^{-2}\} \log \frac{p}{\delta}$ samples, if T^{CL} is the Chow–Liu tree as defined in (2.3), then with probability at least $1 - \delta$ we have $\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P})) < \eta$.*

The main challenge is that the number of samples assumed to be available in Theorem 3.3 is not sufficient for structure learning, as can be seen by comparing with Theorem 3.1. This means that accurate marginals must be computed *using possibly the wrong tree*. The proof is sketched in Section 5.

We also lower bound the number of samples necessary for small $\mathcal{L}^{(2)}$ loss. Let the learning algorithm be $\Psi : \{-1, +1\}^{p\times n} \to \mathcal{P}$ where $\mathcal{P} = \bigcup_{\mathsf{T}} \mathcal{P}_{\mathsf{T}}$ is the set of tree-structured Ising models with no external field defined in (1.1).

THEOREM 3.4 (Samples necessary for small ssTV). *Fix $\eta > 0$. Suppose $\tanh(\beta) > \tanh(\alpha) + 4\eta$ and $n < C[1 - (\tanh(\alpha) + 4\eta)^2]\eta^{-2} \log p$. The worst-case probability of $\mathcal{L}^{(2)}$*

loss greater than η, taken over trees $\mathsf{T} \in \mathcal{T}$ and distributions $P \in \mathcal{P}_{\mathsf{T}}(\alpha, \beta)$, is at least half for any algorithm, that is,

$$\inf_{\Psi} \sup_{\substack{\mathsf{T}\in\mathcal{T} \\ P\in\mathcal{P}_{\mathsf{T}}(\alpha,\beta)}} \mathsf{P}\big[\mathcal{L}^{(2)}\big(P, \Psi\big(X^{(1:n)}\big)\big) > \eta\big] > 1/2.$$

Theorem 3.4 is proved in Section 6.3. As noted earlier, the assumption $\tanh(\beta) > \tanh(\alpha) + 2\eta$ captures the scenario that the precise values of the edge parameters are not known a priori. In the extreme where $\alpha = \beta$, the entire problem is quite different (and we believe less realistic). We state a lower bound for this setting in Appendix B, which is not directly comparable to the above theorems.

If α is bounded above by a constant, then Theorems 3.3 and 3.4 have the same dependence on η. The theorems imply the following bounds on risk.

COROLLARY 3.5 (Upper bound for risk). *For $\mathsf{T} \in \mathcal{T}$, let the distribution $P \in \mathcal{P}_{\mathsf{T}}(0, \beta)$. Given n samples with empirical distribution $\widehat{P}$, if the tree T^{CL} is the Chow–Liu tree as defined in (2.3), then*

$$\mathbb{E}\big[\mathcal{L}^{(2)}\big(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P})\big)\big] < C' p \exp\big(-Cne^{-2\beta}\big) + C''\sqrt{\frac{\log p}{n}}.$$

COROLLARY 3.6 (Lower bound for risk). *Suppose $\tanh(\beta) - \tanh(\alpha) > 1/2$ and one observes n samples. Then the minimax risk over trees $\mathsf{T} \in \mathcal{T}$ and distributions $P \in \mathcal{P}_{\mathsf{T}}(\alpha, \beta)$ is lower bounded by*

$$\inf_{\Psi} \sup_{\substack{\mathsf{T}\in\mathcal{T} \\ P\in\mathcal{P}_{\mathsf{T}}(\alpha,\beta)}} \mathbb{E}\big[\mathcal{L}^{(2)}\big(P, \Psi\big(X^{(1:n)}\big)\big)\big] > \min\left\{\frac{1}{24}\sqrt{\frac{\log p}{n}}, \frac{1}{12}\right\}.$$

Corollary 3.5 is proved in Appendix G. Proof of Corollary 3.6 is immediate from the statement of Theorem 3.4. The upper bound in Corollary 3.5 has an extra term that depends on β compared to the lower bound of Corollary 3.6. We conjecture that the lower bound is tight.

4. Illustrative example and algorithm comparison.

4.1. *Three node Markov chain.* A Markov chain with three nodes captures a few of the key ideas developed in this paper. Let $P(X_1, X_2, X_3)$ be represented by the tree T_1 in Figure 1 in which $X_1 \leftrightarrow X_2 \leftrightarrow X_3$ form a Markov chain with correlations μ_{12}, μ_{23} and $\mu_{13} = \mu_{12}\mu_{23}$. Without loss of generality, we assume $\mu_{12}, \mu_{23} > 0$. Suppose that for some small value ϵ, $\mu_{12} = 1 - \mu_{23} = \epsilon$.

Given n samples from P, the empirical correlations $\widehat{\mu}_{12}$, $\widehat{\mu}_{23}$ and $\widehat{\mu}_{13}$ are concentrated around $\mu_{12} = \epsilon$, $\mu_{23} = 1 - \epsilon$ and $\mu_{13} = \mu_{12}\mu_{23} = \epsilon(1 - \epsilon)$. Let $\widehat{\mu}_{12} = \mu_{12} + z_{12}$, $\widehat{\mu}_{23} = \mu_{23} + z_{23}$ and $\widehat{\mu}_{13} = \mu_{12}\mu_{23} + z_{13}$, where the fluctuations of z_{12}, z_{23} and z_{13} shrink as n grows. It is useful to imagine the typical fluctuations of z_{ij} to be on the order $\epsilon/10$.

Since $\max\{\mu_{12}, \mu_{13}\} = \max\{\epsilon, \epsilon(1 - \epsilon)\} = \epsilon \ll \mu_{23}$, concentration bounds guarantee that with high probability $\widehat{\mu}_{23} > \max\{\widehat{\mu}_{12}, \widehat{\mu}_{13}\}$ and the (greedy implementation of) Chow–Liu algorithm described in (2.3) adds the edge $(2, 3)$ to T^{CL}. However, because $\mu_{12} - \mu_{13} = \epsilon^2$ is smaller than the fluctuations of z_{12} and z_{13} there is no guarantee that $\widehat{\mu}_{12} > \widehat{\mu}_{13}$: if $z_{13} - z_{12} > \epsilon^2$, then edge $(1, 3)$ is added and $\mathsf{T}^{\mathsf{CL}} = \mathsf{T}_3 \neq \mathsf{T}_1$.

The preceding discussion provides the intuition underlying a statistical characterization of the possible errors made by the Chow–Liu algorithm. To make this quantitative, later on in

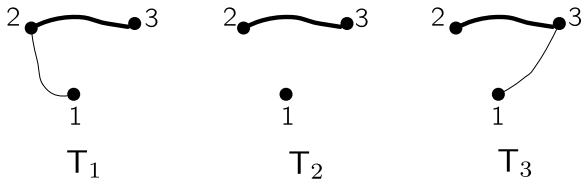

FIG. 1. *The original distribution factors according to* T_1. *The width of the edges corresponds to their strength. The forest approximation algorithm* (*defined in Section* 4.2) *recovers* $\widehat{\mathsf{T}} = \mathsf{T}_2$ *since the correlation between* X_1 *and each of the other variables is not strong enough to confidently recover any edge to node* 1. *The Chow–Liu tree* T^{CL} *is either* T_1 *or* T_3, *depending on the realization of the samples.*

the proof, we determine a value $\tau := \tau(n, \beta, \delta)$ so that any (strong) edge e with $|\mu_e| \geq \tau$ is recovered by the Chow–Liu algorithm with probability at least $1 - \delta$. Equivalently, if there is a mistake made by the Chow–Liu algorithm such that $e \in \mathcal{E}_{\mathsf{T}}$ but $e \notin \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$, then $|\mu_e| \leq \tau$ (i.e., missed edges are weak). This is going to play a key role in bounding the ssTV $\mathcal{L}^{(2)}$ for T^{CL}, whether or not it is equal to T.

In the regime where n is not large enough to guarantee the correct recovery of all the edges in the tree, there are two natural strategies:

I. Forest approximation algorithm. This algorithm attempts to recover a forest F, a good approximation of the original tree T in the sense that $\mathcal{E}_{\mathsf{F}} \subseteq \mathcal{E}_{\mathsf{T}}$. This is accomplished by finding a forest consisting of sufficiently strong edges, that is, having weight at least τ for an appropriate value of τ. As shown by Tan et al. in [40], such a forest can be obtained by running the Chow–Liu algorithm and removing the edges with weight below τ. The details of this algorithm and its sample complexity will be discussed in Section 4.2.

II. Chow–Liu algorithm. We can use the Chow–Liu tree as our estimated structure despite the fact that it may well be incorrect.

For our three-node example, the forest approximation algorithm would return $\widehat{\mathsf{T}} = \mathsf{T}_2$ in Figure 1, whereas Chow–Liu would give $\widehat{\mathsf{T}} = \mathsf{T}_1$ or $\widehat{\mathsf{T}} = \mathsf{T}_3$. To focus on the implication of graph structure estimation (as opposed to parameter estimation), we compare the *loss due to graph estimation error*, defined as $\mathcal{L}^{(2)}(P, \Pi_{\widehat{\mathsf{T}}}(P))$, for the above cases:

$$\begin{aligned}
\widehat{\mathsf{T}} = \mathsf{T}_1 \quad &\to \quad \mathcal{L}^{(2)}\big(P, \Pi_{\widehat{\mathsf{T}}}(P)\big) = 0, \\
\widehat{\mathsf{T}} = \mathsf{T}_2 \quad &\to \quad \mathcal{L}^{(2)}\big(P, \Pi_{\widehat{\mathsf{T}}}(P)\big) = |\mu_{12}| = \epsilon, \\
\widehat{\mathsf{T}} = \mathsf{T}_3 \quad &\to \quad \mathcal{L}^{(2)}\big(P, \Pi_{\widehat{\mathsf{T}}}(P)\big) = |\mu_{12}|\big(1 - \mu_{23}^2\big) = \epsilon^2(2 - \epsilon).
\end{aligned}$$

Evidently, the loss due to graph estimation error in the forest approximation algorithm is bigger than the Chow–Liu algorithm, whether or not the latter recovers the true tree. This is because the Chow–Liu algorithm does not make arbitrary errors in estimating the tree: errors happen when both the original tree and the estimated tree describe the original distribution rather well. Theorem 3.3 makes this formal.

4.2. *Forest approximation.* In the regime where exact recovery of the tree is impossible, a reasonable goal is to instead find a forest approximation to the tree. In this section, we analyze a natural truncation algorithm, which thresholds to zero edges with correlation below a specified value $\tau(\epsilon)$.

There is extensive literature on estimating forests in the fully observed setting of this paper, including [26, 40]. Mossel in [30] studied the problem of learning phylogenetic forests, where samples are only observed at the leaves of the tree. They quantified the idea that most edges of phylogenies are easy to reconstruct. In the regime that the sample complexity of structure learning is too high, they instead estimate a forest. An upper bound on the number of edges

necessary to glue together the forest to get the original tree is provided as a function of the number of leaves, the minimum edge weight, and the metric distortion bounds. Our results in this section are consistent with the asymptotic conditions on the thresholds given in [40] by Tan et al. for forest approximation of general distributions over trees.

The forest approximation algorithm considered in this section thresholds to zero the edges with correlations below $\tau(\epsilon) = \frac{4\epsilon}{\sqrt{1-\tanh\beta}}$ for $\epsilon = \sqrt{2/n \log(2p^2/\delta)}$. This is equivalent to finding the maximum-weight spanning forest over the complete graph with edge weights $|\widehat{\mu}_e| - \tau(\epsilon) - \epsilon$. The output $\widehat{\mathsf{T}} = (\mathcal{V}, \mathcal{E}_{\widehat{\mathsf{T}}})$ is a truncated version of T^{CL} such that the empirical correlation between any pair of nodes $(i, j) \in \mathcal{E}_{\widehat{\mathsf{T}}}$ satisfies $|\widehat{\mu}_{ij}| \geq \tau(\epsilon) + \epsilon$.

