Information Thresholds for Non-Parametric Structure Learning on Tree Graphical Models

Konstantinos Nikolakakis, *Student Member, IEEE,*
Dionysios S. Kalogerias, *Member, IEEE,* and Anand D. Sarwate, *Senior Member, IEEE*

***Abstract*—We provide high probability finite sample complexity guarantees for non-parametric structure learning of tree-shaped graphical models whose nodes are discrete random variables with either finite or countable alphabets, both in the noiseless and noisy regimes. We study a fundamental quantity called the *(noisy) information threshold*, which arises naturally from the error analysis of the Chow-Liu algorithm and, as we discuss, provides explicit necessary and sufficient conditions on sample complexity, by effectively summarizing the difficulty of the tree-structure learning problem. Specifically, we show that finite sample complexity of the Chow-Liu algorithm for ensuring exact structure recovery is inversely proportional to the information threshold (provided it is positive), and scales almost logarithmically relative to the number of nodes over a given probability of failure, also matching relevant asymptotic results in the literature. Conversely, in the noiseless case, we show that, for arbitrarily small information thresholds, the structure recovery task given any finite number of samples becomes impossible for any algorithm whatsoever. Consequently, strict positivity of the information threshold characterizes the feasibility of tree-structure learning, in general terms. Lastly, as a consequence of our analysis, we resolve the problem of tree structure learning in the presence of *non-identically* distributed observation noise, providing conditions for convergence of the Chow-Liu algorithm under this setting, as well.**

***Index Terms*—Tree-Structured Graphical Models, Chow-Liu algorithm, Information Threshold, Hidden Markov Models**

I. Introduction

GRAPHICAL models are a widely-used and powerful tool for analyzing high-dimensional structured data [1], [2]. In those models, variables are represented as nodes of a graph, the edges of which indicate conditional dependencies among the corresponding nodes. In this paper, we consider the problem of *learning acyclic undirected graphs* or *tree-structured Markov Random Fields (MRFs)*. In particular, we study the well-known *Chow-Liu (CL) Algorithm* [3], which, given a dataset of samples drawn from a tree-structured distribution, returns an estimate of the original tree. The importance of the CL algorithm stems from its efficiency, its low computational complexity, and also its optimality in terms of sample complexity (matching information-theoretic limits) for a variety of statistical settings, e.g., such involving Gaussian data [4], binary data [5], as well as for non-parametric models with discrete alphabets [6], [7].

Our contribution centers on a rigorous characterization of a fundamental model-dependent statistic we which call the *information threshold* (denoted as $\mathbf{I}^o$). We show that $\mathbf{I}^o$ quantifies not only the sample complexity of the CL algorithm in particular, but also the complexity of the tree-structure learning problem in general. Information thresholds have been already appeared in fundamental prior work in the area (e.g., [6] and [8]). Precisely, in Tan et al. [8, Theorem 3.5], it is shown that if the information threshold is strictly positive, then the probability of incorrect structure recovery by the CL algorithm decays exponentially with respect to the number of samples, with the corresponding error exponent [7], [9] being *implicitly* dependent on the information threshold. However, no explicit connection exists between the sample complexity of the CL algorithm (*either* finite *or* asymptotic) and the value of the information threshold in the literature (e.g., in [6], [8]). In fact, although it is true that the number of samples required for successful recovery scales logarithmically with respect to the number of nodes p (in the asymptotic sense), this sample complexity can vary significantly for different distributions (for fixed p), and fixed probability of failure δ.

The discussion above naturally raises the following basic question: *Is there any essential statistic (e.g., the information threshold) that determines the sample complexity of the tree-structure learning problem?* Here, we show that, indeed, the information threshold constitutes such a representative statistic; in fact, it provides explicit *necessary and sufficient* conditions on sample complexity, by effectively and completely summarizing the difficulty of the tree-structure learning problem.

Specifically, we consider general tree-structured graphical models with variables taking values in either finite or countable sets. First, we study the noiseless case, in which all variables corresponding to the nodes of the underlying tree structure are perfectly observable. Our contributions are as follows.

- We provide finite sample complexity bounds for the CL algorithm (Theorems 1 and 2), which depend on the underlying tree-structured distribution through the information threshold, the latter essentially acting as a summary statistic. In particular, the sample complexity with respect to the ratio p/δ scales *almost logarithmically* as $\mathcal{O}\big(\log^{1+\zeta}(p/\delta)\big)$, for any $\zeta > 0$, whereas, with respect to the associated information threshold, the same complexity is of the the order of $\mathcal{O}(1/(\mathbf{I}^o)^{2(1+\zeta)})$, for any $\zeta > 0$. Our finite sample complexity bounds hold for arbitrary tree-structured distributions (summarized through the information threshold) and extend

K. Nikolakakis and A.D. Sarwate are with the Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ, 08854 USA (e-mail: {k.nikolakakis, anand.sarwate}@rutgers.edu).

D. Kalogerias is with the Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, 19104, USA (e-mail: dionysis@seas.upenn.edu).

The work of Sarwate and Nikolakakis was supported in part by the United States National Science Foundation under award CCF-1453432, and DARPA and SSC Pacific under contract N66001-15-C-4070.

Manuscript received May xx, 2020; revised xx xx, 2020.

the sample complexity bounds previously developed by Tan et al. [6], [7] (asymptotic) and Bresler and Karzand [5] (finite), in a plausible non-parametric setting.

- We prove that, for *any* data-driven tree-structure estimator, the associated worst-case probability of incorrect structure recovery over all tree-structured MRFs with arbitrarily small information threshold (i.e., as $\mathbf{I}^o \downarrow 0$) is at least $1/2$, for any dataset of finite length (Theorem 3). This result shows that $\mathbf{I}^o$ is fundamental, in the sense that its strict positivity characterizes the feasibility of the tree-structure learning problem. In other words, our result implies that no algorithm can successfully recover the underlying tree structure *uniformly* in any set of tree-structured MRFs with $\mathbf{I}^o \downarrow 0$, from any finite number of samples. For such a task ever to be successful (using any algorithm whatsoever), either $\mathbf{I}^o$ must be uniformly bounded away from 0 (within a postulated model class), or further assumptions on the family of tree-structured models considered are required.

Apart from the noiseless case, we also study the tree structure learning problem in a noisy (partially observable) setting. In many applications, the underlying physical or artificial phenomenon may be well-modeled by a (tree-structured) MRF but the data acquisition device or sensor may itself introduce noise. We wish to understand how sensitive the performance of an algorithm is to noisy inputs. There are two recent works that study the impact of corrupted observations on both binary and Gaussian models [10], [11]. Goel et al. [10] extend the Interaction Screening Objective [12] for the case of Ising models, while our prior work [11] analyzes the performance of the CL algorithm for both noisy Ising and Gaussian models. In particular, this work showed the consistency of the CL algorithm under the assumption of independent and identically distributed noise.

