

MIT Open Access Articles

*Learning a tree-structured ising
model in order to make predictions*

The MIT Faculty has made this article openly available. **Please share**
how this access benefits you. Your story matters.

Citation: Bresler, Guy and Mina Karzand. "Learning a tree-structured ising model in order to make predictions." *Annals of Statistics*, 48, 2 (April 2020): 713-737 © 2020 The Author(s)

As Published: 10.1214/19-AOS1808

Publisher: Institute of Mathematical Statistics

Persistent URL: <https://hdl.handle.net/1721.1/129620>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike



Learning a Tree-Structured Ising Model in Order to Make Predictions

Guy Bresler* and Mina Karzand†

Abstract

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 non-asymptotic 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 [46, 25]. 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 e.g. [30, 11, 44, 1, 35, 29]. In contrast, we consider in this paper the problem of learning a good model *for the purpose of performing subsequent*

*Guy Bresler is with the Department of Electrical Engineering and Computer Science at MIT and a member of LIDS and IDSS. guy@mit.edu

†Mina Karzand is with the Department of Electrical Engineering and Computer Science at MIT and a member of LIDS and IDSS. mkarzand@mit.edu

‡The authors are ordered alphabetically.

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 “infering” (inference + learning) [22], 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.

Structure learning becomes statistically more challenging, meaning more data is required, when interactions between variables are very weak or very strong [38, 9, 42]. It is intuitively clear that very weak edges are difficult to detect, leading to non-identifiability 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

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

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 [36] and has appeared earlier in Dobrushin’s work on Gibbs measures [21]. 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 [26, 33, 46, 47] 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

$$P(x) = \exp \left(\sum_{(i,j) \in \mathcal{E}} \theta_{ij} x_i x_j - \Phi(\theta) \right), \quad (1)$$

where $\Phi(\theta)$ is the normalizing constant. We assume throughout that $\alpha \leq |\theta_{ij}| \leq \beta$ for some $\alpha, \beta \geq 0$ 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). The assumption of no external field ($\theta_i = 0$) implies uniform singleton marginals, *i.e.*, $P(x_i = +1) = 1/2$ 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) with $|\theta_{ij}| \leq \beta$, denote the Chow-Liu tree by T^{CL} . The Ising model Q on T^{CL} obtained by matching correlations on the edges satisfies $\mathcal{L}^{(2)}(P, Q) \leq \eta$ with probability at least $1 - \delta$. Conversely, if $\tanh \alpha + 2\eta \leq \min\{\tanh \beta, 1/2\}$ and $n \leq C' \eta^{-2} \log p$, then no algorithm can find a tree model Q such that $\mathcal{L}^{(2)}(P, Q) \leq \eta$ 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 $\tanh \alpha + 2\eta \leq \min\{\tanh \beta, 1/2\}$ 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) 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, *i.e.*, samples are not from a tree-structured Ising model; external field in the Ising model (1) generating the data; accurate marginals of size $k \geq 3$.

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 [37, 19, 48, 34], artificial intelligence [33], coding theory [20] 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 [38, 42, 9, 10, 49, 28]. Structure learning of trees has been studied by [12, 40]. Learning of generalizations of tree-structured models has been studied, including: forest approximations [41, 27], polytrees [16], bounded treewidth graphs [39, 32], loopy graphs with correlation decay [6, 5], and mixtures of trees [4, 29].

Loopy belief propagation yields accurate marginals in high girth (locally tree-like) graphs with correlation decay. This fact was used by Heinemann and Globerson [22] 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 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 [32] 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. [18] studied sample and time complexity of tree metric based algorithms to reconstruct phylogenetic trees. Daskalakis et al. [17] and Mossel [31] use distorted tree metrics to get approximations of phylogenetic trees when exact reconstruction of the tree is impossible. In [31] 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. [17] 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 \hat{T} with $\|\hat{T} - D\|_\infty \leq 3\epsilon$. They prove that finding a T with $\|T - D\|_\infty \leq 9/8\epsilon$ 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 [45] 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 re-weighted 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.

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) with the restriction $\alpha \leq |\theta_{ij}| \leq \beta$ 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 $\hat{P}(x) = \frac{1}{n} \sum_{l=1}^n \mathbf{1}_{\{X^{(l)}=x\}}$ and $\hat{\mu}_{ij} = \mathbb{E}_{\hat{P}} X_i X_j$ 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 [26]) is that $P(x)$ factorizes as

$$P(x) = \prod_{i \in \mathcal{V}} P(x_i) \prod_{(i,j) \in \mathcal{E}} \frac{P(x_i, x_j)}{P(x_i)P(x_j)}. \quad (2)$$

2.3 Information projection

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

$$\tilde{P}(x) = \arg \min_{\substack{Q \text{ is factorized as (2)} \\ \text{according to } \mathsf{T}}} D(P\|Q)$$

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 \tilde{P} is obtained by matching the first and second order marginals to those of P , i.e., for all $(i, j) \in \mathcal{E}$, and all $x_i, x_j \in \mathcal{X}$, $\tilde{P}(x_i, x_j) = P(x_i, x_j)$.

Let

$$\Pi_{\mathsf{T}}(P) = \arg \min_{Q \in \mathcal{P}_{\mathsf{T}}} D(P\|Q) \quad (3)$$

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}_{T} that $\tilde{P} = \Pi_{\mathsf{T}}(P)$ can be represented as Equation (1) for some $\tilde{\theta}$ supported on T . It is shown in Appendix A in the supplementary material that $\tilde{P} = \Pi_{\mathsf{T}}(P)$ has edge weights $\tilde{\theta}_{ij}$ for each $(i, j) \in \mathcal{E}_{\mathsf{T}}$ satisfying $\tanh \tilde{\theta}_{ij} = \mu_{ij} \triangleq \mathbb{E}_P X_i X_j$ (and $\tilde{\theta}_{ij} = 0$ 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 $P \in \mathcal{P}_T$, one observes n independent samples (configurations) $X^{(1)}, \dots, X^{(n)} \in \{-, +\}^p$ from the Ising model (1). In this context, a *structure learning algorithm* is a (possibly randomized) map $\phi : \{-1, +1\}^{p \times n} \rightarrow \mathcal{T}$ 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 A.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 $P \in \mathcal{P}_T$, we define the Chow-Liu tree to be the maximum likelihood tree:

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

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 $P \in \mathcal{P}_T$ can be described by (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 \hat{P} , $T^{\text{CL}} = \operatorname{argmin}_{T \in \mathcal{T}} \min_{P \in \mathcal{P}_T} D(\hat{P} \| P)$. It is shown in Lemma A.1 in Appendix A that

$$T^{\text{CL}} = \operatorname{argmax}_{\{\text{spanning trees } T'\}} \sum_{e \in \mathcal{E}_{T'}} |\hat{\mu}_e|, \quad (4)$$

where for $e = (i, j)$, $\hat{\mu}_e = \mathbb{E}_{\hat{P}} X_i X_j$ 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, i.e., 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. [40, 41]. A variety of other results and generalizations have appeared, including for example Liu et al.'s work on forest estimation with non-parametric potentials [27] (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,

$$\mathcal{L}^{0-1}(T, T') = \mathbf{1}_{\{T \neq T'\}}, \quad (5)$$

meaning that the exact underlying graph must be learned (see e.g., [11, 38, 41, 27]). The risk, or expected loss, of algorithm ϕ under some distribution $P \in \mathcal{P}_{\mathbf{T}}(\alpha, \beta)$ is then given by the probability of reconstruction error, $\mathbb{E}_P \mathcal{L}^{0-1}(\mathbf{T}, \phi(X^{(1:n)})) = \mathbb{P}(\phi(X^{(1:n)}) \neq \mathbf{T})$, and the maximum risk is $\sup\{\mathbb{P}(\phi(X^{(1:n)}) \neq \mathbf{T}) : \mathbf{T} \in \mathcal{T}, P \in \mathcal{P}_{\mathbf{T}}(\alpha, \beta)\}$ for given α, β, p and n .

The sample complexity of learning the correct tree underlying the distribution increases as edges become weaker, i.e., as $\alpha \rightarrow 0$, 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 [38, 42] 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 $n < \frac{1}{16}e^{2\beta}/(\alpha \tanh \alpha) \log p$ samples, the worst-case probability of error over trees $\mathbf{T} \in \mathcal{T}$ and distributions $P \in \mathcal{P}_{\mathbf{T}}(\alpha, \beta)$ is at least half for any algorithm, i.e.,*

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

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 \mathbf{T} and Ising model $P \in \mathcal{P}_{\mathbf{T}}(\alpha, \beta)$. 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, i.e., $\mathbf{T}^{\text{CL}} = \mathbf{T}$.*

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 9.1a and 9.1c 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, e.g. [1, 27, 22, 24, 41]. 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 non-trivial 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 $P_{\mathcal{S}}(x_{\mathcal{S}}) = \sum_{x_{V \setminus \mathcal{S}}} P(x)$.

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

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

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

$$\mathcal{L}^{(k)}(P, R) + \mathcal{L}^{(k)}(R, Q) \geq \mathcal{L}^{(k)}(P, Q). \quad (6)$$

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 $P(X_i = + | X_S)$. We measure performance of the approximation Q by the expected magnitude of error $|P(x_i = + | X_S) - Q(x_i = + | X_S)|$ averaged over X_S :

$$\begin{aligned} \mathbb{E}_{X_S} |P(X_i = + | X_S) - Q(X_i = + | X_S)| \\ &= \sum_{x_S} |P(X_i = +, X_S = x_S) - Q(X_i = + | X_S = x_S)P(X_S = x_S)| \\ &\leq \sum_{x_S} |P(X_i = +, X_S = x_S) - Q(X_i = +, X_S = x_S)| + \sum_{x_S} |Q(x_S) - P(x_S)| \\ &\leq 2\mathcal{L}^{(|S|+1)}(P, Q). \end{aligned}$$

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

$$\begin{aligned} \mathcal{L}^{(2)}(P, \tilde{P}) &= \max_{w, \tilde{w} \in \mathcal{V}} \frac{1}{2} \sum_{x_w, x_{\tilde{w}} \in \{-, +\}^2} |P(x_w, x_{\tilde{w}}) - \tilde{P}(x_w, x_{\tilde{w}})| \\ &= \max_{w, \tilde{w} \in \mathcal{V}} \frac{1}{2} \left| \prod_{e \in \text{path}_{\mathsf{T}}(w, \tilde{w})} \mu_e - \prod_{e' \in \text{path}_{\tilde{\mathsf{T}}}(w, \tilde{w})} \tilde{\mu}_{e'} \right|. \end{aligned} \quad (7)$$

The second equality is derived by noting that $P(x_i = +) = 1/2$, $P(x_w, x_{\tilde{w}}) = [1 + x_w x_{\tilde{w}} \mathbb{E}_P[X_w X_{\tilde{w}}]] / 4$ and analogously for \tilde{P} . Also, as noted above after (1), it is immediate from Lemma 8.6 that $\mathbb{E}_P X_w X_{\tilde{w}} = \prod_{e \in \text{path}_{\mathsf{T}}(w, \tilde{w})} \mu_e$. The same holds for $\tilde{P} \in \mathcal{P}_{\tilde{\mathsf{T}}}$, which gives (7).

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 (7). The Chow-Liu algorithm produces the maximum likelihood tree, which minimizes the expected zero-one loss in (5). As shown in Theorem 3.3, the maximum likelihood tree also performs well in terms of accuracy of pairwise marginals.

Recall from (3) 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 (4), then with probability at least $1 - \delta$ we have $\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\text{CL}}}(P)) < \eta$.*

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 $\Psi : \{-1, +1\}^{p \times n} \rightarrow \mathcal{P}$ where $\mathcal{P} = \cup_{\mathsf{T}} \mathcal{P}_{\mathsf{T}}$ is the set of tree-structured Ising models with no external field defined in (1).

Theorem 3.4 (Samples necessary for small ssTV). *Fix $\eta > 0$. Suppose $\tanh(\beta) > \tanh(\alpha) + 2\eta$ and $n < C[1 - (\tanh(\alpha) + 2\eta)^2]\eta^{-2} \log p$. 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, i.e.,*

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

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 \hat{P} , if the tree T^{cl} is the Chow-Liu tree as defined in (4), then*

$$\mathbb{E} \left[\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}^{\text{cl}}}(\hat{P})) \right] < C' p \exp(-C n e^{-2\beta}) + C'' \sqrt{\frac{\log p}{n}}.$$

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*

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

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 4.1 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 $\hat{\mu}_{12}$, $\hat{\mu}_{23}$ and $\hat{\mu}_{13}$ are concentrated around $\mu_{12} = \epsilon$, $\mu_{23} = 1 - \epsilon$ and $\mu_{13} = \mu_{12}\mu_{23} = \epsilon(1 - \epsilon)$. Let $\hat{\mu}_{12} = \mu_{12} + z_{12}$, $\hat{\mu}_{23} = \mu_{23} + z_{23}$ and $\hat{\mu}_{13} = \mu_{13} + z_{13}$, 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 $\epsilon/10$.

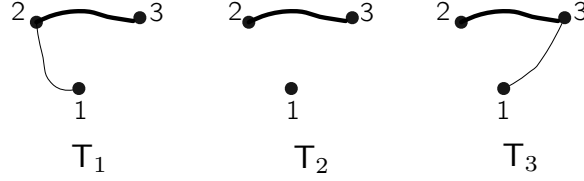


Figure 4.1: The original distribution factors according to T_1 . The width of the edges corresponds to their strength. The forest approximation algorithm (defined in Section 4.2) recovers $\hat{T} = T_2$ since the correlation between X_1 and each of the other variables is not strong enough to confidently 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.

Since $\max\{\mu_{12}, \mu_{13}\} = \max\{\epsilon, \epsilon(1 - \epsilon)\} = \epsilon \ll \mu_{23}$, concentration bounds guarantee that with high probability $\hat{\mu}_{23} > \max\{\hat{\mu}_{12}, \hat{\mu}_{13}\}$ and the (greedy implementation of) Chow-Liu algorithm described in (4) 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 $\hat{\mu}_{12} > \hat{\mu}_{13}$: if $z_{13} - z_{12} > \epsilon^2$, then edge $(1, 3)$ is added and $T^{\text{CL}} = T_3 \neq T_1$.

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 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 $e \in \mathcal{E}_T$ but $e \notin \mathcal{E}_{T^{\text{CL}}}$, 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 $\mathcal{E}_F \subseteq \mathcal{E}_T$. This is accomplished by finding a forest consisting of sufficiently strong edges, i.e. having weight at least τ for an appropriate value of τ . As shown by Tan et al. in [41], 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 $\hat{T} = T_2$ in Figure 4.1, whereas Chow-Liu would give $\hat{T} = T_1$ or $\hat{T} = T_3$. 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 $\mathcal{L}^{(2)}(P, \Pi_{\hat{T}}(P))$, for the above cases:

$$\begin{aligned} \hat{T} = T_1 &\rightarrow \mathcal{L}^{(2)}(P, \Pi_{\hat{T}}(P)) = 0, \\ \hat{T} = T_2 &\rightarrow \mathcal{L}^{(2)}(P, \Pi_{\hat{T}}(P)) = |\mu_{12}| = \epsilon, \\ \hat{T} = T_3 &\rightarrow \mathcal{L}^{(2)}(P, \Pi_{\hat{T}}(P)) = |\mu_{12}|(1 - \mu_{23}^2) = \epsilon^2(2 - \epsilon). \end{aligned}$$

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 [41, 27]. Mossel in [31] 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 [41] 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 $|\hat{\mu}_e| - \tau(\epsilon) - \epsilon$. The output $\hat{T} = (\mathcal{V}, \mathcal{E}_{\hat{T}})$ is a truncated version of T^{CL} such that the empirical correlation between any pair of nodes $(i, j) \in \mathcal{E}_{\hat{T}}$ satisfies $|\hat{\mu}_{ij}| \geq \tau(\epsilon) + \epsilon$.

Proposition 4.1. *Given $n > Ce^{2\beta}\eta^{-2} \log \frac{p}{\delta}$ samples, the forest approximation algorithm guarantees that $\mathcal{E}_{\hat{T}} \subseteq \mathcal{E}_T$ and $\mathcal{L}^{(2)}(P, \Pi_{\hat{T}}(\hat{P})) < \eta$ 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

$$\tilde{d}(P, Q) = \max \{ \mathcal{L}^{(2)}(P, Q), \mathbb{1}[\mathcal{E}_F \not\subseteq \mathcal{E}_T] \}, \quad (8)$$

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 $P_T(0, \beta)$ and recovers a forest F and a distribution $Q \in P_F(0, \beta)$. Let the (asymmetric) distance \tilde{d} be defined in (8). If $n < C \min\{p, e^{2\beta}\}/[\eta \text{atanh}\eta] \log p$, then the worst-case probability of \tilde{d} loss greater than η , taken over trees $T \in \mathcal{T}$ and distributions $P \in \mathcal{P}_T(\alpha, \beta)$, is at least half for any algorithm, i.e.,*

$$\inf_{\Psi} \sup_{\substack{T \in \mathcal{T} \\ P \in \mathcal{P}_T(\alpha, \beta)}} \mathbb{P} \left[\tilde{d}(P, \Psi(X^{(1:n)})) > \eta \right] > 1/2.$$

Comparison with Theorem 3.3 shows that avoiding adding wrong edges (as forced by the loss function (8)) 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 (6):

$$\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}_{\text{cl}}}(\hat{P})) \leq \mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}_{\text{cl}}}(P)) + \mathcal{L}^{(2)}(\Pi_{\mathsf{T}_{\text{cl}}}(P), \Pi_{\mathsf{T}_{\text{cl}}}(\hat{P})). \quad (9)$$

The first term on the right-hand side of Equation (9) 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 (7) tells us that for each pair of nodes $u, v \in \mathcal{V}$, $\text{path}_{\mathsf{T}}(u, v)$ and $\text{path}_{\mathsf{T}_{\text{cl}}}(u, v)$ must be compared.

The second term on the right-hand side of Equation (9) 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 (9).