PROPOSITION 4.1. *Given $n > Ce^{2\beta}\eta^{-2}\log\frac{p}{\delta}$ samples, the forest approximation algorithm guarantees that $\mathcal{E}_{\widehat{\mathsf{T}}} \subseteq \mathcal{E}_{\mathsf{T}}$ and $\mathcal{L}^{(2)}(P, \Pi_{\widehat{\mathsf{T}}}(\widehat{P})) < \eta$ with probability at least $1 - \delta$.*

The proof is presented in Appendix C in the Supplementary Material. The forest approximation algorithm is trying to avoid adding incorrect edges and we then measure its performance according to $\mathcal{L}^{(2)}$. Given this objective, it is natural to consider the loss

$$\widetilde{d}(P, Q) = \max\{\mathcal{L}^{(2)}(P, Q), \mathbb{1}[\mathcal{E}_{\mathsf{F}} \not\subseteq \mathcal{E}_{\mathsf{T}}]\}, \tag{4.1}$$

which is one if the learned forest is not a subgraph of the original tree, and otherwise is equal to $\mathcal{L}^{(2)}$.

It turns out that the forest approximation algorithm is optimal (up to constant factor) for this loss, as shown in the following proposition.

PROPOSITION 4.2. *Fix $\eta > 0$. Let Ψ be an estimator that takes n samples generated from $\mathsf{P}_{\mathsf{T}}(0, \beta)$ and recovers a forest F and a distribution $Q \in \mathsf{P}_{\mathsf{F}}(0, \beta)$. Let the (asymmetric) distance $\widetilde{d}$ be defined in (4.1). If $n < C \min\{p, e^{2\beta}\}/[\eta \operatorname{atanh} \eta)] \log p$, then the worst-case probability of $\widetilde{d}$ loss greater than η, taken over trees $\mathsf{T} \in \mathcal{T}$ and distributions $P \in \mathcal{P}_{\mathsf{T}}(\alpha, \beta)$, is at least half for any algorithm, that is,*

$$\inf_{\Psi} \sup_{\substack{\mathsf{T} \in \mathcal{T} \\ P \in \mathcal{P}_{\mathsf{T}}(\alpha,\beta)}} \mathsf{P}\big[\widetilde{d}\big(P, \Psi\big(X^{(1:n)}\big)\big) > \eta\big] > 1/2.$$

Comparison with Theorem 3.3 shows that avoiding adding wrong edges (as forced by the loss function (4.1)) entails a degradation in performance.

5. Outline of proof. We now sketch the argument for the main result, Theorem 3.3, guaranteeing accurate pairwise marginals in the Chow–Liu tree. The starting point is an application of the triangle inequality (3.2):

$$\mathcal{L}^{(2)}\big(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P})\big) \leq \mathcal{L}^{(2)}\big(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(P)\big) + \mathcal{L}^{(2)}\big(\Pi_{\mathsf{T}^{\mathsf{CL}}}(P), \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P})\big). \tag{5.1}$$

The first term on the right-hand side of equation (5.1) represents the error due to the difference in the structure of T and T^{CL}, so we call this first term the *loss due to graph estimation error*. Equation (3.3) tells us that for each pair of nodes $u, v \in \mathcal{V}$, $\mathsf{path}_{\mathsf{T}}(u, v)$ and $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u, v)$ must be compared.

The second term on the right-hand side of equation (5.1) represents the propagation of error due to inaccuracy in estimated parameters. Recall that the estimated parameters on the Chow–Liu tree are obtained by matching correlations to the empirical values.

Theorem 3.3 follows by separately bounding each term on the right-hand side of equation (5.1).

PROPOSITION 5.1 (Loss due to parameter estimation error). *Given* $n > C\max\{e^{2\beta}, \eta^{-2}\}\log\frac{p}{\delta}$ *samples, with probability at least* $1-\delta$ *we have* $\mathcal{L}^{(2)}(\Pi_{\mathsf{T}^{\mathsf{CL}}}(P), \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P}) \le \eta$.

PROPOSITION 5.2 (Loss due to graph estimation error). *Given* $n > C'\max\{e^{2\beta}, \eta^{-2}\}\log\frac{p}{\delta}$ *samples, with probability at least* $1-\delta$ *we have* $\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(P)) \le \eta$.

These two propositions are proved in full detail in Sections 6.1 and 6.2. In the remainder of this section, we define probabilistic events of interest and sketch the proofs of the propositions.

We define three highly probable events $\mathrm{E}^{\mathrm{corr}}(\epsilon)$, $\mathrm{E}^{\mathrm{strong}}(\epsilon)$ and $\mathrm{E}^{\mathrm{cascade}}(\epsilon)$ as follows. Let $\mathrm{E}^{\mathrm{corr}}(\epsilon)$ be the event that all empirical correlations are within ϵ of population values:

$$\mathrm{E}^{\mathrm{corr}}(\epsilon) = \Big\{ \max_{w,\widetilde{w}\in\mathcal{V}} |\mu_{w,\widetilde{w}} - \widehat{\mu}_{w,\widetilde{w}}| \le \epsilon \Big\}. \tag{5.2}$$

Let

$$\tau(\epsilon) = \frac{4\epsilon}{\sqrt{1-\tanh\beta}} \quad \text{and} \quad \mathcal{E}_{\mathsf{T}}^{\mathrm{strong}}(\epsilon) = \big\{(i,j) \in \mathcal{E}_{\mathsf{T}} : |\mu_{ij}| > \tau(\epsilon)\big\} \tag{5.3}$$

consist of the set of strong edges in tree T. Let T^{CL} be the Chow–Liu tree defined in equation (2.3). Weak edges are those that are not strong, that is, $\mathcal{E}_{\mathsf{T}} \setminus \mathcal{E}_{\mathsf{T}}^{\mathrm{strong}}(\epsilon)$. Let $\mathrm{E}^{\mathrm{strong}}(\epsilon)$ be the event that all strong edges in T as defined in (5.3) are recovered by the Chow–Liu tree:

$$\mathrm{E}^{\mathrm{strong}}(\epsilon) = \big\{\mathcal{E}_{\mathsf{T}}^{\mathrm{strong}}(\epsilon) \subset \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}\big\}. \tag{5.4}$$

Finally, define the event

$$\mathrm{E}^{\mathrm{cascade}}(\epsilon) = \big\{\mathcal{L}^{(2)}\big(P, \Pi_{\mathsf{T}}(\widehat{P})\big) \le \epsilon\big\}. \tag{5.5}$$

Recall that P factorizes according to T and $\widehat{P}$ is the empirical distribution which does not factorize according to any tree. This event controls *cascades* of errors in correlations computed along paths in T.

Since we are interested in the situation that all three events hold, let $\mathrm{E}(\epsilon,\gamma) := \mathrm{E}^{\mathrm{corr}}(\epsilon) \cap \mathrm{E}^{\mathrm{strong}}(\epsilon) \cap \mathrm{E}^{\mathrm{cascade}}(\gamma)$. Lemmas 8.1, 8.5 and 8.7 prove that for

$$\epsilon_0 = \min\big\{e^{-\beta}/24, \eta/16\big\} \quad \text{and} \quad \gamma_0 = \eta/3, \tag{5.6}$$

if $n > \max\{1152e^{2\beta}, 512\eta^{-2}\}\log(6p^3/\delta) := n_0$, then

$$\mathsf{P}\big[\mathrm{E}(\epsilon_0,\gamma_0)\big] \ge 1-\delta. \tag{5.7}$$

Sketch of proof of Proposition 5.1. The proof entails showing that on the event $\mathrm{E}(\epsilon_0,\gamma_0)$ for ϵ_0 and γ_0 defined in (5.6) we have the desired inequality $\mathcal{L}^{(2)}(\Pi_{\mathsf{T}^{\mathsf{CL}}}(P), \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P})) \le \eta$. The result then follows from (5.7).

To bound $\mathcal{L}^{(2)}(\Pi_{\mathsf{T}^{\mathsf{CL}}}(P), \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P}))$ on event $\mathrm{E}(\epsilon_0,\gamma_0)$, we consider parameter estimation errors along paths in T^{CL}. First, observe that on event $\mathrm{E}^{\mathrm{cascade}}(\gamma_0)$ defined in (5.5), the end-to-end error for each path in $\mathcal{E}_{\mathsf{T}^{\mathsf{CL}}} \cap \mathcal{E}_{\mathsf{T}}$ is bounded by γ_0. Next, we study parameter estimation error in paths containing (falsely added) edges in $\mathcal{E}_{\mathsf{T}^{\mathsf{CL}}} \setminus \mathcal{E}_{\mathsf{T}}$.

For any pair of nodes $w, \widetilde{w}$, denote by $t = |\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w}) \setminus \mathcal{E}_{\mathsf{T}}|$ the number of falsely added edges in the path connecting them in T^{CL}. As discussed in the proof, these edges correspond to missed edges in T, and thus $\mathrm{E}^{\mathrm{strong}}(\epsilon_0)$ guarantees that these edges are weak (as defined after (5.3)). These t weak edges break up the $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})$ into at most $t+1$ contiguous segments $\mathcal{F}_0, \mathcal{F}_1, \ldots, \mathcal{F}_t$, each entirely within $\mathcal{E}_{\mathsf{T}^{\mathsf{CL}}} \cap \mathcal{E}_{\mathsf{T}}$.

On event $\mathrm{E}^{\mathrm{cascade}}(\gamma_0)$ the error on each segment $\mathcal{F}_i$, is bounded by γ_0, but now there are $t+1$ such segments and the errors may add up. This effect is counterbalanced by the fact that the falsely added edges are weak and hence scale down the error in a multiplicative fashion.

Sketch of proof of Proposition 5.2. We want to show that $\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\mathsf{CL}}}(P)) \le \eta$ on the event $\mathbb{E}^{\mathrm{corr}}(\epsilon_0) \cap \mathbb{E}^{\mathrm{strong}}(\epsilon_0) \supseteq \mathbb{E}(\epsilon_0, \gamma_0)$ for ϵ_0 defined in (5.6).

The proof of the proposition sets up a careful induction on the distance between nodes (computed in T^{CL}) for which we wish to bound the error in correlation. One of the ingredients is Lemma 8.8, a combinatorial statement relating trees T and T^{CL}. The lemma states that for any two spanning trees on p nodes, and for any two nodes $w, \widetilde{w} \in [p]$, there exists at least one pair of edges $f \in \mathsf{path}_{\mathsf{T}}(w, \widetilde{w})$ and $g \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$ satisfying a collection of properties illustrated in Figure 2 (and specified in the lemma).

A consequence is that the true correlation across g according to P can be expressed in terms of the correlation on f as $\mu_g = \mu_f \mu_{\mathcal{A}} \mu_C \mu_{\widetilde{\mathcal{A}}} \mu_{\widetilde{C}}$, hence $|\mu_g| \le |\mu_f|$. But $|\widehat{\mu}_g| \ge |\widehat{\mu}_f|$, since the Chow–Liu algorithm chose g in T^{CL} instead of f. Tight control on the relationship between $|\widehat{\mu}_g|$ and $|\widehat{\mu}_f|$ yields a recurrence for $\Delta(d)$, where $\Delta(d)$ is an upper bound on the error due to graph estimation error for any pair of nodes w, $\widetilde{w}$ with $|\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})| = d$.