Based on this prior work, it is tempting to think that one can successfully perform structure recovery by running the standard CL algorithm *directly* on noisy data. This is common in practice, because the distribution of the noise may be unknown. However, even for simplest models with binary observations, the CL algorithm with noisy input data may not be consistent (and divergence is guaranteed) without considering a pre-processing procedure. Figure 1 shows a simple binary-valued example of a 3-variable tree with non-identically distributed noise; even for this simple example, structure learning from raw data can be infeasible. However, if we denoise the observations first (a form of data pre-processing), CL will return the true structure with high probability.

This simple example raises another natural question: *Under which circumstances (e.g., using some form of pre-processing) can the CL algorithm be successful when used directly on noisy data?* We resolve this question by defining and analyzing the properties of the *noisy information threshold* (as termed herein, and denoted as $\mathbf{I}^o_\dagger$), which, as it turns out, characterizes the finite sample complexity of the CL algorithm on hidden tree-structured MRFs. Our contributions are as follows.

- We provide the first results on *non-parametric* hidden tree-structure learning, where the distributions of the hidden and observable layers are *unknown* and the mappings between the hidden and observed variables are general. In fact, as

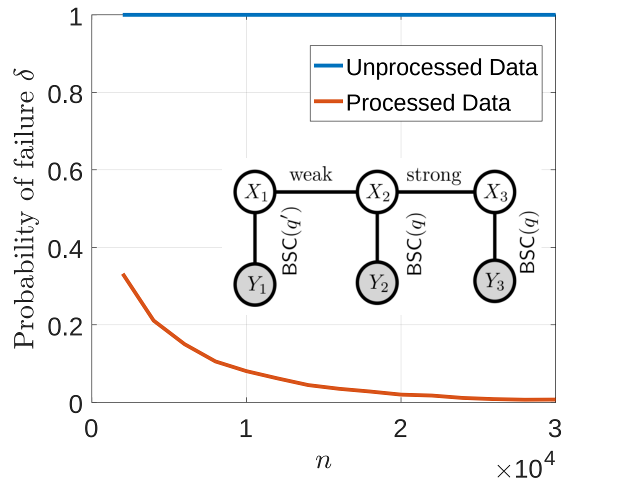

Fig. 1. $\delta \equiv \hat{\mathbb{P}}_n\big(\mathrm{T}^{\mathrm{CL}}_{\dagger} \neq \mathrm{T}\big)$ for 3-node hidden structure whose $\mathbf{I}^o_\dagger < 0$ (non-identically distributed noise). "Weak" and "Strong" refer to correlations between hidden variables: $|\mathbb{E}[X_1X_2]| < |\mathbb{E}[X_2X_3]|$.

with the noiseless case, we show that, as long as $\mathbf{I}^o_\dagger > 0$, the sample complexity of the CL algorithm with respect to p/δ scales as $\mathcal{O}\big(\log^{1+\zeta}(p/\delta)\big)$, for any $\zeta > 0$; also, with respect to the associated *noisy* information threshold, the same sample complexity is of the order of $\mathcal{O}(1/(\mathbf{I}^o_\dagger)^{2(1+\zeta)})$, for any $\zeta > 0$ (Theorem 4).

- Additionally, we show that, if $\mathbf{I}^o_\dagger \leq 0$, then structure recovery from raw data is not possible. Still, whenever $\mathbf{I}^o_\dagger < 0$, by introducing suitable *pre-processing on the noisy observations* and enforcing appropriate conditions on the hidden model (Definition 4), we show that the CL algorithm can be made convergent (with the aforementioned sample complexity). Such conditions can be satisfied for a variety of interesting observation models. Essentially, our results confirm that the CL algorithm is an effective *universal estimator* for tree-structure learning, on the basis of either noiseless or noisy data. We lastly explicitly illustrate how our results capture certain interesting scenarios involving *generalized M-ary* erasure and symmetric channels (i.e., observation noise models). Our framework extends and unifies recent results proved earlier for the binary case [5], [11], by considering *non-identically distributed noise* as well. Note that this latter problem remains open for general graph structures; see e.g., [10, Section 6].

A. Related Work

Structure learning from node observations is a fundamental and well-studied problem in the context of graphical models. For general graph structures, the complexity of the problem has been studied by Karger and Srebro [13]. Under the assumption of bounded degree the problem becomes tractable, leading to a large body of work in the last decade [12], [14]–[19]. These approaches employ greedy algorithms, l_1 regularization methods, or other optimization techniques. The sample complexity of each approach is evaluated based on information theoretic bounds [20]–[22].

For acyclic graphs, the CL algorithm is computationally efficient and it has been shown to be optimal in terms of sample complexity with respect to the number of variables. The error analysis and convergence rates for trees and forests were studied by Tan et al. [4], [6], [7], Liu et al. [23] and Bresler and Karzand [5]. In particular, for finite alphabets the number of samples needed by the CL algorithm is logarithmic with respect to the ratio p/δ for tree structures and polylogarithmic in the case of forests: $\mathcal{O}\big(\log^{1+\zeta}(p/\delta)\big)$, for all $\zeta > 0$ [7, Theorem 5]). Comparatively, our results extend to countable alphabets and noisy observations (hidden models). Our sample complexity bounds are consistent with prior work: the order is poly-logarithmic but remains arbitrarily close to logarithmic ($\mathcal{O}\big(\log^{1+\zeta}(p/\delta)\big)$, for all $\zeta > 0$). In the special case of Ising models, hidden structures were recently considered by Goel et al. [10] and our previous work [11], [24]. Goel et al. [10] consider bounded degree graphs and require the noise model (or an approximation of it) to be known with the noise i.i.d. on each variable (see their Section 6). Our results apply more generally to settings with *non-identically distributed noise* (Section V-C), but are restricted to the tree structure model assumption. Our prior work [11] solves the problem of hidden tree structure learning for (binary) Ising models with (modulo-2) additive noise, while in this paper we generalize prior sample complexity bounds [11] to finite and countable alphabets, encompassing non-parametric tree structures and general noise models.

B. Related Applications

Noise-corrupted structured data appear in various applications for major branches of science as physics, computer science, biology and finance. The assumption of discrete and (arbitrarily) large alphabets is suitable for many of these scenarios. Our work is also motivated by a number of emerging applications in machine learning: differential privacy, distributed learning, and adversarial corruption. For example in *local differential privacy* [25]–[30], one approach perturbs each data sample at the time it is collected: our results therefore characterize the increase in sample complexity for successful structure recovery under a given privacy guarantee. In *distributed learning* problems, communication constraints require data to be quantized, resulting in quantization noise on the samples (c.f. [31]). Finally, we might have an *adversarial attack* on the training data: an adversary could corrupt the dataset and make structure learning infeasible. As our example shows, non-i.i.d. corruption can make the information threshold negative, and the CL algorithm will fail without appropriate preprocessing of the corrupted data. We continue by introducing the notation that we use in our paper.