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 $\mathcal{L}^{(2)}(\Pi_{\mathsf{T}_{\text{cl}}}(P), \Pi_{\mathsf{T}_{\text{cl}}}(\hat{P})) \leq \eta$.*

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 $\mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}_{\text{cl}}}(P)) \leq \eta$.*

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 $\mathsf{E}^{\text{corr}}(\epsilon)$, $\mathsf{E}^{\text{strong}}(\epsilon)$ and $\mathsf{E}^{\text{cascade}}(\epsilon)$ as follows. Let $\mathsf{E}^{\text{corr}}(\epsilon)$ be the event that all empirical correlations are within ϵ of population values:

$$\mathsf{E}^{\text{corr}}(\epsilon) = \left\{ \max_{w, \tilde{w} \in \mathcal{V}} |\mu_{w, \tilde{w}} - \hat{\mu}_{w, \tilde{w}}| \leq \epsilon \right\}. \quad (10)$$

Let

$$\tau(\epsilon) = \frac{4\epsilon}{\sqrt{1 - \tanh \beta}} \quad \text{and} \quad \mathcal{E}_{\mathsf{T}}^{\text{strong}}(\epsilon) = \{(i, j) \in \mathcal{E}_{\mathsf{T}} : |\mu_{ij}| > \tau(\epsilon)\} \quad (11)$$

consist of the set of strong edges in tree T . Let T^{cl} be the Chow-Liu tree defined in Equation (4). Weak edges are those that are not strong, i.e., $\mathcal{E}_{\mathsf{T}} \setminus \mathcal{E}_{\mathsf{T}}^{\text{strong}}(\epsilon)$. Let $\mathsf{E}^{\text{strong}}(\epsilon)$ be the event that all strong edges in T as defined in (11) are recovered by the Chow-Liu tree :

$$\mathsf{E}^{\text{strong}}(\epsilon) = \{\mathcal{E}_{\mathsf{T}}^{\text{strong}}(\epsilon) \subset \mathcal{E}_{\mathsf{T}_{\text{cl}}}\}. \quad (12)$$

Finally, define the event

$$\mathsf{E}^{\text{cascade}}(\epsilon) = \left\{ \mathcal{L}^{(2)}(P, \Pi_{\mathsf{T}}(\hat{P})) \leq \epsilon \right\}. \quad (13)$$

Recall that P factorizes according to T and \hat{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 $\mathsf{E}(\epsilon, \gamma) := \mathsf{E}^{\text{corr}}(\epsilon) \cap \mathsf{E}^{\text{strong}}(\epsilon) \cap \mathsf{E}^{\text{cascade}}(\gamma)$. Lemmas 8.1, 8.5 and 8.7 prove that for

$$\epsilon_0 = \min\{e^{-\beta}/24, \eta/16\} \quad \text{and} \quad \gamma_0 = \eta/3, \quad (14)$$

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

$$\mathsf{P}[\mathsf{E}(\epsilon_0, \gamma_0)] \geq 1 - \delta. \quad (15)$$

Sketch of proof of Proposition 5.1 The proof entails showing that on the event $E(\epsilon_0, \gamma_0)$ for ϵ_0 and γ_0 defined in (14) we have the desired inequality $\mathcal{L}^{(2)}(\Pi_{\mathcal{T}^{\text{cl}}}(P), \Pi_{\mathcal{T}^{\text{cl}}}(\hat{P})) \leq \eta$. The result then follows from (15).

To bound $\mathcal{L}^{(2)}(\Pi_{\mathcal{T}^{\text{cl}}}(P), \Pi_{\mathcal{T}^{\text{cl}}}(\hat{P}))$ on event $E(\epsilon_0, \gamma_0)$, we consider parameter estimation errors along paths in \mathcal{T}^{cl} . First, observe that on event $E^{\text{cascade}}(\gamma_0)$ defined in (13), the end-to-end error for each path in $\mathcal{E}_{\mathcal{T}^{\text{cl}}} \cap \mathcal{E}_{\mathcal{T}}$ is bounded by γ_0 . Next, we study parameter estimation error in paths containing (falsely added) edges in $\mathcal{E}_{\mathcal{T}^{\text{cl}}} \setminus \mathcal{E}_{\mathcal{T}}$.

For any pair of nodes w, \tilde{w} , denote by $t = |\text{path}_{\mathcal{T}^{\text{cl}}}(w, \tilde{w}) \setminus \mathcal{E}_{\mathcal{T}}|$ the number of falsely added edges in the path connecting them in \mathcal{T}^{cl} . As discussed in the proof, these edges correspond to missed edges in \mathcal{T} and thus $E^{\text{strong}}(\epsilon_0)$ guarantees that these edges are weak (as defined after (11)). These t weak edges break up the $\text{path}_{\mathcal{T}^{\text{cl}}}(w, \tilde{w})$ into at most $t + 1$ contiguous segments $\mathcal{F}_0, \mathcal{F}_1, \dots, \mathcal{F}_t$, each entirely within $\mathcal{E}_{\mathcal{T}^{\text{cl}}} \cap \mathcal{E}_{\mathcal{T}}$.

On event $E^{\text{cascade}}(\gamma_0)$ 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 $\mathcal{L}^{(2)}(P, \Pi_{\mathcal{T}^{\text{cl}}}(P)) \leq \eta$ on the event $E^{\text{corr}}(\epsilon_0) \cap E^{\text{strong}}(\epsilon_0) \supseteq E(\epsilon_0, \gamma_0)$ for ϵ_0 defined in (14).

The proof of the proposition sets up a careful induction on the distance between nodes (computed in \mathcal{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 \mathcal{T} and \mathcal{T}^{cl} . The lemma states that for any two spanning trees on p nodes, and for any two nodes $w, \tilde{w} \in [p]$, there exists at least one pair of edges $f \in \text{path}_{\mathcal{T}}(w, \tilde{w})$ and $g \in \text{path}_{\mathcal{T}^{\text{cl}}}(w, \tilde{w})$ satisfying a collection of properties illustrated in Figure 6.1 (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 $\mu_g = \mu_f \mu_A \mu_C \mu_{\tilde{A}} \mu_{\tilde{C}}$, hence $|\mu_g| \leq |\mu_f|$. But $|\hat{\mu}_g| \geq |\hat{\mu}_f|$, since the Chow-Liu algorithm chose g in \mathcal{T}^{cl} instead of f . Tight control on the relationship between $|\hat{\mu}_g|$ and $|\hat{\mu}_f|$ 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, \tilde{w} with $|\text{path}_{\mathcal{T}^{\text{cl}}}(w, \tilde{w})| = d$.

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 $E(\epsilon_0, \gamma_0)$ for ϵ_0 and γ_0 defined in (14) the desired inequality $\mathcal{L}^{(2)}(\Pi_{\mathcal{T}^{\text{cl}}}(P), \Pi_{\mathcal{T}^{\text{cl}}}(\hat{P})) \leq \eta$ holds. Equation (15) gives the result.

Let $\tau(\epsilon_0)$ (defined in (11)) be the threshold to define $\mathcal{E}_{\mathcal{T}}^{\text{strong}}(\epsilon_0)$, the set of strong edges in \mathcal{T} . For any pair of nodes w, \tilde{w} , consider $\text{path}_{\mathcal{T}^{\text{cl}}}(w, \tilde{w})$. Let $0 \leq t < p$ be the number of weak edges $e_1, \dots, e_t \in \text{path}_{\mathcal{T}^{\text{cl}}}(w, \tilde{w})$ such that $|\mu_{e_i}| \leq \tau(\epsilon_0)$. There are at most $t + 1$ contiguous sub-paths in $\text{path}_{\mathcal{T}^{\text{cl}}}(w, \tilde{w})$ 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 $\text{path}_{\mathcal{T}^{\text{cl}}}(w, \tilde{w})$, then $\mathcal{F}_i = \emptyset$, in which case we define $\mu_{\mathcal{F}_i} = \hat{\mu}_{\mathcal{F}_i} = 1$. By definition of \mathcal{F}_i , all edges $f \in \mathcal{F}_i$ are strong.

According to (12), under the event $\mathbf{E}^{\text{strong}}(\epsilon_0)$ all strong edges in \mathbf{T} are recovered in \mathbf{T}^{CL} . Thus, $\mathcal{F}_i \subseteq \mathcal{E}_{\mathbf{T}}$ is a path not only in \mathbf{T}^{CL} but also in \mathbf{T} , which guarantees $|\hat{\mu}_{\mathcal{F}_i} - \mu_{\mathcal{F}_i}| \leq \gamma_0$ under the event $\mathbf{E}^{\text{cascade}}(\gamma_0)$.

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

$$\begin{aligned}
& \left| \prod_{e \in \text{path}_{\mathbf{T}^{\text{CL}}}(w, \tilde{w})} \hat{\mu}_e - \prod_{e \in \text{path}_{\mathbf{T}^{\text{CL}}}(w, \tilde{w})} \mu_e \right| \\
& \stackrel{(a)}{=} \left| \hat{\mu}_{\mathcal{F}_0} \prod_{i=1}^t \hat{\mu}_{\mathcal{F}_i} \hat{\mu}_{e_i} - \mu_{\mathcal{F}_0} \prod_{i=1}^t \mu_{\mathcal{F}_i} \mu_{e_i} \right| \\
& \stackrel{(b)}{\leq} |\hat{\mu}_{\mathcal{F}_0} - \mu_{\mathcal{F}_0}| \prod_{j=1}^t |\mu_{\mathcal{F}_j} \mu_{e_j}| + \sum_{i=1}^t |\hat{\mu}_{\mathcal{F}_i} \hat{\mu}_{e_i} - \mu_{\mathcal{F}_i} \mu_{e_i}| \cdot |\hat{\mu}_{\mathcal{F}_0}| \prod_{j=1}^{i-1} |\hat{\mu}_{\mathcal{F}_j} \hat{\mu}_{e_j}| \prod_{k=i+1}^t |\mu_{\mathcal{F}_k} \mu_{e_k}| \\
& \stackrel{(c)}{\leq} \gamma_0 [\tau(\epsilon_0)]^t + (\tau(\epsilon_0) + \epsilon_0)^{t-1} \sum_{i=1}^t |\hat{\mu}_{\mathcal{F}_i} \hat{\mu}_{e_i} - \mu_{\mathcal{F}_i} \mu_{e_i}| \\
& \stackrel{(d)}{\leq} \gamma_0 [\tau(\epsilon_0)]^t + (\tau(\epsilon_0) + \epsilon_0)^{t-1} \left[\sum_{i=1}^t |\mu_{\mathcal{F}_i} (\hat{\mu}_{e_i} - \mu_{e_i})| + |\hat{\mu}_{e_i} (\hat{\mu}_{\mathcal{F}_i} - \mu_{\mathcal{F}_i})| \right] \\
& \stackrel{(e)}{\leq} (\tau(\epsilon_0) + \epsilon_0)^{t-1} (2t+1) \max\{\gamma_0, \epsilon_0\} \stackrel{(f)}{\leq} \frac{2t+1}{4^{t-1}} \frac{\eta}{3} \stackrel{(g)}{\leq} \eta.
\end{aligned}$$

In (a) we use $\text{path}_{\mathbf{T}^{\text{CL}}}(w, \tilde{w}) = \{\mathcal{F}_0, e_1, \dots, \mathcal{F}_t, e_t, \mathcal{F}_t\}$. (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 $|\mu_{\mathcal{F}_i}|, |\hat{\mu}_{\mathcal{F}_i}| \leq 1$, $|\mu_{e_i}| \leq \tau(\epsilon_0)$, $|\hat{\mu}_{e_i}| \leq \tau(\epsilon_0) + \epsilon_0$ on $\mathbf{E}^{\text{corr}}(\epsilon_0)$ and $|\hat{\mu}_{\mathcal{F}_0} - \mu_{\mathcal{F}_0}| \leq \gamma_0$ on $\mathbf{E}^{\text{cascade}}(\gamma_0)$. (d) uses triangle inequality. In (e), we use $|\hat{\mu}_{\mathcal{F}_i} - \mu_{\mathcal{F}_i}| \leq \gamma_0$ on the event $\mathbf{E}^{\text{cascade}}(\gamma_0)$ and $|\hat{\mu}_{e_i} - \mu_{e_i}| \leq \epsilon_0$ on the event $\mathbf{E}^{\text{corr}}(\epsilon_0)$. (f) is true since $\gamma_0, \epsilon_0 \leq \eta/3$ (as in (14)). Also, $1 - \tanh \beta \geq e^{-2\beta}$ and the definition of $\tau(\epsilon_0)$ in (11) 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 (14)). (g) holds for all $t \geq 1$. \square

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 $|\hat{\mu}_{ij}|$ on each edge (i, j) . Let \mathbf{T}^{CL} be the maximum weight spanning tree of this graph. If edge $(u, \tilde{u}) \notin \mathcal{E}_{\mathbf{T}^{\text{CL}}}$, then $|\hat{\mu}_{u\tilde{u}}| \leq |\hat{\mu}_{ij}|$ for all $(i, j) \in \text{path}_{\mathbf{T}^{\text{CL}}}(u, \tilde{u})$.*

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

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

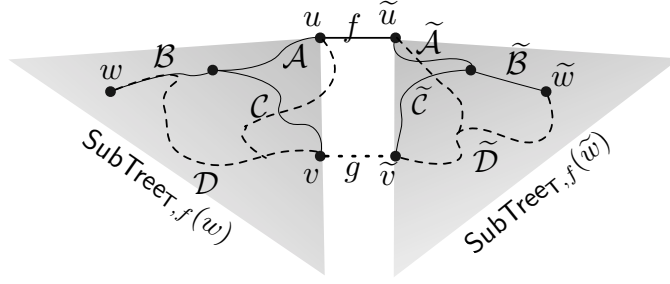


Figure 6.1: Schematic for the proof of Proposition 5.2. The solid lines represent paths in T and the dashed lines represent paths in T^{CL} . The sets of edges \mathcal{D} and $\tilde{\mathcal{D}}$ may overlap with $\mathcal{A} \cup \mathcal{B}$ and $\tilde{\mathcal{A}} \cup \tilde{\mathcal{B}}$.

Our goal is to bound $\mathcal{L}^{(2)}(P, \Pi_{T^{CL}}(P)) = \max_{w, \tilde{w} \in \mathcal{V}} \text{error}_{P, T^{CL}}(w, \tilde{w})$. We will prove that on the event $\mathbf{E}^{\text{corr}}(\epsilon_0) \cap \mathbf{E}^{\text{strong}}(\epsilon_0) \supseteq \mathbf{E}(\epsilon_0, \gamma_0)$ for ϵ_0 defined in (14), $\mathcal{L}^{(2)}(P, \Pi_{T^{CL}}(P)) < \eta$ holds. The result then follows from (15).

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

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

For nodes at distance one in T^{CL} , i.e. $|\text{path}_{T^{CL}}(w, \tilde{w})| = 1$, it follows that $\text{error}_{P, T^{CL}}(w, \tilde{w}) = 0$ from the definition of the projected distribution $\Pi_{T^{CL}}(P)$ (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 $\mathbf{E}^{\text{corr}}(\epsilon_0) \cap \mathbf{E}^{\text{strong}}(\epsilon_0)$ with ϵ_0 defined in (14), if $\Delta(k) \leq \eta$ for all $k < d$, then $\Delta(d) \leq \eta$ which gives the result.

Note that if $\text{path}_{T^{CL}}(w, \tilde{w}) = \text{path}_T(w, \tilde{w})$, then $\text{error}_{P, T^{CL}}(w, \tilde{w}) = 0$ (again because correlations are matched on edges). Thus, we assume $\text{path}_{T^{CL}}(w, \tilde{w}) \neq \text{path}_T(w, \tilde{w})$. Lemma 8.8 shows the existence of a pair of edges $f = (u, \tilde{u}) \in \mathcal{E}_T \setminus \mathcal{E}_{T^{CL}}$ and $g = (v, \tilde{v}) \in \mathcal{E}_{T^{CL}} \setminus \mathcal{E}_T$ such that (See Figure 6.1):

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

Here $\text{SubTree}_{T,f}(w) = \{i \in \mathcal{V}; f \notin \text{path}_T(w, i)\}$ is the set of nodes connected to w in T after removing edge f (see Figure 6.1). We define several sub-paths:

$$\begin{aligned} \mathcal{A} &= \text{path}_T(u, w) \cap \text{path}_T(u, v), & \mathcal{B} &= \text{path}_T(u, w) \setminus \text{path}_T(u, v), \\ \mathcal{C} &= \text{path}_T(u, v) \setminus \text{path}_T(u, w), & \mathcal{D} &= \text{path}_{T^{CL}}(w, v). \end{aligned}$$

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

$$\begin{aligned} \tilde{\mathcal{A}} &= \text{path}_T(\tilde{u}, \tilde{w}) \cap \text{path}_T(\tilde{u}, \tilde{v}), & \tilde{\mathcal{B}} &= \text{path}_T(\tilde{u}, \tilde{w}) \setminus \text{path}_T(\tilde{u}, \tilde{v}), \\ \tilde{\mathcal{C}} &= \text{path}_T(\tilde{u}, \tilde{v}) \setminus \text{path}_T(\tilde{u}, \tilde{w}), & \tilde{\mathcal{D}} &= \text{path}_{T^{CL}}(\tilde{w}, \tilde{v}). \end{aligned}$$

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

Lemma 6.1 applied to $f = (u, \tilde{u}) \notin \mathcal{E}_{T^{\text{cl}}}$ and $g = (v, \tilde{v}) \in \text{path}_{T^{\text{cl}}}(u, \tilde{u})$ gives $|\hat{\mu}_f| \leq |\hat{\mu}_g|$. Also, $f \in \text{path}_T(v, \tilde{v})$, hence $|\mu_g| \leq |\mu_f|$. On the event $\mathbf{E}^{\text{corr}}(\epsilon_0)$, $|\mu_f| - 2\epsilon_0 \leq |\hat{\mu}_f| \leq |\hat{\mu}_g| \leq |\mu_g| + 2\epsilon_0 \leq |\mu_f| + 2\epsilon_0$ which gives $|\mu_f| - 4\epsilon_0 \leq |\mu_f \mu_{\mathcal{A}} \mu_{\mathcal{C}} \mu_{\tilde{\mathcal{A}}} \mu_{\tilde{\mathcal{C}}}| \leq |\mu_f|$. Also, $|\mu_{\mathcal{A}} \mu_{\mathcal{C}} \mu_{\tilde{\mathcal{A}}} \mu_{\tilde{\mathcal{C}}}| \leq |\mu_{\mathcal{C}} \mu_{\tilde{\mathcal{C}}}| \leq 1$. Thus,

$$|\mu_f| \left(1 - \mu_{\mathcal{C}}^2 \mu_{\tilde{\mathcal{C}}}^2\right) \leq 2|\mu_f| (1 - |\mu_{\mathcal{C}} \mu_{\tilde{\mathcal{C}}}|) \leq 2|\mu_f| (1 - |\mu_{\mathcal{A}} \mu_{\mathcal{C}} \mu_{\tilde{\mathcal{A}}} \mu_{\tilde{\mathcal{C}}}|) \leq 8\epsilon_0. \quad (16)$$

Since $f \in \mathcal{E}_T \setminus \mathcal{E}_{T^{\text{cl}}}$, under the event $\mathbf{E}^{\text{strong}}(\epsilon_0)$, f cannot be a strong edge as defined in (11). It follows that $|\mu_f| \leq \tau(\epsilon_0)$ for $\tau(\epsilon_0)$ defined in (11).