6. Proof of main result. As observed in Section 5, Theorem 3.3 is a direct consequence of Propositions 5.1 and 5.2, which we prove in Sections 6.1 and 6.2. We prove Theorem 3.4 in Section 6.3.

6.1. *Loss due to parameter estimation* (*proof of Proposition 5.1*). We will prove that on the event $\mathbb{E}(\epsilon_0, \gamma_0)$ for ϵ_0 and γ_0 defined in (5.6) the desired inequality $\mathcal{L}^{(2)}(\Pi_{\mathsf{T}^{\mathsf{CL}}}(P), \Pi_{\mathsf{T}^{\mathsf{CL}}}(\widehat{P})) \le \eta$ holds. Equation (5.7) gives the result.

Let $\tau(\epsilon_0)$ (defined in (5.3)) be the threshold to define $\mathcal{E}_{\mathsf{T}}^{\mathrm{strong}}(\epsilon_0)$, the set of strong edges in T. For any pair of nodes w, $\widetilde{w}$, consider $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$. Let $0 \le t < p$ be the number of weak edges $e_1, \dots, e_t \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$ such that $|\mu_{e_i}| \le \tau(\epsilon_0)$. There are at most $t+1$ contiguous subpaths in $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$ consisting of strong edges. We call these segments $\mathcal{F}_0, \mathcal{F}_1, \dots, \mathcal{F}_t$. If two weak edges e_i and e_{i+1} are adjacent in $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$, then $\mathcal{F}_i = \varnothing$, in which case we define $\mu_{\mathcal{F}_i} = \widehat{\mu}_{\mathcal{F}_i} = 1$. By definition of $\mathcal{F}_i$, all edges $f \in \mathcal{F}_i$ are strong.

According to (5.4), under the event $\mathbb{E}^{\mathrm{strong}}(\epsilon_0)$ all strong edges in T are recovered in T^{CL}. Thus, $\mathcal{F}_i \subseteq \mathcal{E}_{\mathsf{T}}$ is a path not only in T^{CL} but also in T, which guarantees $|\widehat{\mu}_{\mathcal{F}_i} - \mu_{\mathcal{F}_i}| \le \gamma_0$ under the event $\mathbb{E}^{\mathrm{cascade}}(\gamma_0)$.

Note that if $t = 0$ then $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})$ consists of all strong edges for which Lemma 8.7 gives the desired bound. For $t \ge 1$, we have:

$$
\begin{aligned}
&\Bigg| \prod_{e \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})} \widehat{\mu}_e - \prod_{e \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w, \widetilde{w})} \mu_e \Bigg| \\
&\quad \overset{(a)}{=} \Bigg| \widehat{\mu}_{\mathcal{F}_0} \prod_{i=1}^{t} \widehat{\mu}_{\mathcal{F}_i} \widehat{\mu}_{e_i} - \mu_{\mathcal{F}_0} \prod_{i=1}^{t} \mu_{\mathcal{F}_i} \mu_{e_i} \Bigg| \\
&\quad \overset{(b)}{\le} |\widehat{\mu}_{\mathcal{F}_0} - \mu_{\mathcal{F}_0}| \prod_{j=1}^{t} |\mu_{\mathcal{F}_j} \mu_{e_j}| \\
&\qquad + \sum_{i=1}^{t} |\widehat{\mu}_{\mathcal{F}_i} \widehat{\mu}_{e_i} - \mu_{\mathcal{F}_i} \mu_{e_i}| \cdot |\widehat{\mu}_{\mathcal{F}_0}| \prod_{j=1}^{i-1} |\widehat{\mu}_{\mathcal{F}_j} \widehat{\mu}_{e_j}| \prod_{k=i+1}^{t} |\mu_{\mathcal{F}_k} \mu_{e_k}| \\
&\quad \overset{(c)}{\le} \gamma_0 \big[\tau(\epsilon_0)\big]^t + \big(\tau(\epsilon_0) + \epsilon_0\big)^{t-1} \sum_{i=1}^{t} |\widehat{\mu}_{\mathcal{F}_i} \widehat{\mu}_{e_i} - \mu_{\mathcal{F}_i} \mu_{e_i}| \\
&\quad \overset{(d)}{\le} \gamma_0 \big[\tau(\epsilon_0)\big]^t + \big(\tau(\epsilon_0) + \epsilon_0\big)^{t-1} \Bigg[\sum_{i=1}^{t} |\mu_{\mathcal{F}_i} (\widehat{\mu}_{e_i} - \mu_{e_i})| + |\widehat{\mu}_{e_i} (\widehat{\mu}_{\mathcal{F}_i} - \mu_{\mathcal{F}_i})| \Bigg]
\end{aligned}
$$

$$\overset{(e)}{\leq} \big(\tau(\epsilon_0)+\epsilon_0\big)^{t-1}(2t+1)\max\{\gamma_0,\epsilon_0\} \overset{(f)}{\leq} \frac{2t+1}{4^{t-1}}\frac{\eta}{3} \overset{(g)}{\leq} \eta.$$

In (a), we use $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w}) = \{\mathcal{F}_0, e_1, \ldots, \mathcal{F}_t, e_t, \mathcal{F}_t\}$. (b) uses the bound $|\prod_{i=1}^t a_i - \prod_{i=1}^t b_i| \leq \sum_{i=1}^t |a_i - b_i| \prod_{j=1}^{i-1} |a_j| \prod_{k=i+1}^t |b_k|$ obtained via telescoping sum and triangle inequality. In (c), we use $|\mu_{\mathcal{F}_i}|, |\widehat{\mu}_{\mathcal{F}_i}| \leq 1$, $|\mu_{e_i}| \leq \tau(\epsilon_0)$, $|\widehat{\mu}_{e_i}| \leq \tau(\epsilon_0)+\epsilon_0$ on $\mathbb{E}^{\mathrm{corr}}(\epsilon_0)$ and $|\widehat{\mu}_{\mathcal{F}_0} - \mu_{\mathcal{F}_0}| \leq \gamma_0$ on $\mathbb{E}^{\mathrm{cascade}}(\gamma_0)$. (d) uses triangle inequality. In (e), we use $|\widehat{\mu}_{\mathcal{F}_i} - \mu_{\mathcal{F}_i}| \leq \gamma_0$ on the event $\mathbb{E}^{\mathrm{cascade}}(\gamma_0)$ and $|\widehat{\mu}_{e_i} - \mu_{e_i}| \leq \epsilon_0$ on the event $\mathbb{E}^{\mathrm{corr}}(\epsilon_0)$. (f) is true since $\gamma_0, \epsilon_0 \leq \eta/3$ (as in (5.6)). Also, $1-\tanh\beta \geq e^{-2\beta}$ and the definition of $\tau(\epsilon_0)$ in (5.3) gives $\tau(\epsilon_0) \leq 4\epsilon_0 e^{\beta}$. Hence, $\tau(\epsilon_0)+\epsilon_0 \leq 5\epsilon_0 e^{\beta} \leq 1/4$ where the last inequality uses $\epsilon_0 \leq e^{-\beta}/20$ (according to (5.6)). (g) holds for all $t \geq 1$.

6.2. *Loss due to graph estimation error* (*proof of Proposition* 5.2). The following lemma, proved in Section 8 for completeness, is a well-known consequence of a spanning tree being max-weight [14]. We use Lemma 6.1 to bound the loss due to graph estimation error by the Chow–Liu algorithm.

LEMMA 6.1 (Error characterization in the Chow–Liu tree). *Consider the complete graph on p nodes with weights $|\widehat{\mu}_{ij}|$ on each edge (i,j). Let T^{CL} be the maximum weight spanning tree of this graph. If edge $(u,\widetilde{u}) \notin \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$, then $|\widehat{\mu}_{u\widetilde{u}}| \leq |\widehat{\mu}_{ij}|$ for all $(i,j) \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u,\widetilde{u})$.*

Recall that P factorizes according to T. The error in correlation between any two variables X_w and $X_{\widetilde{w}}$ computed along the $\mathsf{path}_{\mathsf{T}}(w,\widetilde{w})$ as compared to $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})$ is

$$\begin{aligned}\mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w}) &= \frac{1}{2}\cdot\big|\mathbb{E}_P X_w X_{\widetilde{w}} - \mathbb{E}_{\Pi_{\mathsf{T}^{\mathsf{CL}}}(P)} X_w X_{\widetilde{w}}\big| \\ &= \frac{1}{2}\cdot\Big|\prod_{e\in\mathsf{path}_{\mathsf{T}}(w,\widetilde{w})}\mu_e - \prod_{e\in\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})}\mu_e\Big|.\end{aligned}$$

Our goal is to bound $\mathcal{L}^{(2)}(P,\Pi_{\mathsf{T}^{\mathsf{CL}}}(P)) = \max_{w,\widetilde{w}\in\mathcal{V}} \mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})$. We will prove that on the event $\mathbb{E}^{\mathrm{corr}}(\epsilon_0)\cap\mathbb{E}^{\mathrm{strong}}(\epsilon_0) \supseteq \mathbb{E}(\epsilon_0,\gamma_0)$ for ϵ_0 defined in (5.6), $\mathcal{L}^{(2)}(P,\Pi_{\mathsf{T}^{\mathsf{CL}}}(P)) < \eta$ holds. The result then follows from (5.7).

The core of the argument uses induction to derive a recurrence on the maximum of $\mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})$ in terms of the distance (as measured in T^{CL}) between the nodes w, $\widetilde{w}$. Define

$$\Delta(d) \triangleq \max_{\substack{w,\widetilde{w}\in\mathcal{V}\\ |\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})|=d}} \mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w}).$$

For nodes at distance one in T^{CL}, that is, $|\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})| = 1$, it follows that $\mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w}) = 0$ from the definition of the projected distribution $\Pi_{\mathsf{T}^{\mathsf{CL}}}(P)$ (matching pairwise marginals on edges) in Section 2.3. Hence, $\Delta(1) = 0 \leq \eta$. We define $\Delta(0) = 0$. For $d > 1$, we bound $\Delta(d)$ in terms of $\Delta(k)$ for $k < d$: we will show that on the event $\mathbb{E}^{\mathrm{corr}}(\epsilon_0)\cap\mathbb{E}^{\mathrm{strong}}(\epsilon_0)$ with ϵ_0 defined in (5.6), if $\Delta(k) \leq \eta$ for all $k < d$, then $\Delta(d) \leq \eta$ which gives the result.

Note that if $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w}) = \mathsf{path}_{\mathsf{T}}(w,\widetilde{w})$, then $\mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w}) = 0$ (again because correlations are matched on edges). Thus, we assume $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w}) \neq \mathsf{path}_{\mathsf{T}}(w,\widetilde{w})$. Lemma 8.8 shows the existence of a pair of edges $f = (u,\widetilde{u}) \in \mathcal{E}_{\mathsf{T}} \setminus \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$ and $g = (v,\widetilde{v}) \in \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}} \setminus \mathcal{E}_{\mathsf{T}}$ such that (see Figure 2):

- $f \in \mathsf{path}_{\mathsf{T}}(w,\widetilde{w}) \cap \mathsf{path}_{\mathsf{T}}(v,\widetilde{v})$ and $g \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w}) \cap \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u,\widetilde{u})$.
- $f \notin \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})$ and $g \notin \mathsf{path}_{\mathsf{T}}(w,\widetilde{w})$.
- $u, v \in \mathsf{SubTree}_{\mathsf{T},f}(w)$ and $\widetilde{u}, \widetilde{v} \in \mathsf{SubTree}_{\mathsf{T},f}(\widetilde{w})$.