Notation. We denote vectors or tuples by using boldface and we reserve calligraphic face for sets. For an integer M, let $[M] \triangleq \{1, 2, \dots M\}$, and let $[M]^2 \equiv [M] \times [M]$ denote the corresponding Cartesian product. For an odd natural p, we denote the set of even naturals up to $p-1$ as $\mathcal{I}_{p-1} \triangleq \{2, 4, \dots, p-1\}$. The indicator function of a set A is denoted as $\mathbf{1}_A$. The cardinality of the set of nodes $\mathcal{V}$ is assumed to be equal to p, $|\mathcal{V}| = p$. The node variables of the tree are denoted by $\mathbf{X} = (X_1, X_2, \dots, X_p)$. If $\mathbf{X}$ has a finite alphabet we write $\mathbf{X} \in [M]^p$, and $\mathbf{X} \in \mathcal{X}^p$ for countable size alphabets. The probability mass function of $\mathbf{X}$ is denoted as $\mathrm{p}(\cdot)$ and the joint of the pair X_i, X_j as $\mathrm{p}_{i,j}(\cdot,\cdot)$, for any $i, j \in \mathcal{V}$. $\mathrm{T} = (\mathcal{V}, \mathcal{E})$ is a tree (acyclic graph) with set of nodes and edges $\mathcal{V}$ and $\mathcal{E}$ respectively. We denote the unique set of edges in the path of two nodes $u, \bar{u} \in \mathcal{V}$ of T as $\mathrm{path}_{\mathrm{T}}(u, \bar{u})$. The neighborhood of a node $\nu \in \mathcal{V}$ is denoted as $\mathcal{N}_{\mathrm{T}}(\nu)$. $\hat{\mathrm{p}}(\cdot)$ is the estimator of a distribution $\mathrm{p}(\cdot)$. The symbol $\mathbf{X}^{1:n}$ denotes the dataset of n i.i.d samples of $\mathbf{X}$. The estimated mutual information of a pair X_i, X_j is $\hat{I}(X_i; X_j)$ and the resulting tree structure of the CL algorithm is denoted by T^{CL}. The noisy observable is denoted by $\mathbf{Y}$ and we use the symbol † to distinguish between noiseless quantities and the corresponding noisy quantities, for instance $\mathrm{T}^{\mathrm{CL}}_{\dagger}$ is the estimated tree from $\mathbf{Y}^{1:n}$.

II. Model and Problem Statement

First, we provide a complete description of our model including definitions, properties, and assumptions on the underlying distributions.

A. Tree-Structured and Hidden Tree-Structured Models

We consider graphical models over p nodes with variables $\{X_1, X_2, \dots, X_p\} \equiv \mathbf{X}$ and finite or countable alphabet $\mathcal{X}^p$. We assume that the distribution $\mathrm{p}(\cdot)$ of $\mathbf{X}$ is given by a tree-structured *Markov Random Field, (MRF).* Any distribution $\mathrm{p}(\cdot)$ which is Markov with respect to a tree $\mathrm{T} = (\mathcal{V}, \mathcal{E})$ factorizes as [1]

$$\mathrm{p}(\mathbf{x}) = \prod_{i \in V} \mathrm{p}(x_i) \prod_{(i,j) \in \mathcal{E}} \frac{\mathrm{p}(x_i, x_j)}{\mathrm{p}(x_i)\mathrm{p}(x_j)}, \ \mathbf{x} \in \mathcal{X}; \tag{1}$$

we call such distributions $\mathrm{p}(\cdot)$ tree-structured.

The noisy node variables $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_p\}$ are generated by a *randomized mapping* (noisy channel) $\mathcal{F} : \mathcal{X}^p \to \mathcal{Y}^p$. We restrict attention here to mappings that act on each component (not necessarily independently): $\mathcal{F}(\cdot) = \{F_1(\cdot), F_2(\cdot), \dots, F_p(\cdot)\}$ and each $F_i(\cdot) : \mathcal{X} \to \mathcal{Y}$ so $Y_i = F_i(X_i)$ for all $i \in \mathcal{V}$. Let $\mathbb{P}(Y_i = y_i | X_i = x_i)$ be the transition kernel associated with the randomized mapping $F_i(\cdot)$, then the distribution of the output is given by $\mathrm{p}_{\dagger}(y_i) = \sum_{x_i \in \mathcal{X}} \mathbb{P}(Y_i = y_i | X_i = x_i)\mathrm{p}(x_i)$ for $y_i \in \mathcal{Y}$. Note that while the distribution of $\mathbf{X}$ is tree-structured, the distribution of $\mathbf{Y}$ *does not factorize* according to any tree. In general the Markov random field of $\mathbf{Y}$ is a complete graph, which makes learning the hidden structure non-trivial [11], [32].

Given n i.i.d observations $\mathbf{X}^{1:n} \sim \mathrm{p}(\cdot)$, our goal is to learn the underlying structure T. To do this, we use a plug-in estimate of the mutual information $I(X_i; X_j)$ between pairs of variables. In similar fashion, when noise-corrupted observations $\mathbf{Y}^{1:n}$ are available, we aim to learn the hidden tree structure T of $\mathbf{X}$, by estimating the mutual information $I(Y_i; Y_j)$ between pairs of observable variables. Unfortunately, the plug-in mutual information estimate $\hat{I}(X_i; X_j)$ may converge slowly to $I(X_i; X_j)$ in certain cases with countable alphabets [33, Corollary 5], [34]. To avoid such ill-conditioned cases, we make the following assumption.

Assumption 1. *For some $c > 1$ there exist $c_1, c_2 > 0$ such that $c_1/k^c \leq \mathrm{p}_i(k) \leq c_2/k^c$, for $k \in \mathcal{X}$, and $c_1/(k\ell)^c \leq \mathrm{p}_{i,j}(k,\ell) \leq c_2/(k\ell)^c$, for $k,\ell \in \mathcal{X}^2$ and for all $i,j \in \mathcal{V}$. That is, the tuple $\{c, c_1, c_2\}$ satisfies the assumption for all marginal and pairwise joint distributions.*

Assumption 1 holds trivially for finite (fixed) alphabets, where the constants c_1 and c_2 depend on the minimum probability and the size of alphabet. The next assumption guarantees there is a unique tree structure T with exactly p nodes.

Assumption 2. T *is connected; $I(X_i; X_j) > 0$ for all $i,j \in \mathcal{V}$ and the distribution $\mathrm{p}(\cdot)$ of $\mathbf{X}$ is not degenerate.*

Assumption 2 guarantees convergence for the CL algorithm; $\mathrm{T}^{\mathrm{CL}} \to \mathrm{T}$. For a fixed tree T, we use the notation $\mathcal{P}_{\mathrm{T}}(c_1, c_2)$ to denote the set of tree-structured distributions which satisfy Assumption 1 and Assumption 2. In particular, we assume that $X \sim \mathrm{p}(\cdot) \in \mathcal{P}_{\mathrm{T}}(c_1, c_2)$. The set of all trees on p nodes is denoted as $\mathcal{T}$, and we call $\mathcal{P}_{\mathcal{T}}$ the set of all tree-structured distributions that factorize according to (1) for some $\mathrm{T} \in \mathcal{T}$.