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

$$\begin{aligned} \text{error}_{P, T^{\text{cl}}}(w, v) &= |\mu_{\mathcal{D}} - \mu_{\mathcal{B}} \mu_{\mathcal{C}}| \leq \Delta(k), \\ \text{error}_{P, T^{\text{cl}}}(\tilde{w}, \tilde{v}) &= |\mu_{\tilde{\mathcal{D}}} - \mu_{\tilde{\mathcal{B}}} \mu_{\tilde{\mathcal{C}}}| \leq \Delta(\tilde{k}). \end{aligned} \quad (17)$$

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

$$\begin{aligned} \text{error}_{P, T^{\text{cl}}}(w, \tilde{w}) &= |\mu_{\mathcal{D}} \mu_g \mu_{\tilde{\mathcal{D}}} - \mu_{\mathcal{A}} \mu_{\mathcal{B}} \mu_f \mu_{\tilde{\mathcal{A}}} \mu_{\tilde{\mathcal{B}}}| \\ &= |\mu_{\mathcal{D}} \mu_{\mathcal{A}} \mu_{\mathcal{C}} \mu_f \mu_{\tilde{\mathcal{A}}} \mu_{\tilde{\mathcal{C}}} \mu_{\tilde{\mathcal{D}}} - \mu_{\mathcal{A}} \mu_{\mathcal{B}} \mu_f \mu_{\tilde{\mathcal{A}}} \mu_{\tilde{\mathcal{B}}}| \\ &= |\mu_{\mathcal{A}} \mu_f \mu_{\tilde{\mathcal{A}}} \cdot |\mu_{\mathcal{C}} \mu_{\tilde{\mathcal{C}}} (\mu_{\mathcal{D}} - \mu_{\mathcal{B}} \mu_{\mathcal{C}} + \mu_{\mathcal{B}} \mu_{\mathcal{C}}) (\mu_{\tilde{\mathcal{D}}} - \mu_{\tilde{\mathcal{B}}} \mu_{\tilde{\mathcal{C}}} + \mu_{\tilde{\mathcal{B}}} \mu_{\tilde{\mathcal{C}}}) - \mu_{\mathcal{B}} \mu_{\tilde{\mathcal{B}}}| \\ &\leq |\mu_{\mathcal{A}} \mu_f \mu_{\tilde{\mathcal{A}}}| \cdot [|\mu_{\mathcal{C}} \mu_{\tilde{\mathcal{C}}} \mu_{\mathcal{B}} \mu_{\mathcal{C}} \mu_{\tilde{\mathcal{B}}} \mu_{\tilde{\mathcal{C}}} - \mu_{\mathcal{B}} \mu_{\tilde{\mathcal{B}}}| + |\mu_{\mathcal{C}} \mu_{\tilde{\mathcal{C}}} (\mu_{\mathcal{D}} - \mu_{\mathcal{B}} \mu_{\mathcal{C}}) (\mu_{\mathcal{D}} - \mu_{\mathcal{B}} \mu_{\mathcal{C}})| \\ &\quad + |\mu_{\mathcal{C}} \mu_{\tilde{\mathcal{C}}} \mu_{\tilde{\mathcal{B}}} \mu_{\tilde{\mathcal{C}}} (\mu_{\mathcal{D}} - \mu_{\mathcal{B}} \mu_{\mathcal{C}})| + |\mu_{\mathcal{C}} \mu_{\tilde{\mathcal{C}}} \mu_{\mathcal{B}} \mu_{\mathcal{C}} (\mu_{\tilde{\mathcal{D}}} - \mu_{\tilde{\mathcal{B}}} \mu_{\tilde{\mathcal{C}}})|] \\ &\stackrel{(a)}{\leq} |\mu_f \mu_{\mathcal{A}} \mu_{\tilde{\mathcal{A}}} \mu_{\mathcal{B}} \mu_{\tilde{\mathcal{B}}}| |\mu_{\mathcal{C}}^2 \mu_{\tilde{\mathcal{C}}}^2 - 1| + |\mu_f| \left(\Delta(k) \Delta(\tilde{k}) + \Delta(k) + \Delta(\tilde{k}) \right) \\ &\stackrel{(b)}{\leq} 8\epsilon_0 + \tau(\epsilon_0) \left(\Delta(k) + \Delta(\tilde{k}) + \Delta(k) \Delta(\tilde{k}) \right) \stackrel{(c)}{\leq} 8\epsilon_0 + 4\epsilon_0 e^{\beta} (2\eta + \eta^2) \leq \eta. \end{aligned}$$

Inequality (a) follows from (17). (b) uses Equation (16) and $|\mu_f| \leq \tau(\epsilon_0) \leq 4\epsilon_0 e^{\beta}$. We showed that $\Delta(1) = 0$. In (c) we use the inductive assumption $\Delta(k) \leq \eta$ for all $k < d$ and the assumption ϵ_0 defined in (14). Since w and \tilde{w} were arbitrary, this proves $\Delta(d) \leq \eta$, and moreover this holds for all d . \square

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

$$J(\theta \parallel \theta') \triangleq D(\theta \parallel \theta') + D(\theta' \parallel \theta) = \sum_{i < j} (\theta_{ij} - \theta'_{ij}) (\mu_{ij} - \mu'_{ij}). \quad (18)$$

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 [43]). *Assume that $M \geq 2$ and that Θ is a family of models $\theta^0, \theta^1, \dots, \theta^M$. Let Q_{θ^j} denote the probability law of the observation X under model θ^j . Let $\Phi : \{-1, +1\}^{p \times n} \rightarrow \{0, 1, \dots, M\}$ denote an estimator using n i.i.d. samples $X^{(1:n)}$. If*

$$n < (1 - \delta) \frac{\log M}{\frac{1}{M+1} \sum_{j=1}^M J(Q_{\theta^j} \| Q_{\theta^0})}, \quad (19)$$

then the probability of error of any algorithm is bounded as:

$$\inf_{\Phi} \max_{0 \leq j \leq M} Q_{\theta^j} [\Phi(X^{(1:n)}) \neq j] \geq \delta - \frac{1}{\log M}.$$

The following corollary is a restatement of Equation (2.9) in [43] 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 $\theta^0, \dots, \theta^M$ such that $d(\theta^k, \theta^j) \geq 2\eta$ for all $j \neq k$. If n satisfies (19), then any estimator $\Psi : X^{(1:n)} \rightarrow \Theta$ 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 $1/2$:*

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

Proof of Theorem 3.4. We consider a fixed tree structure given by a path, or in other words a Markov chain $X_1 - X_2 - \dots - X_p$. We choose M different parameter vectors θ^m , $0 \leq m \leq M-1$, for $M = p$.

Let $\theta_{i,i+1}^0 = \alpha$ for $i = 1, \dots, p-1$. In the m -th model we have $\theta_{m,m+1}^m = \text{atanh}(\tanh \alpha + 2\eta)$ and the remaining edge weights $\theta_{i,i+1}^m = \alpha$ for $i \neq m$. For $m' \neq m$,

$$\max_{i,j} |\mathbb{E}_{\theta^{m'}} [X_i X_j] - \mathbb{E}_{\theta^m} [X_i X_j]| \geq 2\eta.$$

Also, using (18)

$$J(\theta^m \| \theta^{m'}) \leq 2\eta [\text{atanh}(\tanh \alpha + 2\eta) - \alpha] \leq 2\eta \frac{2\eta}{1 - [\tanh \alpha + 2\eta]^2},$$

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

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 $\theta_{i,i+1}^0 = \alpha$ for odd values of i and $\theta_{i,i+1}^0 = \beta$ for even values of i . For each odd value of $m \leq p-2$ we let θ^m be equal to θ^0 everywhere except $\theta_{m,m+1}^m = 0$ and

$\theta_{m,m+2}^m = \text{atanh}(\tanh(\alpha) \tanh(\beta))$. There are $(p+1)/2$ models in total (including θ^0). A small calculation using (18) leads to

$$J(\theta^m \parallel \theta^0) = \alpha \tanh \alpha [1 - (\tanh \beta)^2] \leq 4\alpha^2 e^{-2\beta}.$$

Here we used $\tanh \alpha \leq \alpha$ and $1 - (\tanh \beta)^2 \leq 4e^{-2\beta}$ for $\alpha, \beta \geq 0$. Plugging the last display into Fano's inequality (Lemma 6.2) completes the proof. \square

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

Consider the original tree T with parameters $\alpha \leq |\theta_{ij}| \leq \beta$ for $(i, j) \in \mathcal{E}_{\mathsf{T}}$. Using Definition 12, the Chow-Liu algorithm recovers strong edges on the event $\mathsf{E}^{\text{strong}}(\epsilon)$, 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 $\mathsf{E}^{\text{strong}}(\epsilon)$ we have $\mathsf{T}^{\text{CL}} = \mathsf{T}$. Note that by Lemma 8.5, the event $\mathsf{E}^{\text{strong}}(\epsilon)$ (defined in (12)) 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 $\mathsf{T}^{\text{CL}} = \mathsf{T}$ with probability at least $1 - \delta$. Using $1 - \tanh \beta \geq e^{-2\beta}$ gives the statement of Theorem 3.2. \square

7.3 Lower bounding sample complexity of forest approximation algorithm, Proof of Proposition 4.2

Since the loss function $\tilde{d}(P, Q)$ defined in (8) 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 $\theta_{12}^1 = 0$, $\theta_{2i}^1 = \beta$ for $3 \leq i \leq p$, and $\theta_{ij}^1 = 0$ for all other edges so that $\mu_{1i} = 0$ for all $i \geq 2$ in this model. For $2 \leq m \leq M$, define θ^m (m -th model) such that $\theta_{12}^m = 0$, $\theta_{2i}^m = \beta$ for $3 \leq i \leq p$ and $\theta_{1,m+1}^m = \text{atanh}(\eta)$. Using (18), since $\mu_{1,m+1} = \eta$ and $\mu_{1i} = \eta \tanh^2(\beta)$ for $i \geq 3$ in model $m \geq 2$,

$$\frac{1}{M} \sum_{m=1}^M J(\theta^m \parallel \theta^M) \leq \eta \text{atanh}(\eta) \left[\frac{1}{M} + \left(1 - \frac{1}{M}\right) 8e^{-2\beta} \right] \leq 16\eta \text{atanh}(\eta) \max \left\{ \frac{1}{p}, e^{-2\beta} \right\}. \quad (20)$$

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 $\Phi : \{-1, +1\}^{p \times n} \rightarrow \{0, 1\}^{\mathcal{V} \setminus \{1\}}$ 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

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

(a) holds since $\{P_{\theta^m}\}_{m=1}^M \subseteq \bigcup_{T \in \mathcal{T}} \mathcal{P}_T(0, \beta)$. (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 $\tilde{d} \geq \eta$. This is because $\tilde{d} = 1$ if any extra edges are added and otherwise, if there are no extra edges and an edge incident to node 1 is missing, then $\tilde{d} \geq \eta$ 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 (20) and specifying the distance. Let $\partial_F(1)$ be the neighborhood of node 1 in forest F . For P defined on forest F and Q defined on F' , we use the distance $\tilde{d}'(P, Q) = \mathbb{1}[\partial_F(1) \neq \partial_{F'}(1)]$ (zero-one loss on neighborhood of node 1).

8 Control of events E^{corr} , E^{strong} and E^{cascade}

We state a standard form of Hoeffding's inequality [23] in Appendix D and use it here.

Lemma 8.1. *The event $E^{\text{corr}}(\epsilon)$ defined in (10) occurs with probability at least $1 - 2p^2 \exp(-n\epsilon^2/2)$.*

Proof. For a given pair of nodes w, \tilde{w} , let $Z^{(i)} = X_w^{(i)} X_{\tilde{w}}^{(i)}$ and apply Hoeffding's inequality (Lemma D.1 in Appendix D) to get $\mathbb{P}[|\mu_{w, \tilde{w}} - \hat{\mu}_{w, \tilde{w}}| > \epsilon] \leq 2 \exp(-n\epsilon^2/2)$. Applying the union bound over $\binom{p}{2}$ pairs $w, \tilde{w} \in \mathcal{V}$ of nodes completes the proof. \square

We next prove Lemma 6.1 for completeness.

Proof of Lemma 6.1. For edge $(u, \tilde{u}) \notin \mathcal{E}_{T^{\text{cl}}}$, if there is an edge $(i, j) \in \text{path}_{T^{\text{cl}}}(u, \tilde{u})$ such that $|\hat{\mu}_{u\tilde{u}}| > |\hat{\mu}_{ij}|$, then T^{cl} cannot be the maximum weight spanning tree. To show that, consider the tree T' identical to T^{cl} except $(u, \tilde{u}) \in \mathcal{E}_{T'}$ and $(i, j) \notin \mathcal{E}_{T'}$ (i.e., $\mathcal{E}_{T'} = (\mathcal{E}_{T^{\text{cl}}} \setminus \{(i, j)\}) \cup \{(u, \tilde{u})\}$). Note that T' is a spanning tree and observe that $\text{weight}(T') \triangleq \sum_{e \in \mathcal{E}_{T'}} |\hat{\mu}_e| = \sum_{e \in \mathcal{E}_{T^{\text{cl}}}} |\hat{\mu}_e| + |\hat{\mu}_{u\tilde{u}}| - |\hat{\mu}_{ij}| > \text{weight}(T^{\text{cl}})$. \square

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, \tilde{v} and edge $f = (u, \tilde{u}) \in \text{path}_T(v, \tilde{v})$, let

$$Z_{f,v,\tilde{v}} = X_u X_{\tilde{u}} - X_v X_{\tilde{v}} = X_u X_{\tilde{u}} (1 - X_u X_v X_{\tilde{v}} X_{\tilde{u}}), \quad (21)$$

$$Y_{f,v,\tilde{v}} = X_u X_{\tilde{u}} + X_v X_{\tilde{v}} = X_u X_{\tilde{u}} (1 + X_u X_v X_{\tilde{v}} X_{\tilde{u}}). \quad (22)$$

Lemma 8.2. *If there exists a pair of edges $f = (u, \tilde{u})$ and $g = (v, \tilde{v})$ such that $f \in \mathcal{E}_T \setminus \mathcal{E}_{T^{\text{cl}}}$, $g \in \mathcal{E}_{T^{\text{cl}}} \setminus \mathcal{E}_T$ and additionally $f \in \text{path}_T(v, \tilde{v})$ and $g \in \text{path}_{T^{\text{cl}}}(u, \tilde{u})$, then,*

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

Proof. Using Lemma 6.1, $f = (u, \tilde{u}) \notin \mathcal{E}_{T^{\text{cl}}}$ and $g = (v, \tilde{v}) \in \text{path}_{T^{\text{cl}}}(u, \tilde{u})$ implies that $|\hat{\mu}_g| \geq |\hat{\mu}_f|$. Hence $\hat{\mu}_g^2 \geq \hat{\mu}_f^2$ and

$$\begin{aligned} 0 &\geq \hat{\mu}_f^2 - \hat{\mu}_g^2 = (\hat{\mu}_f - \hat{\mu}_g)(\hat{\mu}_f + \hat{\mu}_g) \\ &= \frac{1}{n^2} \left(\sum_{i=1}^n X_u^{(i)} X_{\tilde{u}}^{(i)} - X_v^{(i)} X_{\tilde{v}}^{(i)} \right) \left(\sum_{i=1}^n X_u^{(i)} X_{\tilde{u}}^{(i)} + X_v^{(i)} X_{\tilde{v}}^{(i)} \right) \end{aligned}$$

$$= \frac{1}{n^2} \left(\sum_{i=1}^n Z_{f,v,\tilde{v}}^{(i)} \right) \left(\sum_{i=1}^n Y_{f,v,\tilde{v}}^{(i)} \right)$$

where in the last step we used $f \in \text{path}_{\mathbb{T}}(v, \tilde{v})$ and the definition of $Z_{f,w,\tilde{w}}$ and $Y_{f,w,\tilde{w}}$ in (21) and (22). \square

Later we will bound the probability of the event in Lemma 8.2. To this end, we will derive deviation bounds on $Z_{f,v,\tilde{v}}$ and $Y_{f,v,\tilde{v}}$ in Lemma 8.3. We use the standard Bernstein's inequality as quoted from [43] in Appendix D

Lemma 8.3. *For all pairs of nodes $v, \tilde{v} \in \mathcal{V}$ and edges $f = (u, \tilde{u}) \in \text{path}_{\mathbb{T}}(v, \tilde{v})$, let $\mathcal{A}_{f,v,\tilde{v}} = \text{path}_{\mathbb{T}}(v, \tilde{v}) \setminus \{f\}$ such that $\mu_{v\tilde{v}} = \mu_f \mu_{\mathcal{A}_{f,v,\tilde{v}}}$. Given n i.i.d. samples let $Z_{f,v,\tilde{v}}^{(1)}, \dots, Z_{f,v,\tilde{v}}^{(n)}$ be defined in (21), and $Y_{f,v,\tilde{v}}^{(1)}, \dots, Y_{f,v,\tilde{v}}^{(n)}$ be defined in (22). Let $\epsilon = \sqrt{2/n \log(2p^2/\delta)}$. Then, with probability at least $1 - \delta$*

$$\left| \frac{1}{n} \sum_{i=1}^n Z_{f,v,\tilde{v}}^{(i)} - \mu_f(1 - \mu_{\mathcal{A}_{f,v,\tilde{v}}}) \right| \leq \max \{4\epsilon^2, 4\epsilon\sqrt{1 - \mu_{\mathcal{A}_{f,v,\tilde{v}}}}\} \quad \text{and} \quad (23)$$

$$\left| \frac{1}{n} \sum_{i=1}^n Y_{f,v,\tilde{v}}^{(i)} - \mu_f(1 + \mu_{\mathcal{A}_{f,v,\tilde{v}}}) \right| \leq \max \{4\epsilon^2, 4\epsilon\sqrt{1 + \mu_{\mathcal{A}_{f,v,\tilde{v}}}}\}. \quad (24)$$

Proof. We prove that (23) holds with probability at least $1 - \delta/2$. The proof of (24) is analogous.

We use the abbreviation \mathcal{A} instead of $\mathcal{A}_{f,v,\tilde{v}}$ and Z instead of $Z_{f,v,\tilde{v}}$ in this proof. Applying Lemma 8.6, it follows from the fact that P is Markov with respect to \mathbb{T} and $f = (u, \tilde{u}) \in \text{path}_{\mathbb{T}}(v, \tilde{v})$ that $X_u X_{\tilde{u}}$ and $X_u X_v X_{\tilde{u}} X_{\tilde{v}}$ are independent random variables. Note that $P(X_u X_{\tilde{u}} = 1) = (1 + \mu_f)/2$ and $P(X_u X_{\tilde{u}} = -1) = (1 - \mu_f)/2$. Similarly, the distribution of $X_u X_v X_{\tilde{u}} X_{\tilde{v}}$ is a function of $\mu_{\mathcal{A}}$. As a result, the random variable $Z_{f,v,\tilde{v}} \in \{-2, 0, 2\}$ defined in (21) has the following distribution:

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

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

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

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

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

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

Event $\mathbf{E}^{\text{strong}}(\epsilon)$ in (12) occurs if all of the strong edges in \mathbb{T} (defined in (11)) are recovered in \mathbb{T}^{cl} . Lemma 8.4 shows that the deviation bounds for the variables $Z_{f,v,\tilde{v}}$ and $Y_{f,v,\tilde{v}}$ stated in (23) and (24) imply $\mathbf{E}^{\text{strong}}(\epsilon)$.