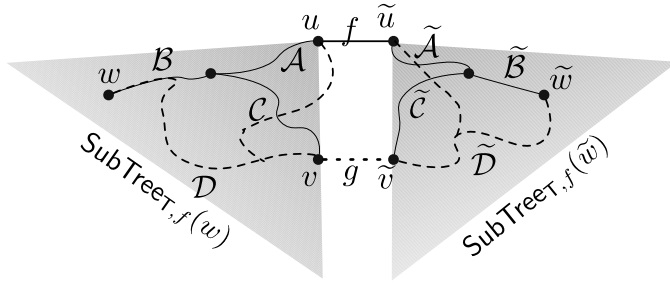

FIG. 2. *Schematic for the proof of Proposition 5.2. The solid lines represent paths in* T *and the dashed lines represent paths in* T^{CL}. *The sets of edges* $\mathcal{D}$ *and* $\widetilde{\mathcal{D}}$ *may overlap with* $\mathcal{A}\cup\mathcal{B}$ *and* $\widetilde{\mathcal{A}}\cup\widetilde{\mathcal{B}}$.

Here, $\mathsf{SubTree}_{\mathsf{T},f}(w)=\{i\in\mathcal{V};\ f\notin\mathsf{path}_{\mathsf{T}}(w,i)\}$ is the set of nodes connected to w in T after removing edge f (see Figure 2). We define several subpaths:

$$\mathcal{A}=\mathsf{path}_{\mathsf{T}}(u,w)\cap\mathsf{path}_{\mathsf{T}}(u,v),\qquad \mathcal{B}=\mathsf{path}_{\mathsf{T}}(u,w)\setminus\mathsf{path}_{\mathsf{T}}(u,v),$$
$$\mathcal{C}=\mathsf{path}_{\mathsf{T}}(u,v)\setminus\mathsf{path}_{\mathsf{T}}(u,w),\qquad \mathcal{D}=\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,v).$$

Recall that for set of edges $\mathcal{S}$, we defined $\mu_{\mathcal{S}}=\prod_{e\in\mathcal{S}}\mu_e$. Since $\mathsf{path}_{\mathsf{T}}(w,v)=\mathcal{B}\cup\mathcal{C}$ and $\mathcal{B}\cap\mathcal{C}=\varnothing$, we have $\mu_{w,v}=\mu_{\mathcal{B}}\mu_{\mathcal{C}}$. Similarly, in $\mathsf{SubTree}_{\mathsf{T},f}(\widetilde{w})$ we define

$$\widetilde{\mathcal{A}}=\mathsf{path}_{\mathsf{T}}(\widetilde{u},\widetilde{w})\cap\mathsf{path}_{\mathsf{T}}(\widetilde{u},\widetilde{v}),\qquad \widetilde{\mathcal{B}}=\mathsf{path}_{\mathsf{T}}(\widetilde{u},\widetilde{w})\setminus\mathsf{path}_{\mathsf{T}}(\widetilde{u},\widetilde{v}),$$
$$\widetilde{\mathcal{C}}=\mathsf{path}_{\mathsf{T}}(\widetilde{u},\widetilde{v})\setminus\mathsf{path}_{\mathsf{T}}(\widetilde{u},\widetilde{w}),\qquad \widetilde{\mathcal{D}}=\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(\widetilde{w},\widetilde{v}).$$

The sets are defined so that $\mathsf{path}_{\mathsf{T}}(v,\widetilde{v})=\mathcal{C}\cup\mathcal{A}\cup\{f\}\cup\widetilde{\mathcal{C}}\cup\widetilde{\mathcal{A}}$ where $f=(u,\widetilde{u})$ and $g=(v,\widetilde{v})\in\mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$. Thus, $\mu_g=\mu_f\mu_{\mathcal{A}}\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{A}}}\mu_{\widetilde{\mathcal{C}}}$. Since $\mathsf{path}_{\mathsf{T}}(w,\widetilde{w})=\mathcal{A}\cup\mathcal{B}\cup\{f\}\cup\widetilde{\mathcal{A}}\cup\widetilde{\mathcal{B}}$ and $\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})=\mathcal{D}\cup\{g\}\cup\widetilde{\mathcal{D}}$, our goal amounts to finding an upper bound for the quantity $|\mu_{\mathcal{D}}\mu_g\mu_{\widetilde{\mathcal{D}}}-\mu_{\mathcal{A}}\mu_{\mathcal{B}}\mu_f\mu_{\widetilde{\mathcal{A}}}\mu_{\widetilde{\mathcal{B}}}|$.

Lemma 6.1 applied to $f=(u,\widetilde{u})\notin\mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$ and $g=(v,\widetilde{v})\in\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u,\widetilde{u})$ gives $|\widehat{\mu}_f|\le|\widehat{\mu}_g|$. Also, $f\in\mathsf{path}_{\mathsf{T}}(v,\widetilde{v})$, hence $|\mu_g|\le|\mu_f|$. On the event $\mathrm{E}^{\mathrm{corr}}(\epsilon_0)$, $|\mu_f|-2\epsilon_0\le|\widehat{\mu}_f|\le|\widehat{\mu}_g|\le|\mu_g|+2\epsilon_0\le|\mu_f|+2\epsilon_0$ which gives $|\mu_f|-4\epsilon_0\le|\mu_f\mu_{\mathcal{A}}\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{A}}}\mu_{\widetilde{\mathcal{C}}}|\le|\mu_f|$. Also, $|\mu_{\mathcal{A}}\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{A}}}\mu_{\widetilde{\mathcal{C}}}|\le|\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{C}}}|\le 1$. Thus,

$$\begin{aligned}|\mu_f|(1-\mu_{\mathcal{C}}^2\mu_{\widetilde{\mathcal{C}}}^2)&\le 2|\mu_f|(1-|\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{C}}}|)\\&\le 2|\mu_f|(1-|\mu_{\mathcal{A}}\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{A}}}\mu_{\widetilde{\mathcal{C}}}|)\le 8\epsilon_0.\end{aligned}\tag{6.1}$$

Since $f\in\mathcal{E}_{\mathsf{T}}\setminus\mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$, under the event $\mathrm{E}^{\mathrm{strong}}(\epsilon_0)$, f cannot be a strong edge as defined in (5.3). It follows that $|\mu_f|\le\tau(\epsilon_0)$ for $\tau(\epsilon_0)$ defined in (5.3).

Let $k=|\mathcal{D}|=|\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(w,v)|$ and $\widetilde{k}=|\widetilde{\mathcal{D}}|=|\mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(\widetilde{w},\widetilde{v})|$, so $d=k+\widetilde{k}+1$. By definition of $\Delta(\cdot)$,

$$\begin{aligned}\mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w,v)&=|\mu_{\mathcal{D}}-\mu_{\mathcal{B}}\mu_{\mathcal{C}}|\le\Delta(k),\\\mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(\widetilde{w},\widetilde{v})&=|\mu_{\widetilde{\mathcal{D}}}-\mu_{\widetilde{\mathcal{B}}}\mu_{\widetilde{\mathcal{C}}}|\le\Delta(\widetilde{k}).\end{aligned}\tag{6.2}$$

We now prove $\mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})\le\eta$ assuming inductively $\mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(i,j)\le\eta$ for all pairs i,j such that $\mathrm{dist}_{\mathsf{T}^{\mathsf{CL}}}(i,j)+1\le d=\mathrm{dist}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})$ (where $\mathrm{dist}_{\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})$ denotes the graph distance in T^{CL}). Using $\mu_g=\mu_{\mathcal{A}}\mu_{\mathcal{C}}\mu_f\mu_{\widetilde{\mathcal{A}}}\mu_{\widetilde{\mathcal{C}}}$,

$$\begin{aligned}&\mathsf{error}_{P,\mathsf{T}^{\mathsf{CL}}}(w,\widetilde{w})\\&\quad=|\mu_{\mathcal{D}}\mu_g\mu_{\widetilde{\mathcal{D}}}-\mu_{\mathcal{A}}\mu_{\mathcal{B}}\mu_f\mu_{\widetilde{\mathcal{A}}}\mu_{\widetilde{\mathcal{B}}}|\\&\quad=|\mu_{\mathcal{D}}\mu_{\mathcal{A}}\mu_{\mathcal{C}}\mu_f\mu_{\widetilde{\mathcal{A}}}\mu_{\widetilde{\mathcal{C}}}\mu_{\widetilde{\mathcal{D}}}-\mu_{\mathcal{A}}\mu_{\mathcal{B}}\mu_f\mu_{\widetilde{\mathcal{A}}}\mu_{\widetilde{\mathcal{B}}}|\end{aligned}$$

$$
\begin{aligned}
&= |\mu_{\mathcal{A}}\mu_f\mu_{\widetilde{\mathcal{A}}}| \cdot \big|\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{C}}}(\mu_{\mathcal{D}} - \mu_{\mathcal{B}}\mu_{\mathcal{C}} + \mu_{\mathcal{B}}\mu_{\mathcal{C}})(\mu_{\widetilde{\mathcal{D}}} - \mu_{\widetilde{\mathcal{B}}}\mu_{\widetilde{\mathcal{C}}} + \mu_{\widetilde{\mathcal{B}}}\mu_{\widetilde{\mathcal{C}}}) - \mu_{\mathcal{B}}\mu_{\widetilde{\mathcal{B}}}\big| \\
&\le |\mu_{\mathcal{A}}\mu_f\mu_{\widetilde{\mathcal{A}}}| \cdot \big[|\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{C}}}\mu_{\mathcal{B}}\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{B}}}\mu_{\widetilde{\mathcal{C}}} - \mu_{\mathcal{B}}\mu_{\widetilde{\mathcal{B}}}| \\
&\quad + \big|\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{C}}}(\mu_{\mathcal{D}} - \mu_{\mathcal{B}}\mu_{\mathcal{C}})(\mu_{\mathcal{D}} - \mu_{\mathcal{B}}\mu_{\mathcal{C}})\big| \\
&\quad + \big|\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{C}}}\mu_{\widetilde{\mathcal{B}}}\mu_{\widetilde{\mathcal{C}}}(\mu_{\mathcal{D}} - \mu_{\mathcal{B}}\mu_{\mathcal{C}})\big| + \big|\mu_{\mathcal{C}}\mu_{\widetilde{\mathcal{C}}}\mu_{\mathcal{B}}\mu_{\mathcal{C}}(\mu_{\widetilde{\mathcal{D}}} - \mu_{\widetilde{\mathcal{B}}}\mu_{\widetilde{\mathcal{C}}})\big|\big] \\
&\overset{(a)}{\le} |\mu_f\mu_{\mathcal{A}}\mu_{\widetilde{\mathcal{A}}}\mu_{\mathcal{B}}\mu_{\widetilde{\mathcal{B}}}|\big|\mu_{\mathcal{C}}^2\mu_{\widetilde{\mathcal{C}}}^2 - 1\big| + |\mu_f|\big(\Delta(k)\Delta(\widetilde{k}) + \Delta(k) + \Delta(\widetilde{k})\big) \\
&\overset{(b)}{\le} 8\epsilon_0 + \tau(\epsilon_0)\big(\Delta(k) + \Delta(\widetilde{k}) + \Delta(k)\Delta(\widetilde{k})\big) \overset{(c)}{\le} 8\epsilon_0 + 4\epsilon_0 e^{\beta}\big(2\eta + \eta^2\big) \le \eta.
\end{aligned}
$$

Inequality (a) follows from (6.2). (b) uses Equation (6.1) and $|\mu_f| \le \tau(\epsilon_0) \le 4\epsilon_0 e^{\beta}$. We showed that $\Delta(1) = 0$. In (c) we use the inductive assumption $\Delta(k) \le \eta$ for all $k < d$ and the assumption ϵ_0 defined in (5.6). Since w and $\widetilde{w}$ were arbitrary, this proves $\Delta(d) \le \eta$, and moreover, this holds for all d.