B. Chow-Liu Algorithm

We consider the classical version of the algorithm [3], due to the non-parametric nature of our model. Given n i.i.d samples of the node variables, we first find the estimates of the pairwise joint distributions and then we evaluate the plug-in mutual information estimates. Finally, a Maximum Spanning Tree (MST) algorithm (for instance Kruskal's or Prim's algorithm) returns the estimated tree. For the rest of the paper, we refer to Algorithm 1 by explicitly mentioning the input data set; if $\mathcal{D} = \mathbf{X}^{1:n}$, then the input consist of noiseless data and we consider $\boldsymbol{Z} \equiv \mathbf{X}$ (see Algorithm 1), furthermore the estimated structure $\mathrm{T}^{\text{estimate}}$ is denoted by T^{CL}. Equivalently, if $\mathcal{D} = \mathbf{Y}^{1:n}$, then the input consists of noisy data, $\boldsymbol{Z} \equiv \mathbf{Y}$, and the estimated structure $\mathrm{T}^{\text{estimate}}$ is named as $\mathrm{T}^{\mathrm{CL}}_{\dagger}$. We compute T^{CL} and $\mathrm{T}^{\mathrm{CL}}_{\dagger}$ by running the MST algorithm on the edge weights $\{\hat{I}(X_i; X_j) : i,j \in \mathcal{V}\}$ and $\{\hat{I}(Y_i; Y_j) : i,j \in \mathcal{V}\}$ respectively. Note that the estimates T^{CL} and $\mathrm{T}^{\mathrm{CL}}_{\dagger}$ depend on the value n, but for brevity, we write $\mathrm{T}^{\mathrm{CL}} \to \mathrm{T}$ and $\mathrm{T}^{\mathrm{CL}}_{\dagger} \to \mathrm{T}$ instead of $\lim_{n\to\infty} \mathrm{T}^{\mathrm{CL}} = \mathrm{T}$ and $\lim_{n\to\infty} \mathrm{T}^{\mathrm{CL}}_{\dagger} = \mathrm{T}$ respectively. We continue by analyzing the event of incorrect reconstruction $\mathrm{T}^{\mathrm{CL}} \neq \mathrm{T}$ (or $\mathrm{T}^{\mathrm{CL}}_{\dagger} \neq \mathrm{T}$), which yields a sufficient condition for exact structure recovery.

Proposition 1. *The estimated tree $\mathrm{T}^{\mathrm{CL}} \neq \mathrm{T}$ if and only if there exist two edges $e \equiv (w, \bar{w}) \in \mathrm{T}$ and $g \equiv (u, \bar{u}) \in \mathrm{T}^{\mathrm{CL}}$ such that $e \notin \mathrm{T}^{\mathrm{CL}}$, $g \notin \mathrm{T}$ and $e \in \mathrm{path}_{\mathrm{T}}(u, \bar{u})$. Then also $g \in \mathrm{path}_{\mathrm{T}^{\mathrm{CL}}}(w, \bar{w})$.*

Intuitively, exact recovery fails when there is at least one edge in the original tree T which does not appears in the estimated tree T^{CL}. We refer the reader to the proof of Proposition 1 in Bresler and Karzand [5, Section F, "Two trees lemma", Lemma F1]. For sake of space, we define the set $\mathcal{EV}^2$.

Definition 1 (Feasibility set $\mathcal{EV}^2$). *Let $e \equiv (w, \bar{w}) \in \mathcal{E}_{\mathrm{T}}$ be an edge and $u, \bar{u} \in \mathcal{V}_{\mathrm{T}}$ be a pair of nodes such that $e \in \mathrm{path}_{\mathrm{T}}(u, \bar{u})$ and $|\mathrm{path}_{\mathrm{T}}(u, \bar{u})| \geq 2$. The set of all such tuples $(e, u, \bar{u})$, is denoted as $\mathcal{EV}^2$,*

$$\mathcal{EV}^2 \triangleq \{e, u, \bar{u} \in \mathcal{E}_{\mathrm{T}} \times \mathcal{V}_{\mathrm{T}} \times \mathcal{V}_{\mathrm{T}} : e \in \mathrm{path}_{\mathrm{T}}(u, \bar{u}) \text{ and } |\mathrm{path}_{\mathrm{T}}(u, \bar{u})| \geq 2\}. \quad (2)$$

Algorithm 1 Chow-Liu (CL)

Require: Data set $\mathcal{D} = \boldsymbol{Z} \in \mathcal{Z}^{|\mathcal{V}| \times n}$

$\hat{\mathrm{p}}_{i,j}(\ell, m) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\{Z_{i,k}=\ell, Z_{j,k}=m\}}$, $\forall i,j \in \mathcal{V}$

$\hat{I}(Z_i; Z_j) = \sum_{\ell,m} \hat{\mathrm{p}}_{i,j}(\ell, m) \log_2 \frac{\hat{\mathrm{p}}_{i,j}(\ell,m)}{\hat{\mathrm{p}}_i(\ell)\hat{\mathrm{p}}_j(m)}$

$\mathrm{T}^{\text{estimate}} \leftarrow \mathrm{MST}\Big(\{\hat{I}(Z_i; Z_j) : i,j \in \mathcal{V}\}\Big)$

For the rest of the paper the pair of nodes w, $\bar{w}$ denotes the edge $e \equiv (w, \bar{w}) \in \mathcal{E}_{\mathrm{T}}$. The error characterization of CL algorithm is expressed as if $\mathrm{T}^{\mathrm{CL}} \neq \mathrm{T}$ then there exists $((w, \bar{w}), u, \bar{u}) \in \mathcal{EV}^2$ such that[1]

$$\hat{I}(X_w; X_{\bar{w}}) \leq \hat{I}(X_u; X_{\bar{u}}).$$

By negating the above statement, we get that if $\hat{I}(X_w; X_{\bar{w}}) > \hat{I}(X_u; X_{\bar{u}})$ for all $((w, \bar{w}), u, \bar{u}) \in \mathcal{EV}^2$ then

$$\mathrm{T}^{\mathrm{CL}} = \mathrm{T}. \quad (3)$$

The latter yields the sufficient condition for accurate structure estimation.

Sufficient condition for exact structure recovery: For exact structure recovery we need $\hat{I}(X_w; X_{\bar{w}}) > \hat{I}(X_u; X_{\bar{u}})$ for all $((w, \bar{w}), u, \bar{u}) \in \mathcal{EV}^2$, or equivalently

$$I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}}) > \hat{I}(X_u; X_{\bar{u}}) - I(X_u; X_{\bar{u}}) - \hat{I}(X_w; X_{\bar{w}}) + I(X_w; X_{\bar{w}}). \quad (4)$$

The latter allows us to derive a sufficient condition based on the error estimation of the mutual information.

Proposition 2. *If*

$$\left|\hat{I}(X_\ell; X_{\bar{\ell}}) - I(X_\ell; X_{\bar{\ell}})\right| < \frac{1}{2} \min_{(e,u,\bar{u}) \in \mathcal{EV}^2} \{I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}})\}, \quad (5)$$

for all $\ell, \ell' \in \mathcal{V}$ then $\mathrm{T}^{\mathrm{CL}} = \mathrm{T}$.