Lemma 8.4. *Under the events described in Lemma 8.3, if there is an edge $f \in \mathcal{E}_T$ missing from the Chow-Liu tree, $f \notin \mathcal{E}_{T^{\text{cl}}}$, then $|\mu_f| \leq \tau(\epsilon) = \frac{4\epsilon}{\sqrt{1-\tanh\beta}}$ (i.e., $\mathbf{E}^{\text{strong}}(\epsilon)$ defined in Equation (12) holds).*

Proof. Applying Lemma 8.8 to $f = (u, \tilde{u})$ shows that for the edge $f \in \mathcal{E}_T \setminus \mathcal{E}_{T^{\text{cl}}}$, there exists an edge $g = (v, \tilde{v}) \in \mathcal{E}_{T^{\text{cl}}} \setminus \mathcal{E}_T$ such that, $f \in \text{path}_T(v, \tilde{v})$ and $g \in \text{path}_{T^{\text{cl}}}(u, \tilde{u})$ (Figure 6.1). Let $Z = Z_{f,v,\tilde{v}}$ defined in 21 and $Y = Y_{f,v,\tilde{v}}$ defined in 22 in the scope of this proof. Applying Lemma 8.2, this implies that $(\sum_{i=1}^n Z^{(i)}) (\sum_{i=1}^n Y^{(i)}) \leq 0$. Note that $\mathbb{E}Z = \mu_f(1 - \mu_{\mathcal{A}})$ and $\mathbb{E}Y = \mu_f(1 + \mu_{\mathcal{A}})$ using the definition $\mathcal{A} = \mathcal{A}_{f,v,\tilde{v}} = \text{path}_T(v, \tilde{v}) \setminus \{f\}$. Hence $(\mathbb{E}Z)(\mathbb{E}Y) = \mu_f^2(1 - \mu_{\mathcal{A}}^2) \geq 0$. Thus, $(\sum_{i=1}^n Z^{(i)}) (\sum_{i=1}^n Y^{(i)}) < 0$ holds only if either one of the following inequalities holds:

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

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

$$\begin{aligned} |\mu_f(1 - \mu_{\mathcal{A}})| &\leq \max \left\{ 4\epsilon^2, 4\epsilon\sqrt{1 - \mu_{\mathcal{A}}} \right\} \quad \text{or} \\ |\mu_f(1 + \mu_{\mathcal{A}})| &\leq \max \left\{ 4\epsilon^2, 4\epsilon\sqrt{1 + \mu_{\mathcal{A}}} \right\}, \end{aligned}$$

which is true if

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

where $\tau(\epsilon)$ is defined in (11). Note that if $\tau(\epsilon) \geq 1$ then the bound on $|\mu_f| \leq 1 \leq \tau(\epsilon)$ is trivial. If $\tau(\epsilon) < 1$, then $\tau^2(\epsilon) < \tau(\epsilon)$ which gives $|\mu_f| \leq \tau(\epsilon)$. \square

Lemma 8.5. *With $\epsilon = \sqrt{2/n \log(2p^2/\delta)}$, event $\mathbf{E}^{\text{strong}}(\epsilon)$ defined in (12) occurs with probability at least $1 - \delta$.*

Proof. Lemma 8.4 shows that, under the events described in Lemma 8.3, the event $\mathbf{E}^{\text{strong}}(\epsilon)$ defined in Equation (12) holds. Using Lemma 8.3, with probability at least $1 - \delta$, all edges $e \in T$ such that $|\mu_e| > \tau(\epsilon)$ are recovered by Chow-Liu algorithm $e \in T^{\text{cl}}$. \square

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

This follows from the factorization of distribution $P(x) \in \mathcal{P}(T)$ in (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 $\mathbf{E}^{\text{cascade}}(\gamma)$ defined in (13) 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, \tilde{w} be a pair of nodes such that $\text{path}_{T_1}(w, \tilde{w}) \neq \text{path}_{T_2}(w, \tilde{w})$. Then there exists a pair of edges $f \triangleq (u, \tilde{u}) \in \text{path}_{T_1}(w, \tilde{w})$ and $g \triangleq (v, \tilde{v}) \in \text{path}_{T_2}(w, \tilde{w})$ such that*

(i) $f \notin \text{path}_{T_2}(w, \tilde{w})$ and $g \notin \text{path}_{T_1}(w, \tilde{w})$

(ii) $f \in \text{path}_{T_1}(v, \tilde{v})$ and $g \in \text{path}_{T_2}(u, \tilde{u})$

Since $f \in \text{path}_{T_1}(w, \tilde{w}) \cap \text{path}_{T_1}(v, \tilde{v})$, w and \tilde{w} (and respectively v and \tilde{v}) are in different subtrees of T_1 after removing edge f , one can label the end points of the edges $f = (u, \tilde{u})$ and $g = (v, \tilde{v})$ such that $u, v \in \text{SubTree}_{T_1, f}(w)$ and $\tilde{u}, \tilde{v} \in \text{SubTree}_{T_1, f}(\tilde{w})$ (Figure 6.1). 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 (5)) and the $\mathcal{L}^{(2)}$ loss defined in (7). 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) 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.

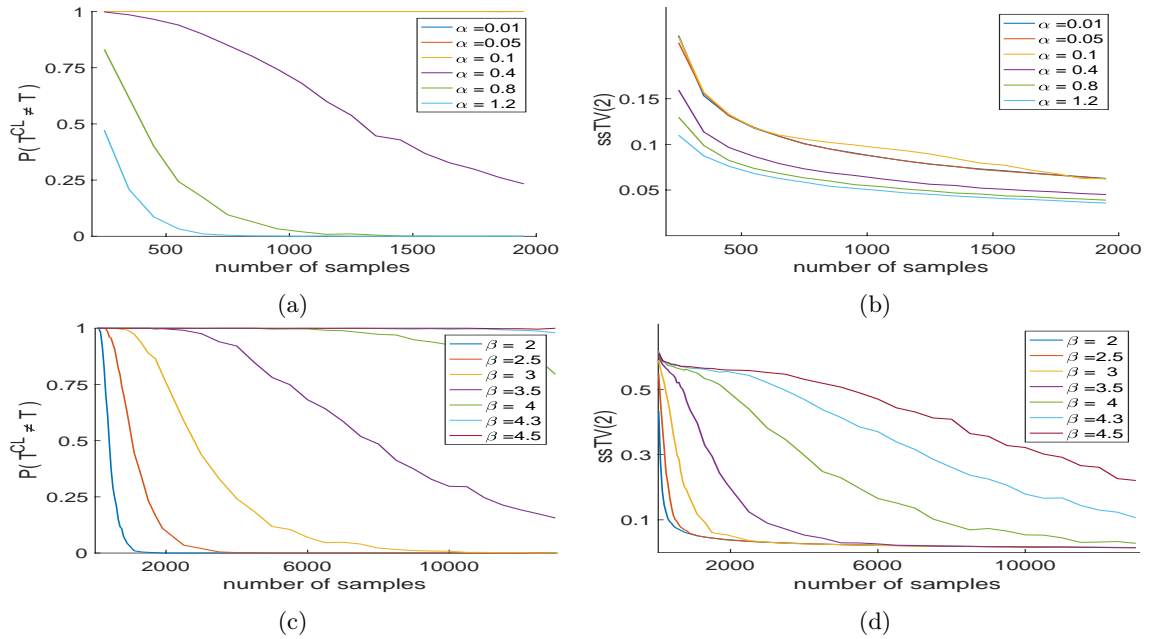


Figure 9.1: The performance of the Chow-Liu algorithm as a function of the number of samples for generative distribution factorized according to (1) with $p = 31$. Figures (a) and (b) use $\beta = 2$ and different values of α . Figures (c) and (d) use $\alpha = .8$ and different values of β . Figures (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.

In Figure 9.1a, we plot $P[T^{\text{CL}} \neq T]$ 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 9.1b 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 9.1c and 9.1d plot the probability of error and the $\mathcal{L}^{(2)}$ loss as a function of n with $p = 31$, $\alpha = .8$, and different values of β . As β increases, Figure 9.1c 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 9.1d 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.

Acknowledgment

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.

This work was supported in part by the grants ONR N00014-17-1-2147, DARPA W911NF-16-1-0551, and NSF CCF-1565516.

References

- [1] ABBEEL, P., KOLLER, D., AND NG, A. Learning factor graphs in polynomial time and sample complexity. *The Journal of Machine Learning Research* 7, Aug (2006), 1743–1788.
- [2] AGARWALA, R., BAFNA, V., FARACH, M., PATERSON, M., AND THORUP, M. On the approximability of numerical taxonomy (fitting distances by tree metrics). *SIAM Journal on Computing* 28, 3 (1998), 1073–1085.

- [3] AMBAINIS, A., DESPER, R., FARACH, M., AND KANNAN, S. Nearly tight bounds on the learnability of evolution. In *Proceedings of the 38th Annual Symposium on Foundations of Computer Science (FOCS)*, (1997), IEEE, pp. 524–533.
- [4] ANANDKUMAR, A., HUANG, F., HSU, D., AND KAKADE, S. Learning mixtures of tree graphical models. In *Advances in Neural Information Processing Systems* (2012), pp. 1052–1060.
- [5] ANANDKUMAR, A., TAN, V., HUANG, F., AND WILLSKY, A. High-dimensional structure estimation in Ising models: Local separation criterion. *The Annals of Statistics* 40, 3 (2012), 1346–1375.
- [6] ANANDKUMAR, A., AND VALLUVAN, R. Learning loopy graphical models with latent variables: Efficient methods and guarantees. *The Annals of Statistics* 41, 2 (2013), 401–435.
- [7] BAXTER, R. J. Exactly solved models in statistical mechanics. In *Integrable systems in statistical mechanics*. World Scientific, 1985, pp. 5–63.
- [8] BENNETT, G. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association* 57, 297 (1962), 33–45.
- [9] BENTO, J., AND MONTANARI, A. Which graphical models are difficult to learn? In *Advances in Neural Information Processing Systems* (2009).
- [10] BRESLER, G. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing* (2015), ACM, pp. 771–782.
- [11] BRESLER, G., MOSSEL, E., AND SLY, A. Reconstruction of Markov random fields from samples: Some observations and algorithms. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques* (2008), Springer, pp. 343–356.
- [12] CHOW, C., AND LIU, C. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information theory* 14, 3 (1968), 462–467.
- [13] CHOW, C., AND WAGNER, T. Consistency of an estimate of tree-dependent probability distributions. *IEEE Transactions on Information Theory*, 19, 3 (1973), 369–371.
- [14] CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L., AND STEIN, C. *Introduction to Algorithms*. McGraw-Hill, 2009.
- [15] CSISZÁR, I., AND SHIELDS, P. C. *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.
- [16] DASGUPTA, S. Learning polytrees. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (1999), pp. 134–141.
- [17] DASKALAKIS, C., MOSSEL, E., AND ROCH, S. Phylogenies without branch bounds: Contracting the short, pruning the deep. In *Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology* (2009), Springer, pp. 451 – 465.
- [18] ERDŐS, P. L., STEEL, M. A., SZÉKELY, L., AND WARNOW, T. J. A few logs suffice to build (almost) all trees: Part ii. *Theoretical Computer Science* 221, 1 (1999), 77–118.
- [19] FREEMAN, W. T., PASZTOR, E. C., AND CARMICHAEL, O. T. Learning low-level vision. *International journal of computer vision* 40, 1 (2000), 25–47.

- [20] GALLAGER, R. Low-density parity-check codes. *IRE Transactions on information theory* 8, 1 (1962), 21–28.
- [21] GEORGH, H. O. *Gibbs measures and phase transitions*, vol. 9. Walter de Gruyter, 2011.
- [22] HEINEMANN, U., AND GLOBERSON, A. Infering with high girth graphical models. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (2014), pp. 1260–1268.
- [23] Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association* 58, 301 (1963), 13–30.
- [24] JOG, V., AND LOH, P.-L. On model misspecification and KL separation for Gaussian graphical models. In *IEEE International Symposium on Information Theory (ISIT)*, (2015), pp. 1174–1178.
- [25] KOLLER, D., AND FRIEDMAN, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.
- [26] LAURITZEN, S. *Graphical models*. Oxford University Press, 1996.
- [27] LIU, H., XU, M., GU, H., GUPTA, A., LAFFERTY, J., AND WASSERMAN, L. Forest density estimation. *The Journal of Machine Learning Research* 12 (2011), 907–951.
- [28] LOH, P.-L., WAINWRIGHT, M. J., ET AL. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *The Annals of Statistics* 41, 6 (2013), 3022–3049.
- [29] MEILA, M., AND JORDAN, M. I. Learning with mixtures of trees. *The Journal of Machine Learning Research* 1 (2001), 1–48.
- [30] MEINSHAUSEN, N., AND BÜHLMANN, P. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* (2006), 1436–1462.
- [31] MOSSEL, E. Distorted metrics on trees and phylogenetic forests. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 4 (2007).
- [32] NARASIMHAN, M., AND BILMES, J. PAC-learning bounded tree-width graphical models. *Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence* (2004), 410–417.
- [33] PEARL, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.
- [34] PORTILLA, J., STRELA, V., WAINWRIGHT, M. J., AND SIMONCELLI, E. P. Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Transactions on Image processing* 12, 11 (2003), 1338–1351.
- [35] RAVIKUMAR, P., WAINWRIGHT, M., AND LAFFERTY, J. High-dimensional Ising model selection using ℓ_1 -regularized logistic regression. *The Annals of Statistics* 38, 3 (2010), 1287–1319.
- [36] REBESCHINI, P., VAN HANDEL, R., ET AL. Can local particle filters beat the curse of dimensionality? *The Annals of Applied Probability* 25, 5 (2015), 2809–2866.

- [37] ROMBERG, J. K., CHOI, H., AND BARANIUK, R. G. Bayesian tree-structured image modeling using wavelet-domain hidden markov models. *IEEE Transactions on image processing* 10, 7 (2001), 1056–1068.
- [38] SANTHANAM, N. P., AND WAINWRIGHT, M. J. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information theory* 58, 7 (2012), 4117–4134.
- [39] SREBRO, N. Maximum likelihood bounded tree-width Markov networks. In *The Conference on Uncertainty in Artificial Intelligence (UAI)* (2001).
- [40] TAN, V. Y., ANANDKUMAR, A., TONG, L., AND WILLSKY, A. S. A large-deviation analysis of the maximum-likelihood learning of Markov tree structures. *IEEE Transactions on Information Theory*, 57, 3 (2011), 1714–1735.
- [41] TAN, V. Y., ANANDKUMAR, A., AND WILLSKY, A. S. Learning high-dimensional Markov forest distributions: Analysis of error rates. *The Journal of Machine Learning Research* 12 (2011), 1617–1653.
- [42] TANDON, R., SHANMUGAM, K., RAVIKUMAR, P. K., AND DIMAKIS, A. G. On the information theoretic limits of learning Ising models. In *Advances in Neural Information Processing Systems* (2014), pp. 2303–2311.
- [43] TSYBAKOV, A. B. *Introduction to nonparametric estimation*. Springer, 2008.
- [44] VUFFRAY, M., MISRA, S., LOKHOV, A., AND CHERTKOV, M. Interaction screening: Efficient and sample-optimal learning of ising models. In *Advances in Neural Information Processing Systems* (2016), pp. 2595–2603.
- [45] WAINWRIGHT, M. Estimating the “wrong” graphical model: Benefits in the computation-limited setting. *The Journal of Machine Learning Research* 7 (2006), 1829–1857.
- [46] WAINWRIGHT, M., AND JORDAN, M. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* 1, 1-2 (2008), 1–305.
- [47] WAINWRIGHT, M. J. Tree-reweighted belief propagation algorithms and approximate ML estimation via pseudo-moment matching. *IN AISTATS* (2003).
- [48] WAINWRIGHT, M. J., SIMONCELLI, E. P., AND WILLSKY, A. S. Random cascades on wavelet trees and their use in analyzing and modeling natural images. *Applied and Computational Harmonic Analysis* 11, 1 (2001), 89–123.
- [49] WU, R., SRIKANT, R., AND NI, J. Learning loosely connected Markov random fields. *Stochastic Systems* 3, 2 (2013), 362–404.

A Information projection lemmas

In the following, we prove two lemmas regarding the maximum likelihood tree.

Lemma A.1. For a given tree $\mathsf{T} = (\mathcal{V}, \mathcal{E})$, the reverse information-projection of the distribution $P(x)$ onto the class of the Ising models on T with no external field (as in Equation (1)), $\tilde{P}(x) = \Pi_{\mathsf{T}}(P) \triangleq \arg \min_{Q \in \mathcal{P}_{\mathsf{T}}} D(P\|Q)$ with the parameter $\tilde{\theta} \in \Omega_{0,\infty}(\mathsf{T})$ has the following property: $\tanh(\tilde{\theta}_{ij}) = \mu_{ij}$ for all $(i, j) \in \mathcal{E}$, where $\mu_{ij} = \mathbb{E}_P X_i X_j$ is the pairwise correlation of the variables under the distribution $P(x)$. Also, we have $D(P\|\tilde{P}) = -H(P) + \sum_{(i,j) \in \mathcal{E}_{\mathsf{T}}} H_B(\frac{1+\mu_{ij}}{2})$ where $H_B(x) = x \log \frac{1}{x} + (1-x) \log \frac{1}{1-x}$ is the binary entropy function.

Proof. This lemma is a direct corollary of Theorem 3.3 in [15]. Note that $\tilde{P}(x)$ is the reverse I-projection of distribution $P(x)$ onto the exponential family with sufficient statistics $X_i X_j$ for $(i, j) \in \mathcal{E}$. It follows that $\mathbb{E}_{\tilde{P}} X_i X_j = \mathbb{E}_P X_i X_j$ for $(i, j) \in \mathcal{E}$ which gives $\tanh(\tilde{\theta}_{ij}) = \mu_{ij}$. Section 3.4.2 in [46] also addresses this problem in depth.

We provide the proof for the sake of completeness here: For the tree $\mathsf{T} = (\mathcal{V}, \mathcal{E})$ and any $Q \in \mathcal{P}_{\mathsf{T}}$ with parameter vector θ , using (1), Q factorizes as $Q(x) = \prod_{(i,j) \in \mathcal{E}} (1 + \tanh \theta_{ij} x_i x_j)/2$, where $\mathbb{E}_Q X_i X_j = \tanh \theta_{ij}$. Let $Y_e = X_i X_j$ for all $e = (i, j) \in \mathcal{E}$ so that $Q(Y_e = +) = (1 + \tanh \theta_{ij})/2$ and $P(Y_e = +) = (1 + \mu_{ij})/2$ where $\mu_{ij} = \mathbb{E}_P X_i X_j$.