6.3. *Necessary samples for accurate pairwise marginals* (*proof of Theorem* 3.4). We construct a family of trees that are difficult to distinguish from one another. Applying the version of Fano's inequality below in Lemma 6.2, gives a lower bound on the error probability. The bound on the sample complexity is in terms of the KL-divergence between pairs of points in the parameter space. The symmetrized KL-divergence between two zero-field Ising models with parameters θ and θ' has the convenient form

$$
J\big(\theta \parallel \theta'\big) \triangleq D\big(\theta \parallel \theta'\big) + D\big(\theta' \parallel \theta\big) = \sum_{i<j}(\theta_{ij} - \theta'_{ij})(\mu_{ij} - \mu'_{ij}). \tag{6.3}
$$

Here, μ_{ij} and μ'_{ij} are the pairwise correlations between nodes i and j computed according to θ and θ', respectively.

LEMMA 6.2 (Fano's inequality, Corollary 2.6 in [42]). *Assume that $M \ge 2$ and that Θ is a family of models $\theta^0, \theta^1, \ldots, \theta^M$. Let Q_{θ^j} denote the probability law of the observation X under model θ^j. Let $\Phi : \{-1, +1\}^{p\times n} \to \{0, 1, \ldots, M\}$ denote an estimator using n i.i.d. samples $X^{(1:n)}$. If*

$$
n < (1-\delta)\frac{\log M}{\frac{1}{M+1}\sum_{j=1}^{M} J(Q_{\theta^j} \parallel Q_{\theta^0})}, \tag{6.4}
$$

then the probability of error of any algorithm is bounded as

$$
\inf_{\Phi} \max_{0\le j\le M} Q_{\theta^j}\big[\Phi\big(X^{(1:n)}\big) \ne j\big] \ge \delta - \frac{1}{\log M}.
$$

The following corollary is a restatement of equation (2.9) in [42] using the above lemma and tailored to our setup.

COROLLARY 6.3. *Let d be a distance* (*i.e., a metric*). *Suppose there are M different parameter vectors $\theta^0, \ldots, \theta^M$ such that $d(\theta^k, \theta^j) \ge 2\eta$ for all $j \ne k$. If n satisfies* (6.4), *then any estimator $\Psi : X^{(1:n)} \to \Theta$ mapping n samples to a set of parameters associated with a tree T and a distribution on $P \in \mathcal{P}_{\mathsf{T}}$ incurs a loss greater than η with probability at least* $1/2$:

$$
\inf_{\Psi} \sup_{\substack{\mathsf{T}\in\mathcal{T} \\ P\in\mathcal{P}_{\mathsf{T}}(\alpha,\beta)}} \mathsf{P}\big[d\big(\Psi\big(X^{(1:n)}\big), \theta\big) \ge \eta\big] \ge \frac{1}{2}.
$$

PROOF OF THEOREM 3.4. We consider a fixed tree structure given by a path, or in other words a Markov chain $X_1 - X_2 - \cdots - X_p$. We choose M different parameter vectors θ^m, $0 \le m \le M-1$, for $M = p$.

Let $\theta^0_{i,i+1} = \alpha$ for $i = 1, \ldots, p-1$. In the mth model, we have $\theta^m_{m,m+1} = \operatorname{atanh}(\tanh\alpha + 4\eta)$ and the remaining edge weights $\theta^m_{i,i+1} = \alpha$ for $i \neq m$. For $m' \neq m$,

$$\max_{i,j} \bigl|\mathbb{E}_{\theta^{m'}}[X_i X_j] - \mathbb{E}_{\theta^m}[X_i X_j]\bigr| \ge 4\eta.$$

Also, using (6.3)

$$J\bigl(\theta^m \parallel \theta^{m'}\bigr) \le 4\eta\bigl[\operatorname{atanh}(\tanh\alpha + 4\eta) - \alpha\bigr] \le 4\eta \frac{4\eta}{1 - [\tanh\alpha + 4\eta]^2},$$

where we used $\frac{d}{dx}\operatorname{atanh}(x) = \frac{1}{1-x^2}$ to get the last inequality. Fano's inequality (Corollary 6.3) with the distance $\mathcal{L}^{(2)}$ gives the bound. □

7. Sample complexities of structure learning and forest approximation.

7.1. *Samples necessary for structure learning.*

PROOF OF THEOREM 3.1. Suppose that p is odd (for simplicity) and let the graph T_0 be a path with associated parameters θ^0 given by $\theta^0_{i,i+1} = \alpha$ for odd values of i and $\theta^0_{i,i+1} = \beta$ for even values of i. For each odd value of $m \le p-2$, we let θ^m be equal to θ^0 everywhere except $\theta^m_{m,m+1} = 0$ and $\theta^m_{m,m+2} = \alpha$. There are $(p+1)/2$ models in total (including θ^0). A small calculation using (6.3) leads to

$$J\bigl(\theta^m \parallel \theta^0\bigr) = \alpha\tanh\alpha[1 - \tanh\beta] \le 2\alpha^2 e^{-2\beta}.$$

Here, we used $\tanh\alpha \le \alpha$ and $1 - \tanh\beta \le 2e^{-2\beta}$ for $\alpha, \beta \ge 0$. Plugging the last display into Fano's inequality (Lemma 6.2) completes the proof. □

7.2. *Samples sufficient for structure learning (proof of Theorem 3.2).*

Consider the original tree T with parameters $\alpha \le |\theta_{ij}| \le \beta$ for $(i,j) \in \mathcal{E}_{\mathsf{T}}$. Using Definition 5.4, the Chow–Liu algorithm recovers strong edges on the event $\mathbb{E}^{\text{strong}}(\epsilon)$, where edge (i,j) is strong if its parameter θ_{ij} satisfies $|\tanh\theta_{ij} > \tau$. Thus, if the edge strength lower bound α in the original tree T satisfies $\tanh\alpha > \tau$, on event $\mathbb{E}^{\text{strong}}(\epsilon)$ we have $\mathsf{T}^{\mathsf{CL}} = \mathsf{T}$. Note that by Lemma 8.5, the event $\mathbb{E}^{\text{strong}}(\epsilon)$ (defined in (5.4)) with $\epsilon = \sqrt{2/n\log(2p^2/\delta)}$ occurs with probability at least $1-\delta$. The bound

$$n > \frac{16}{\tanh^2(\alpha)(1 - \tanh\beta)} \log\frac{2p^2}{\delta}$$

on the number of samples guarantees $\tanh\alpha > \tau$ and $\mathsf{T}^{\mathsf{CL}} = \mathsf{T}$ with probability at least $1-\delta$. Using $1 - \tanh\beta \ge e^{-2\beta}$ gives the statement of Theorem 3.2.

7.3. *Lower bounding sample complexity of forest approximation algorithm, proof of Proposition 4.2.*

Since the loss function $\widetilde{d}(P, Q)$ defined in (4.1) is not symmetric and hence not a distance, one cannot use Corollary 6.3 to lower bound the sample complexity of the forest approximation algorithm.

Define a class of $M = p-1$ models Θ as follows: Let θ^1 be a model in which $\theta^1_{12} = 0$, $\theta^1_{2i} = \beta$ for $3 \le i \le p$, and $\theta^1_{ij} = 0$ for all other edges so that $\mu_{1i} = 0$ for all $i \ge 2$ in this model. For $2 \le m \le M$, define θ^m (mth model) such that $\theta^m_{12} = 0$, $\theta^m_{2i} = \beta$ for $3 \le i \le p$ and

$\theta^m_{1,m+1} = \operatorname{atanh}(\eta)$. Using (6.3), since $\mu_{1,m+1} = \eta$ and $\mu_{1i} = \eta \tanh^2(\beta)$ for $i \geq 3$ in model $m \geq 2$,

$$\begin{aligned}\frac{1}{M}\sum_{m=1}^{M} J\big(\theta^m \parallel \theta^M\big) &\leq \eta \operatorname{atanh}(\eta)\left[\frac{1}{M} + \left(1 - \frac{1}{M}\right)8e^{-2\beta}\right] \\ &\leq 16\eta \operatorname{atanh}(\eta) \max\left\{\frac{1}{p}, e^{-2\beta}\right\}.\end{aligned} \tag{7.1}$$

It will be useful to decompose an estimator into an estimator for the neighborhood of node 1, and an estimator for the rest of the model with neighborhood of node 1 fixed. Let $\Phi : \{-1, +1\}^{p\times n} \to \{0, 1\}^{\mathcal{V}\setminus\{1\}}$ be an estimator for the neighborhood of node 1, and let Ψ'_Φ be an estimator that generates a forest F, constrained to have the same neighborhood for node 1 that the estimator Φ recovers, and a distribution $Q \in \mathcal{P}_{\mathsf{F}}(0, \beta)$. Any estimator Ψ can be decomposed in this way to Φ and Ψ'_Φ. It follows that

$$\begin{aligned}\inf_{\Psi} \sup_{\substack{\mathsf{T}\in\mathcal{T} \\ P\in\mathcal{P}_{\mathsf{T}}(\alpha,\beta)}} \mathsf{P}\big[\widetilde{d}\big(P, \Psi\big(X^{(1:n)}\big)\big) \geq \eta\big] &\overset{(a)}{\geq} \inf_{\Phi}\inf_{\Psi'_\Phi}\max_m \mathsf{P}_{\theta^m}\big[\widetilde{d}\big(P, \Psi\big(X^{(1:n)}\big)\big) \geq \eta\big] \\ &\overset{(b)}{\geq} \inf_{\Phi}\max_m\inf_{\Psi'_\Phi} \mathsf{P}_{\theta^m}\big[\widetilde{d}\big(P, \Psi\big(X^{(1:n)}\big)\big) \geq \eta\big] \\ &\overset{(c)}{\geq} \inf_{\Phi}\max_m \mathsf{P}_{\theta^m}\big[\Phi\big(X^{(1:n)}\big) \neq \partial_{\theta^m}(1)\big].\end{aligned}$$

(a) holds since $\{P_{\theta^m}\}_{m=1}^M \subseteq \bigcup_{\mathsf{T}\in\mathcal{T}} \mathcal{P}_{\mathsf{T}}(0, \beta)$. (b) holds since we swapped the order of infimum and max operations. (c) is justified as follows: If the data is generated from P_θ, $\theta \in \Theta$, incorrect recovery of the neighborhood of node 1 by the estimator implies that $\widetilde{d} \geq \eta$. This is because $\widetilde{d} = 1$ if any extra edges are added and otherwise, if there are no extra edges and an edge incident to node 1 is missing, then $\widetilde{d} \geq \eta$ due to the true correlation η on the edge being zero in the estimated model.

Now, the final quantity in the last display can be lower bounded by a standard application of Fano's inequality (Corollary 6.3) on the family Θ introduced in the beginning of the proof. The main ingredient is the average symmetric KL from (7.1) and specifying the distance. Let $\partial_{\mathsf{F}}(1)$ be the neighborhood of node 1 in forest F. For P defined on forest F and Q defined on F', we use the distance $\widetilde{d}'(P, Q) = \mathbb{1}[\partial_{\mathsf{F}}(1) \neq \partial_{\mathsf{F}'}(1)]$ (zero–one loss on neighborhood of node 1).