In fact (5) implies (4) and (4) implies $\mathrm{T}^{\mathrm{CL}} = \mathrm{T}$. Inequality (5) shows that if the error of mutual information estimates is less than a threshold statistic then exact structured recovery is guaranteed.

C. Information Threshold and Properties

We now define a new quantity for tree structured distributions, which we call the *information threshold*. As well will see shortly, our sample complexity bounds for exact structure recovery via the CL algorithm depend on the distribution only through the information threshold, $\mathbf{I}^o$. We first define

[1]The event $\{\hat{I}(Y_w; Y_{\bar{w}}) = \hat{I}(X_u; X_{\bar{u}})\}$ has non zero probability for certain cases, in this situation the MST arbitrarily chooses one of the edges $(w, \bar{w})$, $(u, \bar{u})$. The choice of $(u, \bar{u})$ yields the error event $\mathrm{T}^{\mathrm{CL}} \neq T$.

$\mathbf{I}^o$ and then show how it affects the difficulty of the structure estimation problem.

Definition 2 (Information Threshold (IT)). *Let $e \equiv (w, \bar{w}) \in \mathcal{E}_{\mathrm{T}}$ be an edge and $u, \bar{u} \in \mathcal{V}_{\mathrm{T}}$ be a pair of nodes such that $e \in \mathrm{path}_{\mathrm{T}}(u, \bar{u})$. The information threshold associated with the model $p(\cdot) \in \mathcal{P}_{\mathrm{T}}(c_1, c_2)$ is defined as*

$$\mathbf{I}^o \triangleq \frac{1}{2} \min_{(e,u,\bar{u}) \in \mathcal{EV}^2} \left(I\left(X_w; X_{\bar{w}}\right) - I\left(X_u; X_{\bar{u}}\right) \right). \quad (6)$$

The minimization in (6) is with respect to all tuples $(e, u, \bar{u}) \in \mathcal{EV}^2$, while the values $I(X_i; X_j) > 0$ are considered fixed for all $i, j \in \mathcal{V}$. Note that the definition in (6) comes form the sufficient condition in (5), which implies that if

$$\left| \hat{I}\left(X_\ell; X_{\bar{\ell}}\right) - I\left(X_\ell; X_{\bar{\ell}}\right) \right| < \mathbf{I}^o \quad \forall \ell, \ell' \in \mathcal{V}, \quad (7)$$

then $\mathrm{T} = \mathrm{T}^{\mathrm{CL}}$. The data processing inequality [9] shows that $\mathbf{I}^o \geq 0$. Also, Assumption 2 guarantees that $\mathbf{I}^o > 0$.

Proposition 3 (Positivity). *If Assumption 2 holds then $\mathbf{I}^o > 0$.*

Note that under the reasonable assumption that the values of $I(X_i, X_j)$ for $(i, j) \in \mathcal{E}$ are constant relative to p [7], $\mathbf{I}^o$ does not depend on p. The latter holds because of the locality property of $\mathbf{I}^o$.

Proposition 4 (Locality). *Assume that Assumption 2 holds. Let $(e^*, u^*, \bar{u}^*) \in \mathcal{EV}^2$ be a tuple such that*

$$(e^*, u^*, \bar{u}^*) = \underset{((w,\bar{w}),u,\bar{u}) \in \mathcal{EV}^2}{\arg\min} I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}}), \quad (8)$$

then $u^* \equiv w^*$ or $u^* \equiv \tilde{w}^*$ and $\bar{u} \in \mathcal{N}_{\mathrm{T}}(w)$ or $\bar{u} \in \mathcal{N}_{\mathrm{T}}(\bar{w})$.

We prove Propositions 3 and 4 in Section A, Appendix.

When the data are noisy, the gap between mutual informations that defines the information threshold will change. If the errors of the mutual information estimates of the noisy variables satisfy the condition

$$\left| \hat{I}\left(Y_\ell; Y_{\bar{\ell}}\right) - I\left(Y_\ell; Y_{\bar{\ell}}\right) \right| < \frac{1}{2} \min_{(e,u,\bar{u}) \in \mathcal{EV}^2} \big(I\left(Y_w Y_{\bar{w}}\right) - I\left(Y_u; Y_{\bar{u}}\right) \big), \quad (9)$$

for all $\ell, \bar{\ell} \in \mathcal{V}$ then $\mathrm{T}^{\mathrm{CL}}_{\dagger} = \mathrm{T}$, and (9) is derived similarly to (5). The definition of the *noisy information threshold* naturally results from the previous condition.

Definition 3 (Noisy IT). *The noisy information threshold is defined as*

$$\mathbf{I}^o_{\dagger} \triangleq \frac{1}{2} \min_{(e,u,\bar{u}) \in \mathcal{EV}^2} \big(I\left(Y_w; Y_{\bar{w}}\right) - I\left(Y_u; Y_{\bar{u}}\right) \big). \quad (10)$$

The minimization in (9) and (10) is with respect to the feasible set $\mathcal{EV}^2$ of the *hidden* tree structure T of $\mathbf{X}$. Note that the distribution of $\mathbf{Y}$ does not factorize according to any tree [32].

III. Recovering the Structure from Noiseless Data

First, we develop a computable sample complexity bound for exact noiseless structure learning with high probability. The structure learning condition in (7), combined with results on concentration of mutual information estimators [33], yields the following result.

Theorem 1. *Assume that $\mathbf{X} \sim \mathrm{p}(\cdot) \in \mathcal{P}_{\mathrm{T}}(c_1, c_2)$ for some $c \in (1, 2)$. Fix a number $\delta \in (0, 1)$. There exists a constant $C > 0$, independent of δ such that, if the number of samples of $\mathbf{X}$ satisfies the inequalities*

$$\frac{n}{\log_2^2 n} \geq \frac{72 \log\left(\frac{p}{\delta}\right)}{\left(\mathbf{I}^o - C n^{\frac{1-c}{c}}\right)^2} \;\; \text{and} \;\; \mathbf{I}^o > C n^{\frac{1-c}{c}}, \quad (11)$$

then Algorithm 1 with input $\mathcal{D} = \mathbf{X}^{1:n}$ returns $\mathrm{T}^{\mathrm{CL}} = \mathrm{T}$ with probability at least $1 - \delta$.

Before proceeding with the proof, we would like to provide some remarks. First, the constant C depends on the values of constants c, c_1, c_2 which are defined in Assumption 1. Specifically, $C = 3c_2\big[c_2^{(1-c)/c} + c^{-1} \int_{c_1}^{\infty} u^{1/c-2} \log\left(eu/c_1\right) + 1/c_1\big]$. The derivation of C has been given by Antos and Kontoyiannis [33, Theorem 7]. Additionally, note that $n/\log^2 n = \Omega\left(n^\epsilon\right)$ for any fixed $\epsilon \in (0, 1)$. Therefore, the required number of samples n with respect to p and δ and for fixed $\mathbf{I}^o$ scale as $\mathcal{O}\big(\log^{1+\zeta}(p/\delta)\big)$, for any choice of $\zeta > 0$, whereas, for fixed p and δ, the complexity is of the order of $\mathbf{I}^o$ is $\mathcal{O}\big((\mathbf{I}^o)^{-2(1+\zeta)}\big)$, for any $\zeta > 0$. The proof of Theorem 1 now follows.