Given probability mass function P on $\{-1, +1\}^p$ and any $Q \in \mathcal{P}_{\mathsf{T}}$ with parameters θ ,

$$\begin{aligned} D(P\|Q) &= \mathbb{E}_P \left[\log \frac{P}{Q} \right] = -H(P) - \mathbb{E}_P [\log Q(X)] \\ &= -H(P) - \mathbb{E}_P \left[\log \prod_{(i,j) \in \mathcal{E}} \frac{1 + \tanh \theta_{ij} X_i X_j}{2} \right] \\ &= -H(P) - \sum_{e=(i,j) \in \mathcal{E}} \mathbb{E}_P \left[\log \frac{1 + \tanh \theta_{ij} Y_e}{2} \right] \\ &= -H(P) - \sum_{e \in \mathcal{E}} \mathbb{E}_P [\log Q(Y_e)] \\ &= -H(P) + \sum_{e \in \mathcal{E}} H_B(P(Y_e)) + \sum_{e \in \mathcal{E}} D(P(Y_e)\|Q(Y_e)). \end{aligned}$$

$\tilde{P} = \arg \min_{Q \in \mathcal{P}_{\mathsf{T}}} D(P\|Q)$ with parameters $\tilde{\theta}$ satisfies the property $\tilde{P}(Y_e) = P(Y_e)$ for all $e \in \mathcal{E}$, hence $\tanh(\tilde{\theta}_{ij}) = \mu_{ij}$ for all $(i, j) \in \mathcal{E}$. Consequently, $D(P\|\tilde{P}) = -H(P) + \sum_{e \in \mathcal{E}} H_B(\frac{1+\mu_{ij}}{2})$. \square

Lemma A.2. Given empirical distribution \hat{P} , the tree T^{CL} defined in Definition 2.1 can be found as the maximum weight spanning tree over a complete weighted graph where the weights of each edge (i, j) is $|\hat{\mu}_{ij}|$.

Proof. Using Lemma A.1, $\mathsf{T}^{\text{CL}} = \arg \min_{\mathsf{T} \in \mathcal{T}} \sum_{e \in \mathcal{E}_{\mathsf{T}}} H_B(\frac{1+\hat{\mu}_e}{2}) - H(\hat{P}) = \arg \min_{\mathsf{T} \in \mathcal{T}} \sum_{e \in \mathcal{E}_{\mathsf{T}}} H_B(\frac{1+\hat{\mu}_e}{2})$. The maximum weight spanning tree can be implemented greedily using Kruskal algorithm or Prim's algorithm [14]. So, finding the maximum weight spanning tree only depends on the sorted order of the edges of the graph.

$H_B(\frac{1+\hat{\mu}_{ij}}{2})$ is a monotonically increasing function of $|\hat{\mu}_{ij}|$. So, sorting all the edges (i, j) in the complete graph based $H_B(\frac{1+\hat{\mu}_{ij}}{2})$ or $|\hat{\mu}_{ij}|$ gives the same order. This gives

$$\mathsf{T}^{\text{CL}} = \arg \min_{\mathsf{T} \in \mathcal{T}} \sum_{(i,j) \in \mathcal{E}_{\mathsf{T}}} H_B\left(\frac{1+\hat{\mu}_{ij}}{2}\right) = \arg \min_{\mathsf{T} \in \mathcal{T}} \sum_{(i,j) \in \mathcal{E}_{\mathsf{T}}} |\hat{\mu}_{ij}|. \quad \square$$

B Samples Necessary for small SSTV

In this theorem, we prove a lower bound for learning a tree-structured Ising model with the ssTV loss function $\mathcal{L}^{(2)}$. Unlike the bound in Theorem 3.4, the following lower bound does not require $\tanh(\beta) > \tanh(\alpha) + 2\eta$. This bound can be used specifically in the extreme case of $\alpha = \beta$.

Theorem B.1. Fix $\eta > 0$. Let $\mu = \tanh \alpha$ and

$$d = \min \left\{ \left\lceil \frac{\log(1 - 2\eta/\mu)}{\log \mu} \right\rceil, p - 1 \right\}. \quad (25)$$

Suppose one observes

$$n < \frac{\log(p - d)}{4\alpha \tanh(\alpha)[1 - \tanh(\alpha)^d]}$$

samples. Then, 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, i.e.,

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

Proof. We construct a family of $M = p - d$ models. Each of the models correspond to a tree-structured model in which the edge parameters on the edges of the tree are taking a fixed value $\theta_e = \alpha$. Let $\theta_{i,i+1}^0 = \alpha$ for $i = 1, \dots, p - 1$. For $m = 1, \dots, p - 1 - d$, in the m -th model, we have $\theta_{m,m+d+1} = \alpha$ and $\theta_{i,i+1} = \alpha$ for $i = 1, \dots, m - 1, m + 1, \dots, p - 1$. Hence, there are $p - d$ models. For $m' \neq m$,

$$\max_{i,j} |\mathbb{E}_{\theta^{m'}}[X_i X_j] - \mathbb{E}_{\theta^m}[X_i X_j]| \geq \tanh(\alpha)[1 - \tanh(\alpha)^d] \geq 2\eta.$$

Also, using (18)

$$J(\theta^m \|\theta^{m'}) \leq 2\alpha \tanh(\alpha)[1 - \tanh(\alpha)^d].$$

Fano's inequality (Corollary 6.3) gives the bound. \square

The following corollary looks into the asymptotic value of the above lower bound when $p \rightarrow \infty$ (p is large).

Corollary B.2. Fix $\eta > 0$. If there exists a constant $0 < c < 1$ such that

$$\tanh(\alpha)[1 - \tanh(\alpha)^{(p^c)}] \geq 2\eta$$

(or equivalently, $d < p - \max\{1, p^c\}$ using the definition in Equation (B.1)), and one observes

$$n < \frac{c}{16} \min \left\{ \frac{1}{\alpha\eta}, \frac{e^{2\alpha}}{\alpha \tanh(\alpha)} \right\} \log p$$

samples, then

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

This corollary is a direct consequence of Theorem B.1. In particular, in the asymptotic regime of interest in which $\alpha = \beta$ (we don't assume any gap between upper and lower bounds of edge parameters), $p \rightarrow \infty$ and $\eta \rightarrow 0$, the above lower bound entails that if $n < Ce^{2\alpha}/[\alpha \tanh(\alpha)] \log p$, then the $\mathcal{L}^{(2)} > \eta$ with probability greater than $1/2$. Interestingly, this lower bound is independent of η . This lower bound is also within multiplicative constant of lower and upper bounds of correct recovery of structure (zero-one loss function) presented in Theorems 3.1 and 3.2 (on the assumption $\alpha = \beta$). To get the matching upper bound in this regime, one could simply run Chow-Liu algorithm to recover \mathcal{T}^{CL} (which is guaranteed to be equal to \mathcal{T} with probability greater than $1 - \delta$ if $n > C'e^{2\alpha} \tanh^{-2}(\alpha) \log(p/\delta)$) and choose $\tilde{\theta}_{ij} = \alpha$ for all $(i, j) \in \mathcal{E}_{\mathcal{T}^{\text{CL}}}$.

C Proof of Sample Complexity Upper Bound of Forest Approximation Algorithm

Proof of Proposition 4.1. Let $\epsilon_1 = \eta e^{-\beta}/8$ and $\gamma_1 = \eta/4$. Using Lemmas 8.1, 8.5 and 8.7, given $n > \max\{128e^{2\beta}/\eta^2 \log \frac{6p^2}{\delta}, 400/\eta^2 \log \frac{12p^2}{\delta}\}$ samples, the event $\mathbf{E}(\epsilon_1, \gamma_1)$ occurs with probability greater than $1 - \delta$. We will show that on the event $\mathbf{E}(\epsilon_1, \gamma_1)$, the statement of the proposition holds. The proof has two steps. In Step 1, we prove that the output of the forest approximation algorithm is a subgraph of the original tree, *i.e.*, $\mathcal{E}_{\hat{\mathcal{T}}} \subseteq \mathcal{E}_{\mathcal{T}}$. In Step 2, we show that the ssTV bound holds.

Step 1 To prove $\mathcal{E}_{\hat{\mathcal{T}}} \subseteq \mathcal{E}_{\mathcal{T}}$, we will show that on the event $\mathbf{E}(\epsilon_1, \gamma_1)$, all the edges in $(i, j) \in \mathcal{E}_{\mathcal{T}^{\text{CL}}}$ with $|\hat{\mu}_{ij}| > \tau(\epsilon_1) + \epsilon_1$ (the edges which survive the truncation algorithm and construct the approximated forest) belong to the original tree, *i.e.*, $(i, j) \in \mathcal{E}_{\mathcal{T}}$.

We prove that if $(i, j) \in \mathcal{E}_{\mathcal{T}^{\text{CL}}}$ and $|\hat{\mu}_{ij}| > \tau(\epsilon_1) + \epsilon_1$, then $(i, j) \in \mathcal{E}_{\mathcal{T}}$. On the event $\mathbf{E}^{\text{corr}}(\epsilon_1)$, defined in (10), $|\hat{\mu}_{ij}| > \tau(\epsilon_1) + \epsilon_1$ implies $|\mu_{ij}| > \tau(\epsilon_1)$. We prove $(i, j) \in \mathcal{E}_{\mathcal{T}}$ by contradiction: Let us assume $(i, j) \notin \mathcal{E}_{\mathcal{T}}$. Then $\text{path}_{\mathcal{T}}(i, j)$ has at least two edges in it. Also, $\tau(\epsilon_1) < |\mu_{ij}| = \prod_{e \in \text{path}_{\mathcal{T}}(i, j)} |\mu_e|$. Then, for all $e \in \text{path}_{\mathcal{T}}(i, j)$, we have $\tau(\epsilon_1) < |\mu_e|$ which implies $e \in \mathcal{E}_{\mathcal{T}}^{\text{strong}}(\epsilon_1)$ as defined in (11). Hence, on the event $\mathbf{E}^{\text{strong}}(\epsilon_1)$, defined in (12) $e \in \mathcal{E}_{\mathcal{T}^{\text{CL}}}$ for all $e \in \text{path}_{\mathcal{T}}(i, j)$. This constructs a path of length at least two between i and j in \mathcal{T}^{CL} , which contradicts $(i, j) \in \mathcal{E}_{\mathcal{T}^{\text{CL}}}$. It follows that $\mathcal{E}_{\hat{\mathcal{T}}} \subseteq \mathcal{E}_{\mathcal{T}}$.

Step 2 Next, we consider the edges $e \in \mathcal{E}_{\mathcal{T}}$, with $|\mu_e| > \tau(\epsilon_1) + 2\epsilon_1$. On event $\mathbf{E}^{\text{strong}}(\epsilon_1)$, these edges are recovered by the Chow-Liu algorithm. On event $\mathbf{E}^{\text{corr}}(\epsilon_1)$, the empirical correlation on these edges are bounded as $|\mu_e| > \tau(\epsilon_1) + \epsilon_1$. Hence, the forest approximation algorithm does not remove these edges. This implies that all the edges in the original tree \mathcal{T} with $|\mu_e| > \tau(\epsilon_1) + 2\epsilon_1$ are in the output of the forest approximation algorithm, $\hat{\mathcal{T}}$.

To bound the ssTV, we use the decomposition into two terms as per (9). We first study the loss due to graph estimation error $\mathcal{L}^{(2)}(P, \Pi_{\hat{\mathcal{T}}}(\mathbf{P}))$. For any pair of nodes w and \tilde{w} in $\hat{\mathcal{T}}$, since $\mathcal{E}_{\hat{\mathcal{T}}} \subseteq \mathcal{E}_{\mathcal{T}}$, if there is a path between w and \tilde{w} , then this is the same as the path between them in \mathcal{T} . Hence looking at (7), $\mathbb{E}_P[X_w X_{\tilde{w}}] = \mathbb{E}_{\Pi_{\hat{\mathcal{T}}}(\mathbf{P})}[X_w X_{\tilde{w}}]$ and the error is zero. If there is no path between w and \tilde{w} in $\hat{\mathcal{T}}$, then $\mathbb{E}_{\Pi_{\hat{\mathcal{T}}}(\mathbf{P})}[X_w X_{\tilde{w}}] = 0$ and the error is $\mathbb{E}_P[X_w X_{\tilde{w}}]$. This occurs whenever there is an edge $e \in \text{path}_{\mathcal{T}}(w, \tilde{w})$ such that $e \notin \mathcal{E}_{\hat{\mathcal{T}}}$. All the missing edges in $\hat{\mathcal{T}}$ satisfy $|\mu_e| \leq \tau(\epsilon_1) + 2\epsilon_1$. This implies that $|\mathbb{E}_P[X_w X_{\tilde{w}}]| \leq \tau(\epsilon_1) + 2\epsilon_1$ and it follows that $\mathcal{L}^{(2)}(P, \Pi_{\hat{\mathcal{T}}}(\mathbf{P})) \leq \tau(\epsilon_1) + 2\epsilon_1$.

To bound the loss due to parameter estimation error $\mathcal{L}^{(2)}(\Pi_{\hat{\mathcal{T}}}(\mathbf{P}), \Pi_{\hat{\mathcal{T}}}(\hat{\mathbf{P}}))$, we again consider an arbitrary pair of nodes w, \tilde{w} . If there is a path between them in $\hat{\mathcal{T}}$, then there is the same path between them in \mathcal{T} since $\mathcal{E}_{\hat{\mathcal{T}}} \subseteq \mathcal{E}_{\mathcal{T}}$. On the event $\mathbf{E}^{\text{cascade}}(\gamma_1)$, for such a pair of nodes, $|\mathbb{E}_{\Pi_{\hat{\mathcal{T}}}(\mathbf{P})}[X_w X_{\tilde{w}}] -$

$\mathbb{E}_{\Pi_{\hat{T}}(\hat{P})}[X_w X_{\tilde{w}}] \leq \gamma_1$. If there is no path between w and \tilde{w} in \hat{T} , then $\mathbb{E}_{\Pi_{\hat{T}}(\hat{P})}[X_w X_{\tilde{w}}] = \mathbb{E}_{\Pi_{\hat{T}}(\hat{P})}[X_w X_{\tilde{w}}] = 0$. Hence, $\mathcal{L}^{(2)}(\Pi_{\hat{T}}(\hat{P}), \Pi_{\hat{T}}(\hat{P})) \leq \gamma_1$.

Combining the two error terms gives $\mathcal{L}^{(2)}(P, \Pi_{\hat{T}}(\hat{P})) \leq \tau(\epsilon_1) + 2\epsilon_1 + \gamma_1$. With the choice of $\epsilon_1 = \eta e^{-\beta}/8$ and $\gamma_1 = \eta/4$ then $\tau(\epsilon_1) = 4\epsilon_1/\sqrt{1 - \tanh \beta} \leq 4\eta e^{\beta} \leq \eta/8$ and $\mathcal{L}^{(2)}(P, \Pi_{\hat{T}}(\hat{P})) \leq \tau(\epsilon_1) + 2\epsilon_1 + \gamma_1 \leq \eta$. \square

D Tail bounds

We state a standard form of Hoeffding's inequality [23]. This bound is used in the proof of Lemma refl:EcorrLemma to show that event $\mathbf{E}^{\text{corr}}(\epsilon)$ defined in (10) occurs with probability at least $1 - 2p^2 \exp(-n\epsilon^2/2)$.

Lemma D.1 (Hoeffding's inequality). *Let $Z^{(1)}, \dots, Z^{(n)}$ be n i.i.d. random variables taking values in the interval $[-1, 1]$ and let $S_n = \frac{1}{n}(Z^{(1)} + \dots + Z^{(n)})$. Then, $\mathbb{P}[|S_n - \mathbb{E}Z^{(1)}| > t] \leq 2 \exp(-nt^2/2)$.*

The next lemma is the standard Bernstein's inequality as quoted from [43]. This bound is used in the proof of Lemma 8.5 to show that the event $\mathbf{E}^{\text{strong}}(\epsilon)$ defined in (12) with $\epsilon = \sqrt{2/n \log(2p^2/\delta)}$ occurs with probability at least $1 - \delta$.

Lemma D.2 (Bernstein's inequality). *Let $Z^{(1)}, \dots, Z^{(n)}$ be i.i.d. random variables. Suppose that for all i , $|Z^{(i)} - \mathbb{E}[Z^{(i)}]| \leq M$ almost surely. Then, for all positive t ,*

$$\mathbb{P}\left[\left|\sum_{i=1}^n Z^{(i)} - n\mathbb{E}[Z^{(1)}]\right| \geq t\right] \leq 2 \exp\left(-\frac{t^2}{2n\text{VAR}(Z) + \frac{2}{3}Mt}\right).$$

The following lemma is essentially a restatement of Theorem 3 of Hoeffding [23]. This bound is used in the proof of Lemma 8.7 in Appendix E to show that if $n > \max\{25/\gamma^2 \log(4p^2/\delta), 108e^{2\beta} \log(2p^3/\delta)\}$, then the event $\mathbf{E}^{\text{cascade}}(\gamma)$ defined in (13) occurs with probability at least $1 - \delta$.

Lemma D.3. *Let $Z^{(1)}, \dots, Z^{(n)}$ be i.i.d. random variables such that $\mathbb{E}[Z^{(i)}] = 0$, $|Z^{(i)}| \leq b$ and $\mathbb{E}[(Z^{(i)})^2] \leq \sigma^2$. Let $\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z^{(i)}$. Then*

$$\mathbb{E}[\exp(\lambda \bar{Z})] \leq \exp\left\{n f\left(\frac{\sigma^2}{b^2}, \frac{b|\lambda|}{n}\right)\right\}, \quad (26)$$

where we defined

$$f(u, c) = \log\left(\frac{1}{1+u}e^{-cu} + \frac{u}{1+u}e^c\right). \quad (27)$$

Also, if $c \geq 0$, the function $f(u, c)$ is concave and nondecreasing in u for $u \geq 0$. And for any $t \geq 0$,

$$\mathbb{P}[|\bar{Z}| > t] \leq 2 \left\{ \left(1 + \frac{bt}{\sigma^2}\right)^{-(1+bt/\sigma^2)\frac{\sigma^2}{b^2+\sigma^2}} \left(1 - \frac{t}{b}\right)^{-(1-t/b)\frac{b^2}{b^2+\sigma^2}} \right\}^n. \quad (28)$$

Proof. Using Equation (11) in Bennett [8], for any λ ,

$$\mathbb{E}[\exp(\lambda Z_i)] \leq \frac{b^2}{b^2 + \sigma^2} \exp\left(-\frac{\sigma^2}{b}|\lambda|\right) + \frac{\sigma^2}{b^2 + \sigma^2} \exp(b|\lambda|) = \exp\left\{f\left(\frac{\sigma^2}{b^2}, b|\lambda|\right)\right\}.$$

Bennett [8] proves this for $\lambda > 0$ with the assumption $Z_i < b$, but the above modification is proved similarly using $|Z_i| < b$. Hence,

$$\mathbb{E} \left[\exp(\lambda \bar{Z}) \right] = \prod_{i=1}^n \mathbb{E} \left[\exp \left(\frac{\lambda}{n} Z_i \right) \right] \leq \exp \left\{ n f \left(\frac{\sigma^2}{b^2}, b \frac{|\lambda|}{n} \right) \right\}.$$

Lemma 3 in [23] shows that $\frac{d^2}{du^2} f(u, c) \leq 0$ for $u \geq 0$ and $c \geq 0$. Following the notation in the same proof, to show that $f(u, c)$ is increasing in u , we write $f(u, c) = c + \log f_1(y, c)$, where $y = 1 + u$ and $f_1(y, c) = (\exp(-cy) - 1 + y)/y$. So, for any $c > 0$,

$$\frac{d}{dy} f_1(y, c) = [1 - (1 + cy)e^{-cy}]y^{-2} \geq 0$$

since $e^{cy} \geq 1 + cy$. This shows that $f(u, c)$ is nondecreasing in u for $u \geq 0$.