8. Control of events $\mathbf{E}^{\mathbf{corr}}$, $\mathbf{E}^{\mathbf{strong}}$ and $\mathbf{E}^{\mathbf{cascade}}$. We state a standard form of Hoeffding's inequality [22] in Appendix D and use it here.

LEMMA 8.1. *The event* $\mathrm{E}^{\mathrm{corr}}(\epsilon)$ *defined in* (5.2) *occurs with probability at least* $1 - 2p^2\exp(-n\epsilon^2/2)$.

PROOF. For a given pair of nodes w, $\widetilde{w}$, let $Z^{(i)} = X_w^{(i)} X_{\widetilde{w}}^{(i)}$ and apply Hoeffding's inequality (Lemma 5 in Appendix D) to get $\mathsf{P}[|\mu_{w,\widetilde{w}} - \widehat{\mu}_{w,\widetilde{w}}| > \epsilon] \leq 2\exp(-n\epsilon^2/2)$. Applying the union bound over $\binom{p}{2}$ pairs $w, \widetilde{w} \in \mathcal{V}$ of nodes completes the proof. □

We next prove Lemma 6.1 for completeness.

PROOF OF LEMMA 6.1. For edge $(u, \widetilde{u}) \notin \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$, if there is an edge $(i, j) \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u, \widetilde{u})$ such that $|\widehat{\mu}_{u\widetilde{u}}| > |\widehat{\mu}_{ij}|$, then T^{CL} cannot be the maximum weight spanning tree. To show

that, consider the tree T' identical to T^{CL} except $(u,\widetilde{u}) \in \mathcal{E}_{\mathsf{T}'}$ and $(i,j) \notin \mathcal{E}_{\mathsf{T}'}$ (i.e., $\mathcal{E}_{\mathsf{T}'} = (\mathcal{E}_{\mathsf{T}^{\mathsf{CL}}} \setminus \{(i,j)\}) \cup \{(u,\widetilde{u})\}$.) Note that T' is a spanning tree and observe that $\text{weight}(\mathsf{T}') \triangleq \sum_{e\in\mathcal{E}_{\mathsf{T}'}} |\widehat{\mu}_e| = \sum_{e\in\mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}} |\widehat{\mu}_e| + |\widehat{\mu}_{u\widetilde{u}}| - |\widehat{\mu}_{ij}| > \text{weight}(\mathsf{T}^{\mathsf{CL}})$. □

We define a pair of random variables that will help to characterize the mistakes made by the Chow–Liu algorithm. For a given pair of nodes $v, \widetilde{v}$ and edge $f = (u,\widetilde{u}) \in \mathsf{path}_{\mathsf{T}}(v,\widetilde{v})$, let

$$Z_{f,v,\widetilde{v}} = X_u X_{\widetilde{u}} - X_v X_{\widetilde{v}} = X_u X_{\widetilde{u}}(1 - X_u X_v X_{\widetilde{v}} X_{\widetilde{u}}), \tag{8.1}$$

$$Y_{f,v,\widetilde{v}} = X_u X_{\widetilde{u}} + X_v X_{\widetilde{v}} = X_u X_{\widetilde{u}}(1 + X_u X_v X_{\widetilde{v}} X_{\widetilde{u}}). \tag{8.2}$$

LEMMA 8.2. *If there exists a pair of edges $f = (u,\widetilde{u})$ and $g = (v,\widetilde{v})$ such that $f \in \mathcal{E}_{\mathsf{T}} \setminus \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$, $g \in \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}} \setminus \mathcal{E}_{\mathsf{T}}$ and additionally $f \in \mathsf{path}_{\mathsf{T}}(v,\widetilde{v})$ and $g \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u,\widetilde{u})$, then*

$$\left(\sum_{i=1}^n Z_{f,v,\widetilde{v}}^{(i)}\right)\left(\sum_{i=1}^n Y_{f,v,\widetilde{v}}^{(i)}\right) \le 0.$$

PROOF. Using Lemma 6.1, $f = (u,\widetilde{u}) \notin \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$ and $g = (v,\widetilde{v}) \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u,\widetilde{u})$ implies that $|\widehat{\mu}_g| \ge |\widehat{\mu}_f|$. Hence $\widehat{\mu}_g^2 \ge \widehat{\mu}_f^2$ and

$$\begin{aligned}
0 &\ge \widehat{\mu}_f^2 - \widehat{\mu}_g^2 = (\widehat{\mu}_f - \widehat{\mu}_g)(\widehat{\mu}_f + \widehat{\mu}_g) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^n X_u^{(i)} X_{\widetilde{u}}^{(i)} - X_v^{(i)} X_{\widetilde{v}}^{(i)}\right)\left(\sum_{i=1}^n X_u^{(i)} X_{\widetilde{u}}^{(i)} + X_v^{(i)} X_{\widetilde{v}}^{(i)}\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^n Z_{f,v,\widetilde{v}}^{(i)}\right)\left(\sum_{i=1}^n Y_{f,v,\widetilde{v}}^{(i)}\right),
\end{aligned}$$

where in the last step we used $f \in \mathsf{path}_{\mathsf{T}}(v,\widetilde{v})$ and the definition of $Z_{f,w,\widetilde{w}}$ and $Y_{f,w,\widetilde{w}}$ in (8.1) and (8.2). □

Later we will bound the probability of the event in Lemma 8.2. To this end, we will derive deviation bounds on $Z_{f,v,\widetilde{v}}$ and $Y_{f,v,\widetilde{v}}$ in Lemma 8.3. We use the standard Bernstein's inequality as quoted from [42] in Appendix D.

LEMMA 8.3. *For all pairs of nodes $v,\widetilde{v} \in \mathcal{V}$ and edges $f = (u,\widetilde{u}) \in \mathsf{path}_{\mathsf{T}}(v,\widetilde{v})$, let $\mathcal{A}_{f,v,\widetilde{v}} = \mathsf{path}_{\mathsf{T}}(v,\widetilde{v}) \setminus \{f\}$ such that $\mu_{v\widetilde{v}} = \mu_f \mu_{\mathcal{A}_{f,v,\widetilde{v}}}$. Given n i.i.d. samples let $Z_{f,v,\widetilde{v}}^{(1)}, \ldots, Z_{f,v,\widetilde{v}}^{(n)}$ be defined in (8.1), and $Y_{f,v,\widetilde{v}}^{(1)}, \ldots, Y_{f,v,\widetilde{v}}^{(n)}$ be defined in (8.2). Let $\epsilon = \sqrt{2/n \log(2p^2/\delta)}$. Then, with probability at least $1-\delta$*

$$\left|\frac{1}{n}\sum_{i=1}^n Z_{f,v,\widetilde{v}}^{(i)} - \mu_f(1 - \mu_{\mathcal{A}_{f,v,\widetilde{v}}})\right| \le \max\{4\epsilon^2, 4\epsilon\sqrt{1 - \mu_{\mathcal{A}_{f,v,\widetilde{v}}}}\} \quad \textit{and} \tag{8.3}$$

$$\left|\frac{1}{n}\sum_{i=1}^n Y_{f,v,\widetilde{v}}^{(i)} - \mu_f(1 + \mu_{\mathcal{A}_{f,v,\widetilde{v}}})\right| \le \max\{4\epsilon^2, 4\epsilon\sqrt{1 + \mu_{\mathcal{A}_{f,v,\widetilde{v}}}}\}. \tag{8.4}$$

PROOF. We prove that (8.3) holds with probability at least $1 - \delta/2$. The proof of (8.4) is analogous.

We use the abbreviation $\mathcal{A}$ instead of $\mathcal{A}_{f,v,\widetilde{v}}$ and Z instead of $Z_{f,v,\widetilde{v}}$ in this proof. Applying Lemma 8.6, it follows from the fact that P is Markov with respect to T and

$f = (u, \widetilde{u}) \in \mathsf{path}_{\mathsf{T}}(v, \widetilde{v})$ that $X_u X_{\widetilde{u}}$ and $X_u X_v X_{\widetilde{u}} X_{\widetilde{v}}$ are independent random variables. Note that $P(X_u X_{\widetilde{u}} = 1) = (1 + \mu_f)/2$ and $P(X_u X_{\widetilde{u}} = -1) = (1 - \mu_f)/2$. Similarly, the distribution of $X_u X_v X_{\widetilde{u}} X_{\widetilde{v}}$ is a function of $\mu_{\mathcal{A}}$. As a result, the random variable $Z_{f,v,\widetilde{v}} \in \{-2, 0, 2\}$ defined in (8.1) has the following distribution:

$$Z = \begin{cases} -2 & \text{w.p. } \dfrac{1-\mu_f}{2}\dfrac{1-\mu_{\mathcal{A}}}{2}, \\ 0 & \text{w.p. } \dfrac{1+\mu_{\mathcal{A}}}{2}, \\ +2 & \text{w.p. } \dfrac{1+\mu_f}{2}\dfrac{1-\mu_{\mathcal{A}}}{2}. \end{cases}$$

The first and second moments of Z are $\mathbb{E}[Z] = \mu_f(1 - \mu_{\mathcal{A}})$ and $\mathrm{Var}[Z] = (1 - \mu_{\mathcal{A}})[2 - \mu_f^2(1 - \mu_{\mathcal{A}})] \le 2(1 - \mu_{\mathcal{A}})$. By Bernstein's inequality (Lemma 6 in Appendix D), with probability at least $1 - \delta/2$,

$$\left|\sum_{i=1}^{n} Z^{(i)} - n\mathbb{E}[Z]\right| \le n \max\left\{\frac{8}{3n}\log\frac{4}{\delta}, \sqrt{\frac{4\,\mathrm{Var}[Z_{f,v,\widetilde{v}}]}{n}\log\frac{4}{\delta}}\right\}.$$

Using a union bound, we show that for any pair of nodes v, $\widetilde{v}$ and any edge $f = (u, \widetilde{u}) \in \mathsf{path}_{\mathsf{T}}(v, \widetilde{v})$,

$$\left|\sum_{i=1}^{n} Z^{(i)} - n\mu_f(1 - \mu_{\mathcal{A}})\right| \le n \max\left\{\frac{8}{3n}\log\frac{4p^3}{\delta}, \sqrt{\frac{8(1-\mu_{\mathcal{A}})}{n}\log\frac{4p^3}{\delta}}\right\}.$$

The definition of ϵ gives $\frac{8}{3n}\log\frac{4p^3}{\delta} \le 4\epsilon^2$ and $\sqrt{\frac{8(1-\mu_{\mathcal{A}})}{n}\log\frac{4p^3}{\delta}} \le 4\epsilon\sqrt{1-\mu_{\mathcal{A}}}$ which gives the lemma. □

Event $\mathrm{E}^{\mathrm{strong}}(\epsilon)$ in (5.4) occurs if all of the strong edges in T (defined in (5.3)) are recovered in T^{CL}. Lemma 8.4 shows that the deviation bounds for the variables $Z_{f,v,\widetilde{v}}$ and $Y_{f,v,\widetilde{v}}$ stated in (8.3) and (8.4) imply $\mathrm{E}^{\mathrm{strong}}(\epsilon)$.