Proof of Theorem 1. To calculate the probability of the exact structure recovery we use a concentration inequality quantifying the rate of convergence of entropy estimators from Antos and Kontoyiannis [33]. In particular, they show ([33, Corollary 1]) how the *plug-in* entropy estimator $\hat{H}_n$ (say) is distributed around its mean $\mathbb{E}[\hat{H}_n]$: For every $n \in \mathbb{N}$ and $\epsilon > 0$,

$$\mathbb{P}\left[\left|\hat{H}_n - \mathbb{E}[\hat{H}_n]\right| > \epsilon\right] \leq 2e^{-n\epsilon^2/2\log_2^2 n}. \quad (12)$$

The plug-in entropy estimator is biased and, actually, $H \geq \mathbb{E}[\hat{H}_n]$. Under their Assumption 1, in Theorem 7 [33] they characterize the bias as follows. For $c \in (1, 2)$ (which is the case of interest in this proof),

$$H - \mathbb{E}[\hat{H}_n] = \Theta\big(n^{\frac{1-c}{c}}\big) \quad (13)$$

and for $c \geq 2$,

$$\Omega\big(n^{\frac{1-c}{c}}\big) = H - \mathbb{E}[\hat{H}_n] = \mathcal{O}\big(n^{-1/2} \log n\big). \quad (14)$$

Then, every $\epsilon > C n^{\frac{1-c}{c}}$, it is true that

$$\begin{aligned}
&\mathbb{P}\left[\left|\hat{H}_n - H\right| > \epsilon\right] \\
&= \mathbb{P}\left[\left|\hat{H}_n - \mathbb{E}[\hat{H}_n] + \mathbb{E}[\hat{H}_n] - H\right| > \epsilon\right] \\
&\leq \mathbb{P}\left[\left|\hat{H}_n - \mathbb{E}[\hat{H}_n]\right| + \left|\mathbb{E}[\hat{H}_n] - H\right| > \epsilon\right] \\
&= \mathbb{P}\left[\left|\hat{H}_n - \mathbb{E}[\hat{H}_n]\right| > \epsilon - \left|\mathbb{E}[\hat{H}_n] - H\right|\right] \\
&\leq \mathbb{P}\left[\left|\hat{H}_n - \mathbb{E}[\hat{H}_n]\right| > \epsilon - C n^{\frac{1-c}{c}}\right] \\
&\leq 2e^{-n\left(\epsilon - C n^{\frac{1-c}{c}}\right)^2/2\log_2^2 n},
\end{aligned} \quad (15)$$

where the last inequality comes from (12) and (13). Notice that for non-trivial bounds we need the condition $\epsilon > Cn^{\frac{1-c}{c}}$. Further, ϵ is free parameter and we choose $\epsilon = \mathbf{I}^o/3$, driven by property (7). This requires than n has to be sufficiently large, such that the following inequality holds

$$\mathbf{I}^o > 3Cn^{\frac{1-c}{c}}. \quad (16)$$

Our goal is to find an upper on the probability of the event $\left\{\left|\hat{I}\left(X_\ell; X_{\bar{\ell}}\right) - I\left(X_\ell; X_{\bar{\ell}}\right)\right| > \mathbf{I}^o\right\}$. By combining the above and applying the property $I(X;Y) = H(X) + H(Y) - H(X,Y)$ we have

$$\begin{aligned}
&\mathbb{P}\left[\left|\hat{I}\left(X_\ell; X_{\bar{\ell}}\right) - I\left(X_\ell; X_{\bar{\ell}}\right)\right| > \mathbf{I}^o\right] \\
&= \mathbb{P}\Big[\Big|\hat{H}(X_\ell) + \hat{H}(X_{\bar{\ell}}) - \hat{H}(X_\ell, X_{\bar{\ell}}) \\
&\qquad - H(X_\ell) - H(X_{\bar{\ell}}) + H(X_\ell, X_{\bar{\ell}})\Big| > \mathbf{I}^o\Big] \\
&\leq \mathbb{P}\Big[\left|\hat{H}(X_\ell) - H(X_\ell)\right| + \left|\hat{H}(X_{\bar{\ell}}) - H(X_{\bar{\ell}})\right| \\
&\qquad + \left|H(X_\ell, X_{\bar{\ell}}) - \hat{H}(X_\ell, X_{\bar{\ell}})\right| > \mathbf{I}^o\Big] \\
&\leq \mathbb{P}\Bigg[\left\{\left|\hat{H}(X_\ell) - H(X_\ell)\right| > \frac{\mathbf{I}^o}{3}\right\} \\
&\qquad \bigcup \left\{\left|\hat{H}(X_{\bar{\ell}}) - H(X_{\bar{\ell}})\right| > \frac{\mathbf{I}^o}{3}\right\} \\
&\qquad \bigcup \left\{\left|H(X_\ell, X_{\bar{\ell}}) - \hat{H}(X_\ell, X_{\bar{\ell}})\right| > \frac{\mathbf{I}^o}{3}\right\}\Bigg] \\
&\leq 6e^{-n\left(\frac{\mathbf{I}^o}{3} - Cn^{\frac{1-c}{c}}\right)^2/2\log_2^2 n}, \quad (17)
\end{aligned}$$

and the last inequality is a consequence of (15). To guarantee that the condition in (7) holds, we apply the union bound on the events $\left\{\left|\hat{I}\left(X_\ell; X_{\bar{\ell}}\right) - I\left(X_\ell; X_{\bar{\ell}}\right)\right| > \mathbf{I}^o\right\}$, for all $\ell, \bar{\ell} \in \mathcal{V}$. Since there are $\binom{p}{2}$ pairs we define

$$\delta \triangleq \binom{p}{2} 6e^{-n\left(\frac{\mathbf{I}^o}{3} - Cn^{\frac{1-c}{c}}\right)^2/2\log_2^2 n}. \quad (18)$$

To conclude, for some fixed $\delta \in (0,1)$ if

$$\frac{n}{\log_2^2 n} \geq \frac{2\log\left(\frac{6\binom{p}{2}}{\delta}\right)}{\left(\frac{\mathbf{I}^o}{3} - Cn^{\frac{1-c}{c}}\right)^2} \quad \text{and} \quad \mathbf{I}^o > 3Cn^{\frac{1-c}{c}}, \quad (19)$$

then the probability of exact recovery is at least $1-\delta$. The latter combined with the inequalities $8\log(p/\delta) > 2\log\left(6\binom{p}{2}/\delta\right)$, $p \geq 3$ completes the proof. □

Theorem 1 characterizes the sample complexity for models with *either* countable *or* finite alphabets. By restricting our setting to finite alphabets, it follows that we may choose $c = 2$ in Theorem 1, and we have the following result. The proof is virtually identical to that of Theorem 1, and is omitted.