Using a Chernoff bound,

$$\mathbb{P}[\bar{Z} > t] \leq \min_{\lambda > 0} \exp \left\{ n \left[f \left(\frac{\sigma^2}{b^2}, b \frac{|\lambda|}{n} \right) - \lambda \frac{t}{n} \right] \right\}.$$

Plugging $\lambda = \lambda^*$ with

$$\lambda^* = \frac{nb}{b^2 + \sigma^2} \log \frac{1 + \frac{tb}{\sigma^2}}{1 - \frac{t}{b}}.$$

with change of variables $u = \sigma^2/b^2$ and $s = tb/\sigma^2$ gives

$$\begin{aligned} \mathbb{P}[\bar{Z} > t] &\leq \left[\frac{1}{1+u} \exp \left(\frac{-u}{1+u} \log \frac{1+s}{1-su} \right) + \frac{u}{1+u} \exp \left(\frac{1}{1+u} \log \frac{1+s}{1-su} \right) \right]^n \\ &\quad \exp \left(-\frac{sun}{u+1} \log \frac{1+s}{1-su} \right) \\ &= (1+s)^{-(1+s)\frac{un}{u+1}} (1-su)^{-(1-su)\frac{n}{u+1}} \left[\frac{1-su}{1+u} + \frac{u(1+s)}{u+1} \right]^n \\ &= (1+s)^{-(1+s)\frac{un}{u+1}} (1-su)^{-(1-su)\frac{n}{u+1}}. \end{aligned}$$

A similar Chernoff bounds with $\lambda = -\lambda^*$ upper bounds $\mathbb{P}[\bar{Z} < -t]$. Combining these two bounds gives (28). \square

E End-to-end error on paths in the original tree (Proof of Lemma 8.10)

For each edge $e \in \mathcal{E}_T$, let

$$\gamma_e = \sqrt{3 \frac{1 - \mu_e^2}{n} \log(2p/\delta)} \quad (29)$$

and

$$\mathbf{E}_e^{\text{edge}} = \left\{ |\hat{\mu}_e - \mu_e| \leq \gamma_e \right\} \quad (30)$$

be the event that the empirical correlation on edge e is within γ_e of the population value. For $\mathcal{A} \subseteq \mathcal{E}_T$, let $\mathbf{E}^{\text{edge}}(\mathcal{A}) = \cap_{e \in \mathcal{A}} \mathbf{E}_e^{\text{edge}}$ (This event will only be used for the purpose of proof of Lemma 8.7).

Now, if $n > 108e^{2\beta} \log(2p/\delta)$, then

$$\gamma_e < \sqrt{\frac{3(1-\mu_e^2)}{108e^{2\beta} \log(2p/\delta)} \log(2p/\delta)} \leq \frac{1}{6}(1-\mu_e^2), \quad (31)$$

where the last inequality is due to $e^{-2\beta} \leq 1 - \tanh(\beta) \leq 1 - |\mu_e| \leq 1 - \mu_e^2$.

Lemma E.1. *If $n > 108e^{2\beta} \log(2p/\delta)$, then $\mathbb{P}[(\mathbf{E}^{\text{edge}}(\mathcal{E}_T))^c] \leq \delta$.*

Proof. For each edge e , the variance of $\hat{\mu}_e$ is $(1 - \mu_e^2)/n$. By Bernstein's inequality (Lemma D.2) and the bound (31) on γ_e ,

$$\mathbb{P}[(\mathbf{E}_e^{\text{edge}})^c] \leq 2 \exp\left(-\frac{n\gamma_e^2}{2(1-\mu_e^2) + 4\gamma_e/3}\right) \leq 2 \exp\left(-\frac{n\gamma_e^2}{3(1-\mu_e^2)}\right)$$

The definition of γ_e in (29) together with (31) and a union bound over p edges prove the lemma. \square

To prove Lemma 8.7, an upper bound on the moment generating function of a random variable with tight dependency on the variance is required. This is given in the Lemma D.3 in Appendix D which is essentially a restatement of Theorem 3 of Hoeffding [23].

Proof of Lemma 8.7. For given pair of nodes w, \tilde{w} , let $\text{path}_T(w, \tilde{w}) = \{e_1, e_2, \dots, e_d\}$ and use the shorthand $\mu_i = \mu_{e_i}$ and $\hat{\mu}_i = \hat{\mu}_{e_i}$. We will show that if $n > \max\{25/\gamma^2 \log(4p^2/\delta), 108e^{2\beta} \log(2p^3/\delta)\}$, then

$$\mathbb{P}\left[\left|\prod_{i=1}^d \hat{\mu}_i - \prod_{i=1}^d \mu_i\right| > \gamma\right] \leq 2\delta/p^2, \quad (32)$$

independent of d . A union bound over $\binom{p}{2} \leq p^2/2$ pairs of nodes w and \tilde{w} provides the desired lower bound for $\mathbb{P}[\mathbf{E}^{\text{cascade}}(\gamma)]$.

If $d = 1$, applying Lemma D.1 gives that if $n > 2/\gamma^2 \log(p^2/\delta)$ then (32) holds. To show (32) for $d \geq 2$, let

$$M_i = (\hat{\mu}_i - \mu_i) \prod_{j=1}^{i-1} \hat{\mu}_j \prod_{j=i+1}^d \mu_j. \quad (33)$$

Then, $\prod_{i=1}^d \hat{\mu}_i - \prod_{i=1}^d \mu_i = \sum_{i=1}^d M_i$. Let $\mathcal{A}_k = \{e_1, \dots, e_k\}$ for $k \leq d$. Conditioning on $\mathbf{E}^{\text{edge}}(\mathcal{A}_{d-1})$ (defined in (30)),

$$\mathbb{P}\left[\left|\sum_{i=1}^d M_i\right| > \gamma\right] \leq \mathbb{P}\left[\left|\sum_{i=1}^d M_i\right| > \gamma \mid \mathbf{E}^{\text{edge}}(\mathcal{A}_{d-1})\right] + \mathbb{P}\left[(\mathbf{E}^{\text{edge}}(\mathcal{A}_{d-1}))^c\right]. \quad (34)$$

The containment $\mathbf{E}^{\text{edge}}(\mathcal{E}_T) \subseteq \mathbf{E}^{\text{edge}}(\mathcal{A}_{d-1})$ gives $\mathbb{P}[(\mathbf{E}^{\text{edge}}(\mathcal{A}_{d-1}))^c] \leq \mathbb{P}[(\mathbf{E}^{\text{edge}}(\mathcal{E}_T))^c]$ and Lemma E.1 shows that if $n > 108e^{2\beta} \log(2p^3/\delta)$, then $\mathbb{P}[(\mathbf{E}^{\text{edge}}(\mathcal{E}_T))^c] \leq \delta/p^2$.

We bound the first term in the last display via bounding the moment generating function and applying Chernoff bound: Given a real number λ , the recurrence below upper bounds $\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^d M_i\right) \mid \mathbf{E}^{\text{edge}}(\mathcal{A}_{d-1})\right]$. For any $k \leq d$,

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^k M_i\right) \mid \mathbf{E}^{\text{edge}}(\mathcal{A}_{k-1})\right]$$

$$\begin{aligned}
&= \mathbb{E} \left[\mathbb{E} \left[\exp \left(\lambda \sum_{i=1}^{k-1} M_i \right) \exp(\lambda M_k) \mid \hat{\mu}_1, \dots, \hat{\mu}_{k-1} \right] \mid \mathbf{E}^{\text{edge}}(\mathcal{A}_{k-1}) \right] \\
&\stackrel{(a)}{=} \mathbb{E} \left[\exp \left(\lambda \sum_{i=1}^{k-1} M_i \right) \mathbb{E} \left[\exp(\lambda M_k) \mid \hat{\mu}_1, \dots, \hat{\mu}_{k-1} \right] \mid \mathbf{E}^{\text{edge}}(\mathcal{A}_{k-1}) \right] \\
&\stackrel{(b)}{\leq} \mathbb{E} \left[\exp \left(\lambda \sum_{i=1}^{k-1} M_i \right) \exp \left\{ n f \left(\frac{1 - \mu_k^2}{4} \prod_{j=1}^{k-1} \hat{\mu}_j^2 \prod_{j=k+1}^d \mu_j^2, 2 \frac{|\lambda|}{n} \right) \right\} \mid \mathbf{E}^{\text{edge}}(\mathcal{A}_{k-1}) \right] \\
&\stackrel{(c)}{\leq} \exp \left\{ n f \left(\frac{1 - \mu_k^2}{4} \prod_{\substack{j=1 \\ j \neq k}}^d (\mu_j^2 + 2\gamma_e), 2 \frac{|\lambda|}{n} \right) \right\} \mathbb{E} \left[\exp \left(\lambda \sum_{i=1}^{k-1} M_i \right) \mid \mathbf{E}^{\text{edge}}(\mathcal{A}_{k-1}) \right]
\end{aligned}$$

where (a) holds since for $i \leq k-1$, $M_i \in \sigma(\hat{\mu}_1, \dots, \hat{\mu}_{k-1})$. (b) is derived by applying Lemma D.3 to random variable M_k conditional on $\hat{\mu}_1, \dots, \hat{\mu}_{k-1}$. Lemma 8.6 shows that $\hat{\mu}_k$ is independent of $\hat{\mu}_1, \dots, \hat{\mu}_{k-1}$. Thus, $M_k = \frac{1}{n} \sum_{i=1}^n Z^{(i)}$ where $Z^{(i)}$'s are independent and

$$Z^{(i)} = \begin{cases} (1 - \mu_k) \prod_{j=1}^{k-1} \hat{\mu}_j \prod_{j=k+1}^d \mu_j, & \text{with probability } \frac{1 + \mu_k}{2} \\ -(1 + \mu_k) \prod_{j=1}^{k-1} \hat{\mu}_j \prod_{j=k+1}^d \mu_j, & \text{with probability } \frac{1 - \mu_k}{2}. \end{cases}$$

So, $|Z^{(i)}| \leq 2$. Conditional on $\hat{\mu}_1, \dots, \hat{\mu}_{k-1}$, the mean of $Z^{(i)}$ is zero, and the variance of $Z^{(i)}$ is $(1 - \mu_k^2) \prod_{j=1}^{k-1} \hat{\mu}_j^2 \prod_{j=k+1}^d \mu_j^2$. Inequality (c) holds since conditional on event $\mathbf{E}^{\text{edge}}(\mathcal{A}_{k-1})$ (defined in 30), $\hat{\mu}_j^2 \leq \mu_j^2 + 2\gamma_j$ for all $j \leq k-1$. Also, the function $f(u, c)$ is nondecreasing in $u, c \geq 0$ (Lemma D.3).

To get a recurrence,

$$\begin{aligned}
&\mathbb{E} \left[\exp \left(\lambda \sum_{i=1}^{k-1} M_i \right) \mid \mathbf{E}^{\text{edge}}(\mathcal{A}_{k-1}) \right] \mathbb{P}[\mathbf{E}^{\text{edge}}(\mathcal{A}_{k-1})] \\
&= \mathbb{E} \left[\exp \left(\lambda \sum_{i=1}^{k-1} M_i \right) \mid \mathbf{E}_{e_{k-1}}^{\text{edge}}, \mathbf{E}^{\text{edge}}(\mathcal{A}_{k-2}) \right] \mathbb{P}[\mathbf{E}_{e_{k-1}}^{\text{edge}} \mid \mathbf{E}^{\text{edge}}(\mathcal{A}_{k-2})] \mathbb{P}[\mathbf{E}^{\text{edge}}(\mathcal{A}_{k-2})] \\
&= \mathbb{E} \left[\exp \left(\lambda \sum_{i=1}^{k-1} M_i \right), \mathbb{1}[\mathbf{E}_{e_{k-1}}^{\text{edge}}] \mid \mathbf{E}^{\text{edge}}(\mathcal{A}_{k-2}) \right] \mathbb{P}[\mathbf{E}^{\text{edge}}(\mathcal{A}_{k-2})] \\
&\leq \mathbb{E} \left[\exp \left(\lambda \sum_{i=1}^{k-1} M_i \right) \mid \mathbf{E}^{\text{edge}}(\mathcal{A}_{k-2}) \right] \mathbb{P}[\mathbf{E}^{\text{edge}}(\mathcal{A}_{k-2})].
\end{aligned}$$

By the last two displays,

$$\begin{aligned}
&\mathbb{E} \left[\exp \left(\lambda \sum_{k=1}^d M_i \right) \mid \mathbf{E}^{\text{edge}}(\mathcal{A}_{d-1}) \right] \mathbb{P}[\mathbf{E}^{\text{edge}}(\mathcal{A}_{d-1})] \\
&\leq \exp \left\{ n \sum_{k=1}^d f \left((1 - \mu_k^2) \prod_{\substack{j=1 \\ j \neq k}}^d (\mu_j^2 + 2\gamma_e), 2 \frac{|\lambda|}{n} \right) \right\} \\
&\leq \exp \left\{ n d f \left(\frac{1}{d} \sum_{k=1}^d (1 - \mu_k^2) \prod_{\substack{j=1 \\ j \neq k}}^d (\mu_j^2 + 2\gamma_e), 2 \frac{|\lambda|}{n} \right) \right\} \tag{35}
\end{aligned}$$

where the second inequality is due to the concavity of $f(u, c)$ in u for $u, c \geq 0$ (Lemma D.3).

Now, we upper bound the first argument to f , $\sum_{k=1}^d (1 - \mu_k^2) \prod_{\substack{j=1 \\ j \neq k}}^d (\mu_j^2 + 3\gamma_e)$, for any sequence of μ_j and $\gamma_j \leq (1 - \mu_j^2)/6$. Let $x_j = \mu_j^2 + 2\gamma_j$. Since according to (31), $2\gamma_j \leq (1 - \mu_j^2)/3$, we have that $x_j < 1$ and $\frac{3}{2}(1 - x_j) = \frac{3}{2}(1 - \mu_j^2 - 2\gamma_j) \geq 1 - \mu_j^2$. Therefore,

$$\sum_{i=1}^d (1 - \mu_i^2) \prod_{\substack{j=1 \\ j \neq i}}^d (\mu_j^2 + 2\gamma_e) < \sum_{i=1}^d \frac{3}{2} (1 - x_i) \prod_{\substack{j=1 \\ j \neq i}}^d x_j \leq \frac{3}{2} dx^{d-1} (1 - x). \quad (36)$$

Symmetry of the function $\sum_{i=1}^d (1 - x_i) \prod_{j=1, j \neq i}^d x_j$ in x_j 's implies that it is maximized when $x_i = x$ for all i . Setting the partial derivatives of this term with respect to x_i equal to zero for any i shows this. This gives the second inequality in the above display.

Now, we find an upper bound for $\frac{3}{2} dx^{d-1} (1 - x)$ for any $d \geq 2$ and $x < 1$. The function $x^{d-1} (1 - x)$ takes its maximum value at $x^* = 1 - 1/d$. So,

$$\frac{3}{2} dx^{d-1} (1 - x) \leq \frac{3}{2} \left(1 - \frac{1}{d}\right)^{d-1} \leq \frac{3}{4},$$

where the last inequality holds for all $d \geq 2$. Since $f(u, c)$ is nondecreasing in u (Lemma D.3), plugging this into (35) and (36) gives

$$\mathbb{E} \left[\exp \left(\lambda \sum_{k=1}^d M_i \right) \middle| \mathbf{E}^{\text{edge}}(\mathcal{A}_{d-1}) \right] \mathbf{P} \left[\mathbf{E}^{\text{edge}}(\mathcal{A}_{d-1}) \right] \leq \exp \left\{ nd f \left(\frac{3}{4d}, 2 \frac{|\lambda|}{n} \right) \right\}.$$

With $n > 108e^{2\beta} \log(4p)$ (as assumed in the statement of lemma), $\mathbf{P} \left[\left(\mathbf{E}^{\text{edge}}(\mathcal{A}_{d-1}) \right)^c \right] \leq 1/2$ (Lemma E.1) which gives

$$\mathbb{E} \left[\exp \left(\lambda \sum_{i=1}^d M_i \right) \middle| \mathbf{E}^{\text{edge}}(\mathcal{A}_{d-1}) \right] \leq 2 \exp \left\{ nd f \left(\frac{3}{4d}, 2 \frac{|\lambda|}{n} \right) \right\}.$$

This yields

$$\begin{aligned} \mathbf{P} \left[\sum_{i=1}^d M_i \geq \gamma \middle| \mathbf{E}^{\text{edge}}(\mathcal{A}_{d-1}) \right] &\leq 2 \min_{\lambda > 0} \exp \left\{ nd f \left(\frac{3}{4d}, 2 \frac{\lambda}{n} \right) - \lambda \gamma \right\} \\ &\stackrel{(a)}{\leq} 2 \left\{ \left(1 + \frac{2\gamma}{3} \right)^{-\left(1 + \frac{2\gamma}{3} \right) \frac{3}{3+4d}} \left(1 - \frac{\gamma}{2d} \right)^{-\left(1 - \frac{\gamma}{2d} \right) \frac{4d}{4d+3}} \right\}^{nd} \\ &\stackrel{(b)}{\leq} 2 \exp(-n\gamma^2/116), \end{aligned}$$

where (a) uses $\lambda = \lambda^*$ with

$$\lambda^* = \frac{n/2}{1 + 3/(4d)} \log \frac{1 + 2\gamma/3}{1 - \gamma/(2d)},$$

and (b) uses the following bound valid for any $d \geq 2$:

$$\frac{3d}{3+4d} \left(1 + \frac{2\gamma}{3} \right) \log \left(1 + \frac{2\gamma}{3} \right) + \frac{4d^2}{3+4d} \left(1 - \frac{\gamma}{2d} \right) \log \left(1 - \frac{\gamma}{2d} \right)$$

$$\begin{aligned}
&\stackrel{(i)}{\geq} \frac{3d}{3+4d} \left(1 + \frac{2\gamma}{3}\right) \left(\frac{2\gamma}{3} - \frac{2\gamma^2}{9}\right) - \frac{4d^2}{3+4d} \left(1 - \frac{\gamma}{2d}\right) \log \left(1 + \frac{\gamma/2d}{1 - \gamma/2d}\right) \\
&\stackrel{(ii)}{\geq} \frac{d}{3+4d} \left[3 \left(1 + \frac{2\gamma}{3}\right) \left(\frac{2\gamma}{3} - \frac{2\gamma^2}{9}\right) - 2\gamma\right] \stackrel{(iii)}{\geq} \gamma^2/25.
\end{aligned}$$

Here we used $x \geq \log(1+x) \geq x - x^2/2$ for $x > 0$ in (i) and (ii), and $\gamma < 1$ gives (iii).