LEMMA 8.4. *Under the events described in Lemma 8.3, if there is an edge $f \in \mathcal{E}_{\mathsf{T}}$ missing from the Chow–Liu tree, $f \notin \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$, then $|\mu_f| \le \tau(\epsilon) = \frac{4\epsilon}{\sqrt{1-\tanh\beta}}$ (i.e., $\mathrm{E}^{\mathrm{strong}}(\epsilon)$ defined in equation (5.4) holds).*

PROOF. Applying Lemma 8.8 to $f = (u, \widetilde{u})$ shows that for the edge $f \in \mathcal{E}_{\mathsf{T}} \setminus \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}}$, there exists an edge $g = (v, \widetilde{v}) \in \mathcal{E}_{\mathsf{T}^{\mathsf{CL}}} \setminus \mathcal{E}_{\mathsf{T}}$ such that, $f \in \mathsf{path}_{\mathsf{T}}(v, \widetilde{v})$ and $g \in \mathsf{path}_{\mathsf{T}^{\mathsf{CL}}}(u, \widetilde{u})$ (Figure 2). Let $Z = Z_{f,v,\widetilde{v}}$ defined in 8.1 and $Y = Y_{f,v,\widetilde{v}}$ defined in 8.2 in the scope of this proof. Applying Lemma 8.2, this implies that $(\sum_{i=1}^n Z^{(i)})(\sum_{i=1}^n Y^{(i)}) \le 0$. Note that $\mathbb{E}Z = \mu_f(1 - \mu_{\mathcal{A}})$ and $\mathbb{E}Y = \mu_f(1 + \mu_{\mathcal{A}})$ using the definition $\mathcal{A} = \mathcal{A}_{f,v,\widetilde{v}} = \mathsf{path}_{\mathsf{T}}(v, \widetilde{v}) \setminus \{f\}$. Hence $(\mathbb{E}Z)(\mathbb{E}Y) = \mu_f^2(1 - \mu_{\mathcal{A}}^2) \ge 0$. Thus, $(\sum_{i=1}^n Z^{(i)})(\sum_{i=1}^n Y^{(i)}) < 0$ holds only if either one of the following inequalities holds:

$$\left|\sum_{i=1}^{n} Z^{(i)} - n\mathbb{E}Z\right| \ge n|\mathbb{E}Z| \quad \text{or} \quad \left|\sum_{i=1}^{n} Y^{(i)} - n\mathbb{E}Y\right| \ge n|\mathbb{E}Y|.$$

On the events described in Lemma 8.3, there is an upper bound on $|\sum_{i=1}^n Z^{(i)} - n\mathbb{E}Z|$ and $|\sum_{i=1}^n Y^{(i)} - n\mathbb{E}Y|$. Hence, on these events, the property in above display holds only if either one of these inequalities holds:

$$|\mu_f(1 - \mu_{\mathcal{A}})| \le \max\{4\epsilon^2, 4\epsilon\sqrt{1 - \mu_{\mathcal{A}}}\} \quad \text{or}$$
$$|\mu_f(1 + \mu_{\mathcal{A}})| \le \max\{4\epsilon^2, 4\epsilon\sqrt{1 + \mu_{\mathcal{A}}}\},$$

which is true if

$$|\mu_f| \le \max\left\{\frac{4\epsilon}{\sqrt{1-\mu_{\mathcal{A}}}}, \frac{4\epsilon^2}{1-\mu_{\mathcal{A}}}, \frac{4\epsilon}{\sqrt{1+\mu_{\mathcal{A}}}}, \frac{4\epsilon^2}{1+\mu_{\mathcal{A}}}\right\} \le \max\{\tau(\epsilon), \tau^2(\epsilon)\},$$

where $\tau(\epsilon)$ is defined in (5.3). Note that if $\tau(\epsilon) \ge 1$ then the bound on $|\mu_f| \le 1 \le \tau(\epsilon)$ is trivial. If $\tau(\epsilon) < 1$, then $\tau^2(\epsilon) < \tau(\epsilon)$ which gives $|\mu_f| \le \tau(\epsilon)$. □

LEMMA 8.5. *With* $\epsilon = \sqrt{2/n \log(2p^2/\delta)}$, *event* $\mathrm{E}^{\mathrm{strong}}(\epsilon)$ *defined in* (5.4) *occurs with probability at least* $1-\delta$.

PROOF. Lemma 8.4 shows that, under the events described in Lemma 8.3, the event $\mathrm{E}^{\mathrm{strong}}(\epsilon)$ defined in equation (5.4) holds. Using Lemma 8.3, with probability at least $1-\delta$, all edges $e \in \mathsf{T}$ such that $|\mu_e| > \tau(\epsilon)$ are recovered by the Chow–Liu algorithm $e \in \mathsf{T}^{\mathsf{CL}}$. □

LEMMA 8.6. *Let the distribution* $P(x) \in \mathcal{P}(\mathsf{T})$ *be a zero-field Ising model on the tree* $\mathsf{T} = (\mathcal{V}, \mathcal{E})$. *For all* $e = (i, j) \in \mathcal{E}$, *let* $Y_e = X_i X_j$. *Then the random variables* $\{Y_e\}_{e \in \mathsf{T}}$ *are jointly independent*.

This follows from the factorization of distribution $P(x) \in \mathcal{P}(\mathsf{T})$ in (1.1).

Next, we prove an upper bound on the end-to-end error on paths in the tree T. Interestingly, the bound is dimension-free: the error is independent of the length of path. Appendix E contains the proof of Lemma 8.7.

LEMMA 8.7. *Suppose* $\gamma < 1$. *If* $n > \max\{25/\gamma^2 \log(4p^2/\delta), 108e^{2\beta} \log(2p^3/\delta)\}$, *then the event* $\mathrm{E}^{\mathrm{cascade}}(\gamma)$ *defined in* (5.5) *occurs with probability at least* $1-\delta$.

LEMMA 8.8. *Let* T_1 *and* T_2 *be two spanning trees on a set of nodes* $\mathcal{V}$. *Let* $w, \widetilde{w}$ *be a pair of nodes such that* $\mathsf{path}_{\mathsf{T}_1}(w, \widetilde{w}) \ne \mathsf{path}_{\mathsf{T}_2}(w, \widetilde{w})$. *Then there exists a pair of edges* $f \triangleq (u, \widetilde{u}) \in \mathsf{path}_{\mathsf{T}_1}(w, \widetilde{w})$ *and* $g \triangleq (v, \widetilde{v}) \in \mathsf{path}_{\mathsf{T}_2}(w, \widetilde{w})$ *such that*:

(i) $f \notin \mathsf{path}_{\mathsf{T}_2}(w, \widetilde{w})$ *and* $g \notin \mathsf{path}_{\mathsf{T}_1}(w, \widetilde{w})$
(ii) $f \in \mathsf{path}_{\mathsf{T}_1}(v, \widetilde{v})$ *and* $g \in \mathsf{path}_{\mathsf{T}_2}(u, \widetilde{u})$.

Since $f \in \mathsf{path}_{\mathsf{T}_1}(w, \widetilde{w}) \cap \mathsf{path}_{\mathsf{T}_1}(v, \widetilde{v})$, w and $\widetilde{w}$ (and, resp., v and $\widetilde{v}$) are in different subtrees of T_1 after removing edge f, one can label the end points of the edges $f = (u, \widetilde{u})$ and $g = (v, \widetilde{v})$ such that $u, v \in \mathsf{SubTree}_{\mathsf{T}_1, f}(w)$ and $\widetilde{u}, \widetilde{v} \in \mathsf{SubTree}_{\mathsf{T}_1, f}(\widetilde{w})$ (Figure 2). Lemma 8.8 is proved in Appendix F.

9. Numerical simulations. We use numerical simulations to demonstrate the performance of the Chow–Liu algorithm in terms of both the probability of incorrect recovery of underlying structure (zero–one loss defined in (3.1)) and the $\mathcal{L}^{(2)}$ loss defined in (3.3). We are specifically interested in the regime in which the number of samples is not large enough to guarantee the correct recovery of the underlying tree.

In these simulations, the generative probability distributions of the samples are factorized according to (1.1) for a randomly chosen tree uniform over the set of trees on p nodes. To observe the effect of upper and lower bounds on the edge parameters, for each edge $(i, j) \in \mathcal{E}$, the edge parameter θ_{ij} takes one of the values α or β with equal probability.

In Figure 3(a), we plot $\mathsf{P}[\mathsf{T}^{\mathsf{CL}} \ne \mathsf{T}]$ as a function of the number of samples with $p = 31$, $\beta = 2$ and different values of α. One can observe that the probability of error is higher for smaller values of α (it increases to one as α decays to zero for any value of n). Figure 3(b)

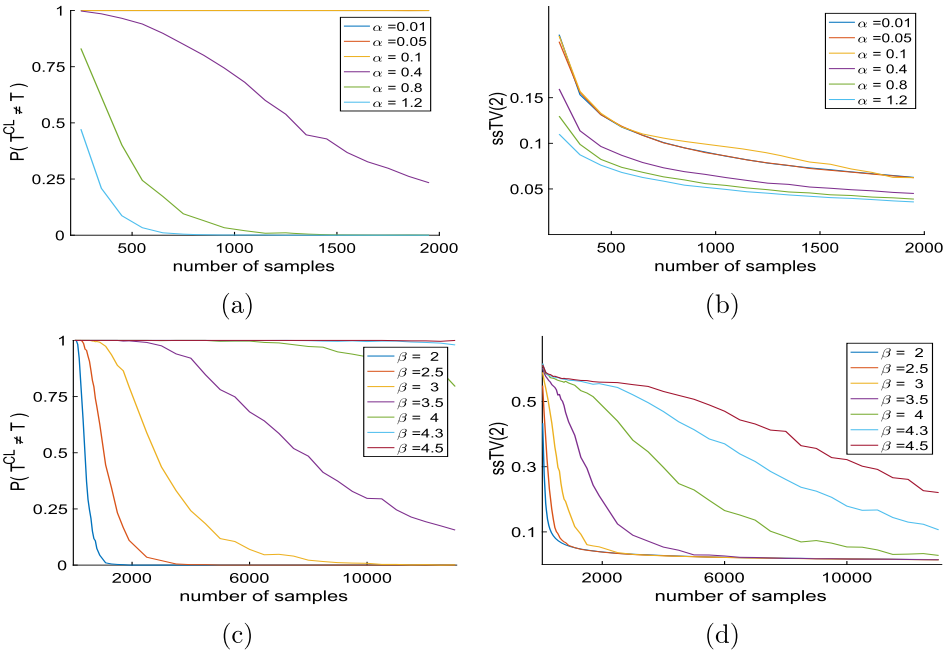

FIG. 3. *The performance of the Chow–Liu algorithm as a function of the number of samples for generative distribution factorized according to* (1.1) *with* $p = 31$. *Figures* (*a*) *and* (*b*) *use* $\beta = 2$ *and different values of* α. *Figures* (*c*) *and* (*d*) *use* $\alpha = .8$ *and different values of* β. *Figures* (*a*) *and* (*c*) *depict the probability of error and Figures* (*b*) *and* (*d*) *show the small set TV* $\mathcal{L}^2$ *as the performance metric*.

illustrates the $\mathcal{L}^{(2)}$ loss in the same setup. This plot shows that as α decays, the loss remains bounded (although not necessarily monotonic in α), consistent with Theorem 3.3.

Figures 3(c) and 3(d) plot the probability of error and the $\mathcal{L}^{(2)}$ loss as a function of n with $p = 31$, $\alpha = 0.8$, and different values of β. As β increases, Figure 3(c) shows that the probability of error in learning the tree increases (for any n, the probability of error goes to one as β grows large enough). Figure 3(d) is consistent with the statement of Theorem 3.3, which states that expected $\mathcal{L}^{(2)}$ loss decays as $C' p \exp(-Cne^{-2\beta}) + C''\sqrt{\log p/n}$.