Theorem 2. *Assume that the random variable* $\mathbf{X}$ *has finite support* ($|\mathcal{X}| < \infty$) *and* $\mathbf{X} \sim \mathrm{p}(\cdot) \in \mathcal{P}_{\mathrm{T}}(c_1, c_2)$. *Fix a number* $\delta \in (0,1)$. *There exists a constant* $C > 0$*, independent of* δ *such that, if the number of samples of* $\mathbf{X}$ *satisfies the inequalities*

$$\frac{n}{\log_2^2 n} \geq \frac{72\log\left(\frac{p}{\delta}\right)}{\left(\mathbf{I}^o - C\frac{1}{\sqrt{n}}\right)^2} \quad \textit{and} \quad \mathbf{I}^o > C\frac{1}{\sqrt{n}}, \quad (20)$$

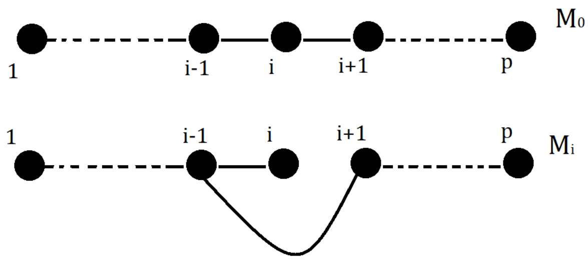

Fig. 2. Construction of the set of tree-structured models for the proof of Theorem 3. The structure of each model M_i, $i \in \mathcal{I}_{p-1}$ differs with M_0 in one edge.

then Algorithm 1 with input $\mathcal{D} = \mathbf{X}^{1:n}$ *returns* $\mathrm{T}^{\mathrm{CL}} = \mathrm{T}$ *with probability at least* $1-\delta$.

Lastly, the corresponding variation of Theorem 1 when Assumption 1 holds for $c \geq 2$ (in the general case of countable alphabets) may be found in Appendix A-C (Theorem 5).

IV. Converse: Information Threshold as a Fundamental Quantity

In this section we provide the statement of the converse of our main result Theorem 1. For some $\eta > 0$, we define the class of all tree-structured models $\mathrm{M} \in \mathcal{P}_{\mathcal{T}}$ with positive, η-bounded information threshold as

$$\mathcal{C}_\eta^{\mathcal{T}} \triangleq \{\mathrm{M} : 0 < \mathbf{I}_{\mathrm{M}}^o \leq \eta\}. \quad (21)$$

Theorem 1 shows that the structure estimation problem is hard for the CL algorithm when $\mathbf{I}^o \to 0$. Our next result shows that this is also true for any other algorithm whatsoever.

Theorem 3. *For any finite* $n < \infty$*, and for any estimator* $\Phi : \mathbf{X}^{1:n} \to \mathcal{T}$*, the worst-case probability of incorrect structure recovery over all tree-structured models in* $\mathcal{C}_\eta^{\mathcal{T}}$*, as* $\eta \downarrow 0$*, is at least* $1/2$. *In other words, it is true that*

$$\inf_{\Phi:\mathbf{X}^{1:n}\to\mathcal{T}} \lim_{\eta\downarrow 0} \sup_{\mathrm{M}\in\mathcal{C}_\eta^{\mathcal{T}}} \mathbb{P}\left(\Phi(\mathbf{X}_{\mathrm{T_M}}^{1:n}) \neq \mathrm{T_M}\right) \geq \frac{1}{2}. \quad (22)$$

Proof of Theorem 3. We define a collection of $(p+1)/2$ Markov models $\mathrm{M}_0, \mathrm{M}_2, \mathrm{M}_4, \ldots, \mathrm{M}_{p-1}$ on p variables (p is odd integer greater than 1), each taking values in a common finite set $\mathcal{X}$. The M_0 model is Markov chain. For each even $i > 0$, the structure of the M_i differs from the structure of M_0 in one edge, as illustrated in Figure 2. We denote the distributions of each model M_i for $i \geq 0$ by $\mathrm{p}_{\mathrm{M}_i}(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}^p$. We denote the sets of edges as $\mathcal{E}_{\mathrm{M}_i}$. Further the marginal distributions $\mathrm{p}(\cdot)$ are uniform for all the models and the node variables, and the pair-wise marginals are denoted as $\mathrm{p}_{\mathrm{M}_i}(x_s, x_t)$. Thus the distribution of M_0 is

$$\mathrm{p}_{\mathrm{M}_0}(\mathbf{x}) = \prod_{s\in V} \mathrm{p}\left(x_s\right) \prod_{(s,t)\in\mathcal{E}_{\mathrm{M}_0}} \frac{\mathrm{p}_{\mathrm{M}_0}(x_s, x_t)}{\mathrm{p}(x_s)\mathrm{p}(x_t)}, \ \mathbf{x} \in \mathcal{X}, \quad (23)$$

and the distribution of M_i for $i \in \mathcal{I}_{p-1}$ is

$$\mathrm{p}_{\mathrm{M}_i}(\mathbf{x}) = \prod_{s\in V} \mathrm{p}\left(x_s\right) \prod_{(s,t)\in\mathcal{E}_{\mathrm{M}_i}} \frac{\mathrm{p}_{\mathrm{M}_i}(x_s, x_t)}{\mathrm{p}(x_s)\mathrm{p}(x_t)}, \ \mathbf{x} \in \mathcal{X}. \quad (24)$$

Recall that $\mathcal{I}_{p-1} \equiv \{2, 4, \dots, p-1\}$. In model M_0, the pair-wise marginal distributions $\mathrm{p}_{\mathrm{M}_0}(x_i, x_{i-1})$ coincide for all $i \in \mathcal{I}_{p-1}$ and the marginal distributions $\mathrm{p}_{\mathrm{M}_0}(x_{i+1}, x_i)$ coincide for all $i \in \mathcal{I}_{p-1}$, i.e.,

$$\mathrm{p}_{\mathrm{M}_0}(x_i, x_{i-1}) \equiv \mathrm{p}_{\mathrm{M}_0}(x_2, x_1),\ \forall i \in \mathcal{I}_{p-1} \quad \text{and} \tag{25}$$
$$\mathrm{p}_{\mathrm{M}_0}(x_{i+1}, x_i) \equiv \mathrm{p}_{\mathrm{M}_0}(x_3, x_2),\ \forall i \in \mathcal{I}_{p-1}. \tag{26}$$