Similarly, $\mathbb{P}\left[\sum_{i=1}^d M_i \leq -\gamma \mid \mathbf{E}^{\text{edge}}(\mathcal{A}_{d-1})\right] \leq 2 \exp(-n\gamma^2/25)$. Plugging these into (34), we get that if $n > \max\{25/\gamma^2 \log(4p^2/\delta), 108e^{2\beta} \log(2p^3/\delta)\}$, then (32) holds. \square

F Two trees lemma

Lemma F.1. *Let T_1 and T_2 be two spanning trees on set of nodes \mathcal{V} . Let w, \tilde{w} be a pair of nodes such that $\text{path}_{T_1}(w, \tilde{w}) \neq \text{path}_{T_2}(w, \tilde{w})$. Then there exists a pair of edges $f \triangleq (u, \tilde{u}) \in \text{path}_{T_1}(w, \tilde{w})$ and $g \triangleq (v, \tilde{v}) \in \text{path}_{T_2}(w, \tilde{w})$ such that*

(i) $f \notin \text{path}_{T_2}(w, \tilde{w})$ and $g \notin \text{path}_{T_1}(w, \tilde{w})$

(ii) $f \in \text{path}_{T_1}(v, \tilde{v})$ and $g \in \text{path}_{T_2}(u, \tilde{u})$

Since $f \in \text{path}_{T_1}(w, \tilde{w}) \cap \text{path}_{T_1}(v, \tilde{v})$, w and \tilde{w} (and respectively v and \tilde{v}) are in different subtrees of T_1 after removing edge f , one can label the end points of the edges $f = (u, \tilde{u})$ and $g = (v, \tilde{v})$ such that $u, v \in \text{SubTree}_{T_1, f}(w)$ and $\tilde{u}, \tilde{v} \in \text{SubTree}_{T_1, f}(\tilde{w})$ (See Figure 6.1).

Proof. We assume that T_1 and T_2 are two spanning trees with no common edges over a set of nodes \mathcal{V} . Otherwise, if T_1 and T_2 have any common edges, we can contract them and construct a new pair of spanning trees over set of nodes \mathcal{V}' in a way that preserves paths. Note that by assumption $\text{path}_{T_1}(w, \tilde{w}) \neq \text{path}_{T_2}(w, \tilde{w})$, so $w, \tilde{w} \in \mathcal{V}'$ are not merged in the contraction process.

Let u_1, \dots, u_{K+1} be the vertices of $\text{path}_{T_1}(w, \tilde{w})$ in order with $u_1 = w$ and $u_{K+1} = \tilde{w}$. Similarly, let v_1, \dots, v_{L+1} be the vertices of $\text{path}_{T_2}(w, \tilde{w})$ with $v_1 = w$ and $v_{L+1} = \tilde{w}$. Now, for $1 \leq k \leq K$, let

$$h_{T_1}(k) = \min\{t : (u_k, u_{k+1}) \in \text{path}_{T_1}(v_t, v_{t+1})\}.$$

We first show that $h_{T_1}(k)$ is well-defined (i.e., for all $1 \leq k \leq K$, there exists a $1 \leq t \leq L$ such that $(u_k, u_{k+1}) \in \text{path}_{T_1}(v_t, v_{t+1})$). For each $1 \leq k \leq K$, let $\mathcal{S}_k = \text{SubTree}_{T_1, (u_k, u_{k+1})}(u_k)$ and $\tilde{\mathcal{S}}_k = \text{SubTree}_{T_1, (u_k, u_{k+1})}(u_{k+1})$. Note that $w \in \mathcal{S}_k$ and $\tilde{w} \in \tilde{\mathcal{S}}_k$ for all k . Also, $(u_k, u_{k+1}) \in \text{path}_{T_1}(v_t, v_{t+1})$ for some $0 \leq t \leq L$ if and only if $v_t \in \mathcal{S}_k$ and $v_{t+1} \in \tilde{\mathcal{S}}_{k+1}$. For any given k , if such t does not exist, then either $v_{t'} \in \mathcal{S}_k$ or $v_{t'} \in \tilde{\mathcal{S}}_k$ for all $1 \leq t' \leq L+1$, which is a contradiction as $v_1 \in \mathcal{S}_k$ and $v_{L+1} \in \tilde{\mathcal{S}}_k$. Hence $h_{T_1}(k)$ is well-defined for all k . Equivalently, we could define $h_{T_1}(k) = \min\{t : v_{t+1} \in \tilde{\mathcal{S}}_k\}$. As $\tilde{\mathcal{S}}_k \subseteq \tilde{\mathcal{S}}_{k+1}$, the function $h_{T_1}(k)$ is non-decreasing in k .

Similarly, the function $h_{T_2}(l) = \min\{t : (v_l, v_{l+1}) \in \text{path}_{T_2}(u_t, u_{t+1})\}$ is well-defined and non-decreasing. To prove the result, we first show that there exists a pair of integers l^* and k^* such that $h_{T_1}(k^*) = l^*$ and $h_{T_2}(l^*) = k^*$. To do so, let function $h(l) : \{1, \dots, L\} \rightarrow \{1, \dots, L\}$ be $h(l) = h_{T_1}(h_{T_2}(l))$. This is a nondecreasing function over its domain. We now show that $h(l)$ has a fixed point. Assume that $h(1) > 1$ and $h(L) < L$, since otherwise the result is proven. Define $m = \min\{t : h(t) \leq t\}$ and note that $1 < m < L$. Thus, $h(m) \leq m$ while $h(m-1) > m-1$. Since $h(l)$ is non-decreasing, we have $m-1 < h(m-1) \leq h(m) \leq m$, which implies that $h(m) = m$.

Let l^* be a fixed point of $h(l)$. According to the definition of $h(l)$, the pair of integers l^* and $k^* = h_{T_2}(l^*)$ satisfies $h_{T_1}(k^*) = l^*$. Hence, the edge pair $(u_{k^*}, u_{k^*+1}) \in \text{path}_{T_1}(v_{l^*}, v_{l^*+1})$ and $(v_{l^*}, v_{l^*+1}) \in \text{path}_{T_2}(u_{k^*}, u_{k^*+1})$ satisfies the properties given in the Lemma. \square

G Bounds for risk (proof of Corollaries 3.5 and 3.6)

We provide a proof for Corollaries 3.5 on the risk which is a direct application of Theorem 3.3. By Theorem 3.3, for any $0 \leq \eta \leq 1$, given n samples, $\mathcal{L}^{(2)}(P, \Pi_{\text{Tcl}}(\hat{P})) \geq \eta$ with probability at most $p \exp(-Cne^{-2\beta}) + p/\eta \exp(-Cn\eta^2)$. Hence,

$$\begin{aligned}
\mathbb{E} \left[\mathcal{L}^{(2)}(P, \Pi_{\text{Tcl}}(\hat{P})) \right] &= \int_0^1 \mathbb{P} \left[\mathcal{L}^{(2)}(P, \Pi_{\text{Tcl}}(\hat{P})) \geq \eta \right] d\eta \\
&\stackrel{(a)}{\leq} \int_0^1 \min \left\{ 1, p \exp(-Cne^{-2\beta}) + \frac{p}{\eta} \exp(-Cn\eta^2) \right\} d\eta \\
&\stackrel{(b)}{\leq} \eta^* + \int_{\eta^*}^1 p \exp(-Cne^{-2\beta}) + p \exp(-Cn\eta^2) d\eta \\
&\leq \eta^* + p \exp(-Cne^{-2\beta}) + p \int_{\eta^*}^1 \exp(-Cn\eta^2) d\eta \\
&\stackrel{(c)}{\leq} \eta^* + p \exp(-Cne^{-2\beta}) + p \sqrt{\frac{\pi}{Cn}} \exp(-Cn\eta^{*2}/2) \\
&\stackrel{(d)}{\leq} 3\eta^*,
\end{aligned}$$

where (a) is a direct application of Theorem 3.3. Inequality (b) is true for any $0 < \eta^* \leq 1$. (c) uses the inequality $1/\sqrt{2\pi} \int_a^\infty \exp(-x^2/2) dx \leq \exp(-a^2/2)$. (d) holds for $\eta^* = p \exp(-Cne^{-2\beta}) + C' \sqrt{\frac{\log p}{n}}$ with a constant C' .

H Accurate k -wise marginals

We first describe our result, expressing marginals in terms of correlations, and then delineate implications to learning Ising models in order to make subsequent predictions from partial observations.

Let $\mathcal{S} \subset \mathcal{V}$ be a subset of variables. We are interested in the marginal $P(x_{\mathcal{S}})$ for an Ising model $P(x)$ of the form (1). Note that the variables have mean zero by symmetry of the model, hence $P(x_i) = \frac{1}{2}$ for each $i \in \mathcal{V}$. When $|\mathcal{S}| = 2$, the marginal over any pair of nodes $\mathcal{S} = \{i, j\}$ can be expressed as $P(X_i = x_i, X_j = x_j) = \frac{1}{4}[1 + x_i x_j \mu_{ij}]$ for $x_i, x_j \in \{-, +\}$.

We seek a simple closed form for the marginal distribution $P(x_{\mathcal{S}})$ as a function of pairwise correlations between pairs of nodes $i, j \in \mathcal{S}$. By the previous paragraph, we have already accomplished this when $|\mathcal{S}| = 2$. The general case $|\mathcal{S}| > 2$ is (as one might expect) rather more complicated and involves a certain maximization over matchings.

Let $\binom{\mathcal{V}}{2}$ denote the collection of pairs of nodes in \mathcal{V} . For $m \in \binom{\mathcal{V}}{2}$, if $m = \{i, j\}$, then we also write the correlation $\mu_m = \mathbb{E}_P X_i X_j$.

Definition 1. For subset of nodes $\mathcal{C} \subseteq \mathcal{V}$ of even cardinality $|\mathcal{C}| = 2t$, we define the set $\mathcal{M}_{\mathcal{C}}$ to be the collection of all partitions of \mathcal{C} into t disjoint pairs of nodes, i.e. the set of all perfect matchings in the complete graph with node set \mathcal{C} :

$$\mathcal{M}_{\mathcal{C}} = \left\{ M = \{m_1, \dots, m_t\} : m_l \in \binom{\mathcal{C}}{2} \text{ for each } l \in [t], m_l \cap m_k = \emptyset \text{ for } l \neq k \right\}.$$

Definition 2. Let $t \geq 1$ be a positive integer and $\mathcal{C} \subseteq \mathcal{V}$ be of size $|\mathcal{C}| = 2t$. The set-valued function $G(\mathcal{C})$ is defined as the set of matchings M in $\mathcal{M}_{\mathcal{C}}$ that maximize the product of correlations for

elements in M :

$$G(\mathcal{C}) = \arg \max_{M \in \mathcal{M}_{\mathcal{C}}} \prod_{m \in M} |\mu_m|. \quad (37)$$

We define $G^*(\mathcal{C})$ to be an arbitrary element in $G(\mathcal{C})$, and $G(\emptyset) = \emptyset$ and $G^*(\emptyset) = \emptyset$. We also use the notation $G^*(\mathcal{C}, P)$ to be optimal matching $G^*(\mathcal{C})$ under the distribution P to explicitly determine the underlying distribution.

We emphasize that often the maximum in (37) is not uniquely achieved and the set $G(\mathcal{C})$ can be large. In what follows we use the convention that the product $\prod_{m \in \mathcal{I}} \mu_m = 1$ whenever $\mathcal{I} = \emptyset$. Our main result is the following:

Theorem H.1. *If $P(x)$ is the probability mass function for an Ising model (1) on any tree T , then the marginal $P(x_{\mathcal{S}})$ over the subset $\mathcal{S} \subseteq \mathcal{V}$ can be written as:*

$$P(x_{\mathcal{S}}) = 2^{-|\mathcal{S}|} \sum_{t=0}^{\lfloor |\mathcal{S}|/2 \rfloor} \sum_{\substack{\mathcal{C} \subseteq \mathcal{S} \\ |\mathcal{C}|=2t}} \prod_{c \in \mathcal{C}} x_c \prod_{m \in G^*(\mathcal{C})} \mu_m. \quad (38)$$

Proof. For subset of variables $\mathcal{S} \subseteq \mathcal{V}$, the marginal distribution $P_{\mathcal{S}}$ is

$$\begin{aligned} P_{\mathcal{S}}(x_{\mathcal{S}}) &= \sum_{x_{\mathcal{S}^c}} P(x) \stackrel{(a)}{=} \sum_{x_{\mathcal{S}^c}} \prod_{(i,j) \in \mathcal{E}} \frac{1 + \mu_{ij} x_i x_j}{2} \stackrel{(b)}{=} \frac{1}{2^p} \sum_{x_{\mathcal{S}^c}} [1 + \sum_{\mathcal{F} \subseteq \mathcal{E}, \mathcal{F} \neq \emptyset} \prod_{(i,j) \in \mathcal{F}} \mu_{ij} x_i x_j] \\ &\stackrel{(c)}{=} \frac{1}{2^{|\mathcal{S}|}} [1 + \sum_{\substack{\mathcal{F} \subseteq \mathcal{E}, \\ \mathcal{F} \in \Omega(\mathcal{S}), \mathcal{F} \neq \emptyset}} \prod_{(i,j) \in \mathcal{F}} \mu_{ij} x_i x_j] \end{aligned}$$

where (a) uses the Equation (2) for $P \in \mathcal{P}_{\mathsf{T}}$. In (b) we expand the product term over all the edges in the tree. In (c), $\Omega(\mathcal{S}) \subseteq 2^{\mathcal{E}}$ is defined such that for $\mathcal{F} \subseteq \mathcal{E}$, if $\mathcal{F} \in \Omega(\mathcal{S})$, then for all nodes $i \in \mathcal{S}^c$, there are even number of edges $e \in \mathcal{F}$ that have i as their end point. Equivalently, any $\mathcal{F} \in \Omega(\mathcal{S})$ is the collection of edges in the set of edge-disjoint paths between distinct pairs of nodes in \mathcal{S} . \square

The utility of this expression is in part due to the fact that it depends on the tree topology T only through the optimization implicit in $G(\mathcal{C})$, which depends only on end-to-end correlations between variables in $\mathcal{C} \subseteq \mathcal{S}$. In particular, the expression is agnostic to the topology over variables not in \mathcal{S} : even the *number* of such variables is irrelevant. This is important because there can be several (non-isomorphic) subtrees which give the same marginal over \mathcal{S} . This means that it is in general impossible to detect the tree topology just by observing end-to-end correlations, but this expression is robust to this.

Lemma 1. *For any set $\mathcal{C} \subseteq \mathcal{V}$ with $|\mathcal{C}| = 2t$, and any pair of tree structured Ising distributions P and Q as in (1) (with possibly different underlying tree), let $\mu_{ij} = \mathbb{E}_P X_i X_j$ and $\tilde{\mu}_{ij} = \mathbb{E}_Q X_i X_j$. If $|\mu_{ij} - \tilde{\mu}_{ij}| \leq \epsilon$ for all $i, j \in \mathcal{C}$, then*

$$\left| \prod_{m \in G^*(\mathcal{C}, P)} \mu_m - \prod_{m' \in G^*(\mathcal{C}, Q)} \tilde{\mu}_{m'} \right| \leq 2t\epsilon.$$

Proof. First, we will show that for all $M \in \mathcal{M}_{\mathcal{C}}$, $\prod_{m \in M} \mu_m$ takes the same sign (i.e., for $M, M' \in \mathcal{M}_{\mathcal{C}}$, $\prod_{m \in M} \mu_m \prod_{m' \in M'} \mu_{m'} \geq 0$). Similar statement holds for $\tilde{\mu}$. To do so, note that for all $M, M' \in \mathcal{M}_{\mathcal{C}}$, one can make a sequence of matchings $M_0 = M, M_1, \dots, M_r = M'$ such that M_r and M_{r+1} differ only at two pairs of nodes (i.e., There exists $(i, j), (k, l) \in M_r$ such that $(i, k), (j, l) \in M_{r+1}$).

The product of correlations on M_r and M_{r+1} have the same sign (i.e., $\prod_{m \in M_r} \mu_m \prod_{m \in M_{r+1}} \mu_m \geq 0$). The same argument applies for all r and hence $\prod_{m \in M} \mu_m \prod_{m' \in M'} \mu_{m'} \geq 0$.

Next, we show that if $|\mu_{ij} - \tilde{\mu}_{ij}| \leq \epsilon$ for all $i, j \in \mathcal{C}$, then for any $M = \{m_1, m_2, \dots, m_t\} \in \mathcal{M}_{\mathcal{C}}$, $|\prod_{m \in M} \mu_m - \prod_{m' \in M} \tilde{\mu}_{m'}| \leq t\epsilon$.

$$\begin{aligned} \left| \prod_{i=1}^t \mu_{m_i} - \prod_{i=1}^t \tilde{\mu}_{m_i} \right| &= \left| \sum_{i=1}^t \prod_{j=1}^{i-1} \mu_{m_j} \prod_{j'=i+1}^t \tilde{\mu}_{m_{j'}} (\mu_{m_i} - \tilde{\mu}_{m_i}) \right| \\ &\leq \sum_{i=1}^t \left| \prod_{j=1}^{i-1} \mu_{m_j} \prod_{j'=i+1}^t \tilde{\mu}_{m_{j'}} (\mu_{m_i} - \tilde{\mu}_{m_i}) \right| \leq \sum_{i=1}^t |\mu_{m_i} - \tilde{\mu}_{m_i}| \leq t\epsilon. \end{aligned}$$

Let's define

$$a = \prod_{m \in G^*(\mathcal{C}, P)} \mu_m, \quad a' = \prod_{m \in G^*(\mathcal{C}, Q)} \mu_m, \quad b = \prod_{m \in G^*(\mathcal{C}, P)} \tilde{\mu}_m, \quad b' = \prod_{m \in G^*(\mathcal{C}, Q)} \tilde{\mu}_m.$$

Hence, we showed that $aa' \geq 0, bb' \geq 0, |a - b| \leq t\epsilon$ and $|a' - b'| \leq t\epsilon$. We also know that $|a| \geq |a'|$ and $|b| \geq |b'|$ according to the definition of $G^*(\mathcal{C}, P)$ and $G^*(\mathcal{C}, Q)$. To bound $|a - b'|$, we study two different cases: Either $ab \geq 0$ in which case $|a - b'| = ||a| - |b'|| \leq \max\{|a - b|, |a' - b'|\} \leq t\epsilon$. Or $ab \leq 0$ in which case $|a - b'| = |a| + |b'| \leq |a - b| + |a' - b'| \leq 2t\epsilon$. \square

Proposition 1. Let $\mathbb{T} = (\mathcal{V}, \mathcal{E})$ be a tree and P an Ising model on \mathbb{T} . Let Q be an Ising model on tree $\mathbb{T}' = (\mathcal{V}, \mathcal{E}')$. If $|\mathbb{E}_P X_i X_j - \mathbb{E}_Q X_i X_j| \leq \eta$ for all $i, j \in \mathcal{V}$ (i.e., $\text{ssTV}^{(2)}(P, Q) \leq \eta/2$), then $\text{ssTV}^{(k)}(P, Q) \leq k2^k \eta$.