Appendix I contains numerical simulations on implementation of forest approximation algorithm and its comparison with the Chow–Liu algorithm. It also contains the simulation results depicting the performance of Chow–Liu algorithm with misspecified models. A tree-structured Ising model is changed isotropically on the space of distributions by a small offset to construct the generative distribution. This suggests that the the output of the Chow–Liu algorithm is robust to misspecification in the model and close to the generative distribution with respect to $\mathcal{L}^{(2)}$ loss. The performance of the Chow–Liu algorithm in term of $\mathcal{L}^{(k)}$ loss for general value of k and for generative tree-structured Ising models in presence of external field are also studied in Appendix I.

Discussion. In this paper, we prove guarantees on accuracy of prediction for a learned tree-structured Ising model. There is a large literature on learning tree-structured Markov random fields, and it is useful to carefully compare the guarantees obtained by each when applied to our setting. In the Supplementary Material, we review different approaches that could be taken toward learning a tree-structured distribution. We also review some known algorithms and their sample complexity.

There are many interesting questions remaining, including those mentioned in the Introduction: model misspecification, how to close the gap between the upper and lower bound on sample complexity, Ising models with external field, and obtaining tight guarantees for

$\mathcal{L}^{(k)}$ (marginals of order k) for $k > 2$. Of course, it is also of great interest to go beyond tree models and study other classes of models.

Acknowledgments. We thank Jerry Li for pointing out an error in an early version of the manuscript and the anonymous reviewers for their feedback which greatly improved the quality of the paper. We also thank Gregory Wornell, Lizhong Zheng, Gabor Lugosi, Devavrat Shah, David Gamarnik, Elchanan Mossel and Andrea Montanari for stimulating discussions.

The list of authors is in alphabetical order.

SUPPLEMENTARY MATERIAL

Supplement to "Learning a tree-structured Ising model in order to make predictions" (DOI: 10.1214/19-AOS1808SUPP; .pdf). Appendices contain additional proofs, numerical simulations and discussions on related work, available as a Supplementary Material [10].

REFERENCES

[1] ABBEEL, P., KOLLER, D. and NG, A. Y. (2006). Learning factor graphs in polynomial time and sample complexity. *J. Mach. Learn. Res.* **7** 1743–1788. MR2274423

[2] AGARWALA, R., BAFNA, V., FARACH, M., PATERSON, M. and THORUP, M. (1999). On the approximability of numerical taxonomy (fitting distances by tree metrics). *SIAM J. Comput.* **28** 1073–1085. MR1670078 https://doi.org/10.1137/S0097539795296334

[3] AMBAINIS, A., DESPER, R., FARACH, M. and KANNAN, S. (1997). Nearly tight bounds on the learnability of evolution. In *Proceedings of the* 38*th Annual Symposium on Foundations of Computer Science* (*FOCS*) 524–533. IEEE.

[4] ANANDKUMAR, A., HUANG, F., HSU, D. and KAKADE, S. (2012). Learning mixtures of tree graphical models. In *Advances in Neural Information Processing Systems* 1052–1060.

[5] ANANDKUMAR, A., TAN, V. Y. F., HUANG, F. and WILLSKY, A. S. (2012). High-dimensional structure estimation in Ising models: Local separation criterion. *Ann. Statist.* **40** 1346–1375. MR3015028 https://doi.org/10.1214/12-AOS1009

[6] ANANDKUMAR, A. and VALLUVAN, R. (2013). Learning loopy graphical models with latent variables: Efficient methods and guarantees. *Ann. Statist.* **41** 401–435. MR3099108 https://doi.org/10.1214/12-AOS1070

[7] BAXTER, R. J. (1985). Exactly solved models in statistical mechanics. In *Integrable Systems in Statistical Mechanics*. *Series on Advances in Statistical Mechanics* **1** 5–63. World Sci. Publishing, Singapore. MR0826537

[8] BENTO, J. and MONTANARI, A. (2009). Which graphical models are difficult to learn? In *Advances in Neural Information Processing Systems*.

[9] BRESLER, G. (2015). Efficiently learning Ising models on arbitrary graphs [extended abstract]. In *STOC'*15*—Proceedings of the* 2015 *ACM Symposium on Theory of Computing* 771–782. ACM, New York. MR3388257

[10] BRESLER, G. and KARZAND, M. (2020). Supplement to "Learning a tree-structured Ising model in order to make predictions." https://doi.org/10.1214/19-AOS1808SUPP.

[11] BRESLER, G., MOSSEL, E. and SLY, A. (2008). Reconstruction of Markov random fields from samples: Some observations and algorithms. In *Approximation, Randomization and Combinatorial Optimization*. *Lecture Notes in Computer Science* **5171** 343–356. Springer, Berlin. MR2538799 https://doi.org/10.1007/978-3-540-85363-3_28

[12] CHOW, C. and LIU, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inform. Theory* **14** 462–467.

[13] CHOW, C. and WAGNER, T. (1973). Consistency of an estimate of tree-dependent probability distributions. *IEEE Trans. Inform. Theory* **19** 369–371.

[14] CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L. and STEIN, C. (2009). *Introduction to Algorithms*. McGraw-Hill.

[15] DASGUPTA, S. (1999). Learning polytrees. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* 134–141.

[16] DASKALAKIS, C., MOSSEL, E. and ROCH, S. (2009). Phylogenies without branch bounds: Contracting the short, pruning the deep. In *Proceedings of the* 13*th Annual International Conference on Research in Computational Molecular Biology* 451–465. Springer.

[17] ERDŐS, P. L., STEEL, M. A., SZÉKELY, L. A. and WARNOW, T. J. (1999). A few logs suffice to build (almost) all trees: Part ii. *Theoret. Comput. Sci.* **221** 77–118. MR1700821 https://doi.org/10.1016/S0304-3975(99)00028-6

[18] FREEMAN, W. T., PASZTOR, E. C. and CARMICHAEL, O. T. (2000). Learning low-level vision. *Int. J. Comput. Vis.* **40** 25–47.

[19] GALLAGER, R. G. (1962). Low-density parity-check codes. *IRE Trans. Inf. Theory* **IT-8** 21–28. MR0136009 https://doi.org/10.1109/tit.1962.1057683

[20] GEORGII, H.-O. (2011). *Gibbs Measures and Phase Transitions*, 2nd ed. *De Gruyter Studies in Mathematics* **9**. de Gruyter, Berlin. MR2807681 https://doi.org/10.1515/9783110250329

[21] HEINEMANN, U. and GLOBERSON, A. (2014). Inferning with high girth graphical models. In *Proceedings of the* 31*st International Conference on Machine Learning* (*ICML*-14) 1260–1268.

[22] HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

[23] JOG, V. and LOH, P.-L. (2015). On model misspecification and KL separation for Gaussian graphical models. In *IEEE International Symposium on Information Theory* (*ISIT*) 1174–1178.

[24] KOLLER, D. and FRIEDMAN, N. (2009). *Probabilistic Graphical Models*: *Principles and Techniques*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR2778120

[25] LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Oxford University Press, New York. MR1419991

[26] LIU, H., XU, M., GU, H., GUPTA, A., LAFFERTY, J. and WASSERMAN, L. (2011). Forest density estimation. *J. Mach. Learn. Res.* **12** 907–951. MR2786914 https://doi.org/10.1016/j.micres.2009.11.010

[27] LOH, P.-L. and WAINWRIGHT, M. J. (2013). Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *Ann. Statist.* **41** 3022–3049. MR3161456 https://doi.org/10.1214/13-AOS1162

[28] MEILĂ, M. and JORDAN, M. I. (2001). Learning with mixtures of trees. *J. Mach. Learn. Res.* **1** 1–48. MR1882241 https://doi.org/10.1162/153244301753344605

[29] MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363 https://doi.org/10.1214/009053606000000281

[30] MOSSEL, E. Distorted metrics on trees and phylogenetic forests. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **4** 108–116. https://doi.org/10.1109/TCBB.2007.1010

[31] NARASIMHAN, M. and BILMES, J. (2004). PAC-learning bounded tree-width graphical models. In *Proceedings of the* 20*th Conference in Uncertainty in Artificial Intelligence* 410–417.

[32] PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*: *Networks of Plausible Inference*. *The Morgan Kaufmann Series in Representation and Reasoning*. Morgan Kaufmann, San Mateo, CA. MR0965765

[33] PORTILLA, J., STRELA, V., WAINWRIGHT, M. J. and SIMONCELLI, E. P. (2003). Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Process.* **12** 1338–1351. MR2026777 https://doi.org/10.1109/TIP.2003.818640

[34] RAVIKUMAR, P., WAINWRIGHT, M. J. and LAFFERTY, J. D. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319. MR2662343 https://doi.org/10.1214/09-AOS691

[35] REBESCHINI, P. and VAN HANDEL, R. (2015). Can local particle filters beat the curse of dimensionality? *Ann. Appl. Probab.* **25** 2809–2866. MR3375889 https://doi.org/10.1214/14-AAP1061

[36] ROMBERG, J. K., CHOI, H. and BARANIUK, R. G. (2001). Bayesian tree-structured image modeling using wavelet-domain hidden Markov models. *IEEE Trans. Image Process.* **10** 1056–1068.

[37] SANTHANAM, N. P. and WAINWRIGHT, M. J. (2012). Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Trans. Inform. Theory* **58** 4117–4134. MR2943079 https://doi.org/10.1109/TIT.2012.2191659

[38] SREBRO, N. (2001). Maximum likelihood bounded tree-width Markov networks. In *The Conference on Uncertainty in Artificial Intelligence* (*UAI*).

[39] TAN, V. Y. F., ANANDKUMAR, A., TONG, L. and WILLSKY, A. S. (2011). A large-deviation analysis of the maximum-likelihood learning of Markov tree structures. *IEEE Trans. Inform. Theory* **57** 1714–1735. MR2815845 https://doi.org/10.1109/TIT.2011.2104513

[40] TAN, V. Y. F., ANANDKUMAR, A. and WILLSKY, A. S. (2011). Learning high-dimensional Markov forest distributions: Analysis of error rates. *J. Mach. Learn. Res.* **12** 1617–1653. MR2813149

[41] TANDON, R., SHANMUGAM, K., RAVIKUMAR, P. K. and DIMAKIS, A. G. (2014). On the information theoretic limits of learning Ising models. In *Advances in Neural Information Processing Systems* 2303–2311.

[42] TSYBAKOV, A. B. (2004). *Introduction to Nonparametric Estimation*. Springer.

[43] VUFFRAY, M., MISRA, S., LOKHOV, A. and CHERTKOV, M. (2016). Interaction screening: Efficient and sample-optimal learning of Ising models. In *Advances in Neural Information Processing Systems* 2595–2603.

[44] WAINWRIGHT, M. and JORDAN, M. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.

[45] WAINWRIGHT, M. J. (2003). Tree-reweighted belief propagation algorithms and approximate ML estimation via pseudo-moment matching. In *AISTATS*.

[46] WAINWRIGHT, M. J. (2006). Estimating the "wrong" graphical model: Benefits in the computation-limited setting. *J. Mach. Learn. Res.* **7** 1829–1859. MR2274425

[47] WAINWRIGHT, M. J., SIMONCELLI, E. P. and WILLSKY, A. S. (2001). Random cascades on wavelet trees and their use in analyzing and modeling natural images. *Appl. Comput. Harmon. Anal.* **11** 89–123. MR1841335 https://doi.org/10.1006/acha.2000.0350

[48] WU, R., SRIKANT, R. and NI, J. (2013). Learning loosely connected Markov random fields. *Stoch. Syst.* **3** 362–404. MR3353207 https://doi.org/10.1214/12-SSY073