Proof. For $i, j \in \mathcal{V}$, let $\mu_{ij} \triangleq \mathbb{E}_P X_i X_j$ and $\tilde{\mu}_{ij} \triangleq \mathbb{E}_Q X_i X_j$. We also use the notation $G^*(\mathcal{C}, P)$ to be optimal matching $G^*(\mathcal{C})$ under the distribution P .

To bound $\mathcal{S} \subseteq \mathcal{V}$ with $|\mathcal{S}| = k$, we use Lemma 1 to get:

$$\begin{aligned} d_{\text{TV}}(P_{\mathcal{S}}, Q_{\mathcal{S}}) &= \sum_{x_{\mathcal{S}} \in \{-, +\}^k} |P(x_{\mathcal{S}}) - Q(x_{\mathcal{S}})| \\ &= \frac{1}{2^k} \sum_{x_{\mathcal{S}} \in \{-, +\}^k} \left| \sum_{t=0}^{\lfloor \frac{k}{2} \rfloor} \sum_{\substack{\mathcal{C} \subseteq \mathcal{S} \\ |\mathcal{C}|=2t}} \prod_{c \in \mathcal{C}} x_c \prod_{m \in G^*(\mathcal{C}, P)} \mu_m - \sum_{t=0}^{\lfloor \frac{k}{2} \rfloor} \sum_{\substack{\mathcal{C} \subseteq \mathcal{S} \\ |\mathcal{C}|=2t}} \prod_{c \in \mathcal{C}} x_c \prod_{m \in G^*(\mathcal{C}, Q)} \tilde{\mu}_m \right| \\ &\leq \sum_{t=0}^{\lfloor \frac{k}{2} \rfloor} \sum_{\substack{\mathcal{C} \subseteq \mathcal{S} \\ |\mathcal{C}|=2t}} \frac{1}{2^k} \sum_{x_{\mathcal{S}} \in \{-, +\}^k} \left| \prod_{c \in \mathcal{C}} x_c \right| \cdot \left| \prod_{m \in G^*(\mathcal{C}, P)} \mu_m - \prod_{m \in G^*(\mathcal{C}, Q)} \tilde{\mu}_m \right| \\ &= \sum_{t=0}^{\lfloor \frac{k}{2} \rfloor} \sum_{\substack{\mathcal{C} \subseteq \mathcal{S} \\ |\mathcal{C}|=2t}} \left| \prod_{m \in G^*(\mathcal{C}, P)} \mu_m - \prod_{m \in G^*(\mathcal{C}, Q)} \tilde{\mu}_m \right| \\ &\leq \sum_{t=0}^{\lfloor \frac{k}{2} \rfloor} \sum_{\substack{\mathcal{C} \subseteq \mathcal{S} \\ |\mathcal{C}|=2t}} 2t\eta \leq \sum_{t=0}^{\lfloor \frac{k}{2} \rfloor} \binom{k}{2t} 2t\eta \leq k2^k \eta. \quad \square \end{aligned}$$

\square

Note that learning the tree T^{CL} from n samples using the variant of Chow-Liu algorithm has time complexity of $O((n+1)p^2)$. Finding low-order marginals and posteriors of any subset of variables on a tree-structured distribution takes $O(p)$ operations using belief-propagation algorithm.

I Numerical simulations

In this section we give numerical simulations on the following setups: 1) comparison of ssTV loss for the forest approximation algorithm with the Chow-Liu algorithm; 2) approximation of higher-order marginals, beyond pairwise; 3) tree-structured Ising model with external field; 4) model misspecification.

Forest approximation algorithm We simulate the forest approximation algorithm introduced in Section 4.2 in the paper. In this simulation, the data is generated from a distribution factorized as Equation (1) using a tree with $p = 31$ nodes and maximum edge strength $\beta = 1.5$ and various values of α .

The forest approximation algorithm implemented in our simulations is slightly different from the one introduced in Section 4.2. The algorithm introduced in Section 4.2 chooses a threshold, $\tau(\epsilon) = 4\epsilon/\sqrt{1 - \tanh \beta}$ for $\epsilon = \sqrt{2/n \log(2p^2/\delta)}$ as function of the parameters of the model and removes the edges added by the Chow-Liu algorithm whose weights are smaller than the threshold $\tau(\epsilon)$. This process guarantees that with probability $1 - \delta$ the reconstructed forest is a subgraph of the original tree. In our simulations, we use a slightly different approach in choosing the value of the threshold. Instead of fixing (a conservative) value for the threshold, for each sample and corresponding run of the algorithm, the threshold used is the best choice for that specific sample. The threshold $\tilde{\tau}^{(i)}$ for sample i is the minimum number such that removing the edges with weights at most $\tilde{\tau}^{(i)}$ from the the maximum weight tree results in a subgraph of the original tree. We claim that the $\mathcal{L}^{(2)}$ loss calculated for this algorithm (which uses aid from an oracle) is strictly smaller than the loss according to \tilde{d} defined in (8). To see that, note that if $\tau(\epsilon) < \tilde{\tau}^{(i)}$, then the output of the original forest approximation algorithm is not a subgraph of the original tree (in which case $\tilde{d} = 1$). If $\tau(\epsilon) \geq \tilde{\tau}^{(i)}$, then the original forest approximation algorithm (which uses $\tau(\epsilon)$) might remove more correct edges from the Chow-Liu tree than necessary. We verified that in each simulated instance $\tilde{\tau}^{(i)}$ yields better performance in terms of $\mathcal{L}^{(2)}$ (We ran the original forest approximation algorithm introduced in Section 4.2 for various values of δ and this was the case for all δ).

In Figure I.1, the $\mathcal{L}^{(2)}$ loss for the output of this forest approximation algorithm is compared with the Chow-Liu algorithm for various values of α . Again, note that the output of our variation of oracle-aided forest approximation algorithm is always a subgraph of the original tree and its $\mathcal{L}^{(2)}$ loss is no larger than the original forest approximation algorithm (*i.e.*, it is more powerful as it uses the help from the oracle). Figure I.1 suggests that the $\mathcal{L}^{(2)}$ loss entailed by the Chow-Liu algorithm is smaller than both the original forest approximation and the variation used in the simulations. This confirms our main message of the paper that learning possibly incorrect tree by the Chow-Liu algorithm performs better than the forest approximation algorithm in terms of $\mathcal{L}^{(2)}$ loss.

Accurate k -wise marginals for general k Theorem 1.1 bounds the number of samples necessary to guarantee small $\mathcal{L}^{(2)}$ loss using the Chow-Liu algorithm in learning tree-structured Ising models with no external field. Corollary 1.2 states the preliminary results to bound the $\mathcal{L}^{(k)}$ loss for general k . Next, we use numerical simulations to bound the $\mathcal{L}^{(k)}$ for general values of k as a function of number of samples. The generative distribution $P(x)$ is factorized according to (1) on a random

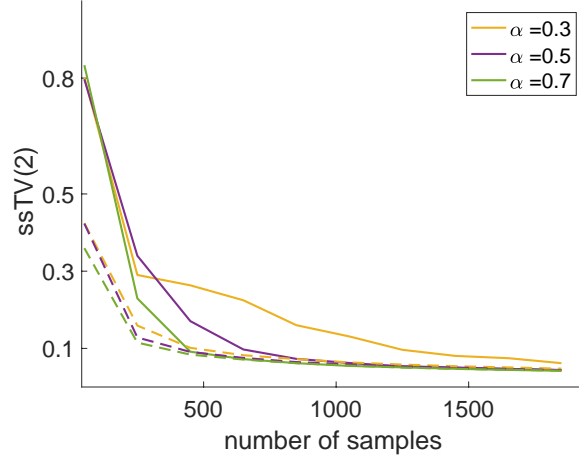


Figure I.1: The generative distribution is factorized as Equation (1) on a tree with $p = 31$ nodes, $\beta = 1.5$ and various values of α . The solid lines depict the $\mathcal{L}^{(2)}$ loss of the forest approximation algoirhtm. The dashed lines are the $\mathcal{L}^{(2)}$ loss of the Chow-Liu algorithm.

tree with $p = 7$ nodes, $\alpha = 0.8$ and $\beta = 2$. Figure I.2 depicts the $\mathcal{L}^{(k)}(P, Q)$ as a function of number of samples where $Q(x)$ is the output of the Chow-Liu algorithm.

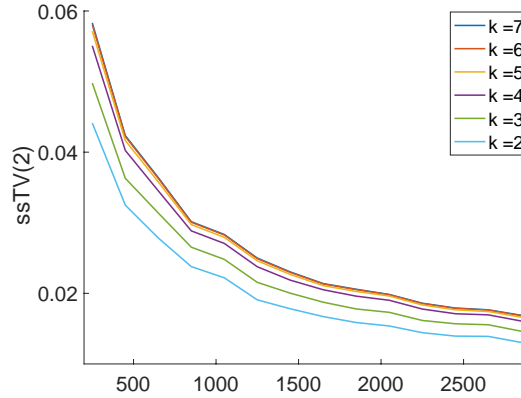


Figure I.2: The generative distribution $P(x)$ is factorized according to (1) on a random tree on $p = 7$ nodes, with $\alpha = 0.8$ and $\beta = 2$. The distribution $Q(x)$ is the output of the Chow-Liu algorithm. This figure depicts the $\mathcal{L}^{(k)}(P, Q)$ loss as a function of number of samples for various values of k .

Ising model with external field We use the numerical simulations to study the performance of Chow-Liu algorithm on tree-structured Ising models in presence of external field. In this case, the singleton marginals over each variable are non-uniform. In our simulation, we generate a random tree and choose random values of parameters to make a generative tree-structured distribution $P(x)$. The output of Chow-Liu algorithm is a tree structured Ising model $Q(x)$. Figure I.3 depicts the $\mathcal{L}^{(2)}(P, Q)$ loss as function of number of samples used by the Chow-Liu algorithm. It conforms the

main message of the paper which states that the output of the Chow-Liu algorithms has accurate pairwise marginals and can be used for subsequent predictions.

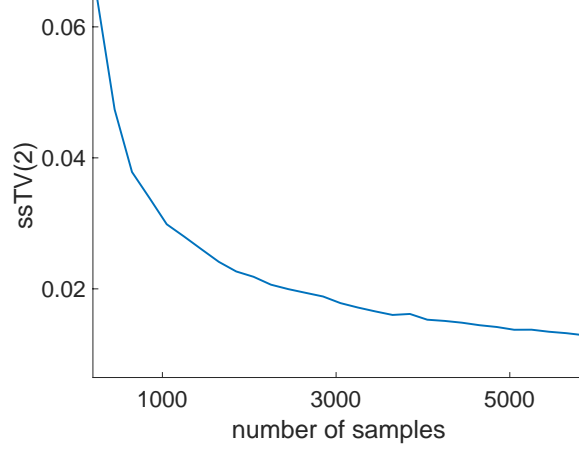


Figure I.3: The generative distribution is a tree structured Ising model with external field over a random tree and random values of parameters. The distribution $Q(x)$ is the output of Chow-Liu algorithm. This figure depicts $\mathcal{L}^{(2)}(P, Q)$ loss as function of number of samples.

Misspecified model We study the effect of misspecification in the modelling on the performance of Chow-Liu algorithm using numerical simulations. The distribution $\tilde{P}(x)$ is a tree structured Ising model factorized as Equation (1) over a tree (uniformly randomly chosen on the set of trees on $p = 6$ nodes) with $\alpha = 0.1$ and $\beta = 0.8$. The generative distribution P is constructed by adding the random isotropic offset to the distribution $\tilde{P}(x)$. Hence, the distribution $P(x)$ used to generate samples is not a tree structured Ising model in this simulation. Let $\gamma \triangleq \mathcal{L}^{(2)}(P, \tilde{P})$. The output of Chow-Liu algorithm on n samples is a tree structured distribution $Q(x)$. Figure I.4 depicts the $\mathcal{L}^{(2)}(P, Q)$ loss for various values of γ . As the number of samples n grows, $\mathcal{L}^{(2)}(P, Q)$ decays and $\lim_{n \rightarrow \infty} \mathcal{L}^{(2)}(P, Q) = \gamma$. The simulation results suggest that the Chow-Liu algorithm is robust to the misspecification in the model.

J Comparison of various tree-learning algorithms

Here, we review different approaches that could be taken toward learning a tree-structured distribution. We also review some known algorithms and their sample complexity. The input of these algorithm is n i.i.d. samples from a tree-structured distribution $P(x) \in \mathcal{P}_{\mathsf{T}}(\alpha, \beta)$ with $\mathsf{T} = (\mathcal{V}, \mathcal{E})$ being a tree. The guarantees mentioned below are with high probability.

- *Guaranteed recovery of the structure*

Algorithm: Chow-Liu algorithm [12]

Output: T^{CL}

Guarantee: $\mathsf{T} = \mathsf{T}^{\text{CL}}$

Requirement: $n > C \frac{e^{2\beta}}{\alpha^2} \log p$

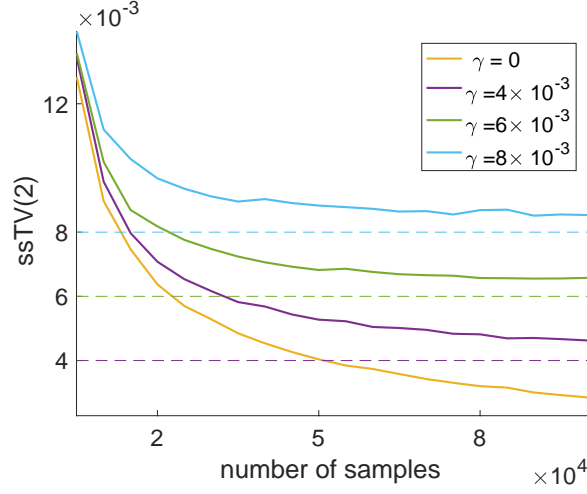


Figure I.4: The distribution $\tilde{P}(x)$ is factorized as Equation (1) over a random tree with $p = 6$ nodes with $\alpha = 0.1$ and $\beta = 0.8$. The generative distribution $P(x)$ is constructed by adding the random isotropic offset to the distribution $\tilde{P}(x)$. The distribution $Q(x)$ is the output of Chow-Liu algorithm. The figure depicts $\mathcal{L}^{(2)}(P, Q)$ as function of number of samples for various values of $\gamma \triangleq \mathcal{L}^{(2)}(P, \tilde{P})$.

In Theorems 3.2 and 3.1 we provided the sample complexity of this problem (tight up to a constant factor).

- *Guaranteed recovery of the structure and accurate pairwise marginal*

Algorithm: Chow-Liu algorithm [12]

Output: T^{CL} and $Q \in \mathcal{P}_{T^{\text{CL}}}$

Guarantee: $T = T^{\text{CL}}$ and $\mathcal{L}^{(2)}(P, Q) \leq \eta$.

Requirement: $n > C \max\{\frac{e^{2\beta}}{\alpha^2}, \frac{1}{\eta^2} \log\} \log p$

- *Forest Approximation and accurate pairwise marginal*

Algorithm: Chow-Liu algorithm with a proper threshold over the weight of edges

Output: $\hat{T} = (\mathcal{V}, \mathcal{E}_{\hat{T}})$ and $Q \in \mathcal{P}_{\hat{T}}$

Guarantee: $\mathcal{E}_{\hat{T}} \subseteq \mathcal{E}_T$ and $\mathcal{L}^{(2)}(P, Q) \leq \eta$.

Requirement: $n > C \frac{e^{2\beta}}{\eta^2} \log p$

Proved in Proposition 4.1

- *Accurate pairwise marginals*

Algorithm: Chow-Liu Algorithm for the purpose of inference

Output: $T^{\text{CL}} = (\mathcal{V}, \mathcal{E})$ and $Q \in \mathcal{P}_{T^{\text{CL}}}$

Guarantee: $\mathcal{L}^{(2)}(P, Q) \leq \eta$.

Requirement: $n > C \max\{e^{2\beta}, \eta^{-2}\} \log p$

Proved in Theorem 3.3

- *Latent tree with accurate pairwise marginals*

Algorithm: Agarwala, Bafna, Farach, Paterson, and Thorup [2].

Output: $S = (\mathcal{V} \cup \tilde{\mathcal{V}}, \mathcal{E}_S)$ and $Q \in \mathcal{P}_S$ such that \mathcal{V} is the set of leaves of S .

Guarantee: $\mathcal{L}^{(2)}(P, Q_{\mathcal{V}}) \leq \eta$.

Requirement: $n > C \frac{1}{\alpha^2 \eta^2} \log p$

Discussed in Section 1.1.

- *Latent tree with accurate joint distribution over the leaves*

Algorithm: Ambainis, Desper, Farach, and Kannan [3].

Output: $S = (\mathcal{V} \cup \tilde{\mathcal{V}}, \mathcal{E}_S)$ and $Q \in \mathcal{P}_S$ such that \mathcal{V} is the set of leaves of S .

Guarantee: $\mathcal{L}^{(p)}(P, Q_{\mathcal{V}}) \leq \eta$.

Requirement: $n > C \max\{\frac{1}{\alpha^2}, p \frac{e^{2\beta}}{\eta^2}\}$

Discussed in Section 1.1.

The last two introduced problems are examples of improper learning in which the class of models from which we learn is bigger than the class of models assumed over the original distribution. In this case, we assume that the original distribution is tree-structured over the nodes in \mathcal{V} whereas the learned model is represented with a latent tree such that its leaves are the nodes in \mathcal{V} . Specific applications determine the requirements over the learned models and the guarantees which are necessary for the subsequent purposes. But in general proper learning is necessary when the identification of the qualitative structure behind the distributions is necessary. The extreme case of which is when one is only interested in the underlying structure which corresponds to the first problem in above category.

Note that every tree-structured distribution over p nodes in \mathcal{V} can be equivalently represented using a latent tree with \mathcal{V} as its leaves. To do so, one has to add some dummy variables and infinite strength edges to the original tree. But the converse is not true: A distribution represented on the leaves of a latent tree cannot be factorized as a tree-structured distribution in general.