Further, we may choose the distributions $\mathrm{p}_{\mathrm{M}_0}(x_3, x_2)$ and $\mathrm{p}_{\mathrm{M}_0}(x_2, x_1)$ to satisfy[2]

$$0 < I_{\mathrm{M}_0}(X_3; X_2) < I_{\mathrm{M}_0}(X_2; X_1). \tag{27}$$

Note that pair-wise marginal distributions that satisfy (27) exist. We construct the set of models M_i ($i \in \mathcal{I}_{p-1}$) in terms of M_0: The pairwise marginals for the common parts coincide, i.e.,

$$\mathrm{p}_{\mathrm{M}_i}(x_{k+1}, x_k) = \mathrm{p}_{\mathrm{M}_0}(x_{k+1}, x_k),\ k \in [p-1] \setminus i, \tag{28}$$

whereas, for the distributions near the variable X_i, we set

$$p_{\mathrm{M}_i}(x_{i+1}, x_{i-1}) = p_{\mathrm{M}_0}(x_{i+1}, x_{i-1}). \tag{29}$$

The construction described above exhibits a number of key features, as follows. First, we show that

$$\begin{aligned}
&\boldsymbol{D}_{\mathrm{KL}}(\mathrm{p}_{\mathrm{M}_0}(\mathbf{X}), \mathrm{p}_{\mathrm{M}_i}(\mathbf{X})) \\
&= I_{\mathrm{M}_0}(X_{i+1}; X_i) - I_{\mathrm{M}_0}(X_{i+1}; X_{i-1}) && (30)\\
&= I_{\mathrm{M}_0}(X_3; X_2) - I_{\mathrm{M}_0}(X_3; X_1), && (31)
\end{aligned}$$

for all $i \in \mathcal{I}_{p-1}$. The right-hand side of (31) comes from (25) and (26), and we find (30) by using (23), (24), (28), (29) and the definition of the KL divergence as

$$\begin{aligned}
&\boldsymbol{D}_{\mathrm{KL}}(\mathrm{p}_{\mathrm{M}_0}(\mathbf{X}), \mathrm{p}_{\mathrm{M}_i}(\mathbf{X})) \\
&= \sum_{\mathbf{x} \in \mathcal{X}} \mathrm{p}_{\mathrm{M}_0}(\mathbf{x}) \log \frac{\mathrm{p}_{\mathrm{M}_0}(\mathbf{x})}{\mathrm{p}_{\mathrm{M}_i}(\mathbf{x})} \\
&= \sum_{\mathbf{x} \in \mathcal{X}} \mathrm{p}_{\mathrm{M}_0}(\mathbf{x}) \log \frac{\mathrm{p}_{\mathrm{M}_0}(x_{i+1}|x_i)}{\mathrm{p}_{\mathrm{M}_i}(x_{i+1}|x_{i-1})} \\
&= \sum_{x_{i-1}, x_i, x_{i+1}} \mathrm{p}_{\mathrm{M}_0}(x_{i-1}, x_i, x_{i+1}) \log \frac{\mathrm{p}_{\mathrm{M}_0}(x_{i+1}|x_i)}{\mathrm{p}_{\mathrm{M}_i}(x_{i+1}|x_{i-1})} \\
&= \sum_{x_{i-1}, x_i, x_{i+1}} \mathrm{p}_{\mathrm{M}_0}(x_{i-1}, x_i, x_{i+1}) \log \frac{\mathrm{p}_{\mathrm{M}_0}(x_{i+1}|x_i)}{\mathrm{p}_{\mathrm{M}_0}(x_{i+1}|x_{i-1})} \\
&= \sum_{x_i, x_{i+1}} \mathrm{p}_{\mathrm{M}_0}(x_i, x_{i+1}) \log \frac{\mathrm{p}_{\mathrm{M}_0}(x_{i+1}, x_i)}{\mathrm{p}_{\mathrm{M}_0}(x_i)} \\
&\quad - \sum_{x_{i-1}, x_{i+1}} \mathrm{p}_{\mathrm{M}_0}(x_{i-1}, x_{i+1}) \log \frac{\mathrm{p}_{\mathrm{M}_0}(x_{i+1}, x_{i-1})}{\mathrm{p}_{\mathrm{M}_0}(x_{i-1})} \\
&= \sum_{x_i, x_{i+1}} \mathrm{p}_{\mathrm{M}_0}(x_{i+1}, x_i) \log \frac{\mathrm{p}_{\mathrm{M}_0}(x_{i+1}, x_i)}{\mathrm{p}_{\mathrm{M}_0}(x_{i+1})\mathrm{p}_{\mathrm{M}_0}(x_i)} \\
&\quad - \sum_{x_{i-1}, x_{i+1}} \mathrm{p}_{\mathrm{M}_0}(x_{i-1}, x_{i+1}) \log \frac{\mathrm{p}_{\mathrm{M}_0}(x_{i+1}, x_{i-1})}{\mathrm{p}_{\mathrm{M}_0}(x_{i+1})\mathrm{p}_{\mathrm{M}_0}(x_{i-1})} \\
&= I(X_{i+1}; X_i) - I(X_{i+1}; X_{i-1}). && (32)
\end{aligned}$$

[2]For the rest of this section we denote by $I_{\mathrm{M}_0}(\cdot\,;\cdot)$ the mutual information with respect to the distribution of model M_0.

We then combine (25), (26), (27), the locality of information threshold (Proposition 4) and Definition 6 to show that

$$\mathbf{I}^o_{\mathrm{M}_0} = \frac{1}{2}\left(I_{\mathrm{M}_0}(X_3; X_2) - I_{\mathrm{M}_0}(X_3; X_1)\right). \tag{33}$$

The latter and (31) give that for all $i \in \mathcal{I}_{p-1}$

$$\boldsymbol{D}_{\mathrm{KL}}(\mathrm{p}_{\mathrm{M}_0}(\mathbf{X}), \mathrm{p}_{\mathrm{M}_i}(\mathbf{X})) = 2\mathbf{I}^o_{\mathrm{M}_0}. \tag{34}$$

Similarly we can find the information threshold of the models M_i for $i \geq 2$,

$$\begin{aligned}
\mathbf{I}^o_{\mathrm{M}_i} &= \frac{1}{2}\left(I_{\mathrm{M}_i}(X_{i+1}; X_{i-1}) - I_{\mathrm{M}_i}(X_{i+1}; X_i)\right) \\
&= \frac{1}{2}\left(I_{\mathrm{M}_i}(X_3; X_1) - I_{\mathrm{M}_i}(X_3; X_2)\right). && (35)
\end{aligned}$$

Further, we show that

$$\begin{aligned}
&\boldsymbol{D}_{\mathrm{KL}}\left(\mathrm{p}_{\mathrm{M}_i}(\mathbf{X})||\mathrm{p}_{\mathrm{M}_0}(\mathbf{X})\right) \\
&= 2\mathbf{I}^o_{\mathrm{M}_i} + \boldsymbol{D}_{\mathrm{KL}}\left(\mathrm{p}_{\mathrm{M}_i}(x_{i+1}, x_i)||\mathrm{p}_{\mathrm{M}_0}(x_{i+1}, x_i)\right) && (36)
\end{aligned}$$

by direct evaluation, that is,

$$\begin{aligned}
&\boldsymbol{D}_{\mathrm{KL}}(\mathrm{p}_{\mathrm{M}_i}(\mathbf{X})||\mathrm{p}_{\mathrm{M}_0}(\mathbf{X})) - \boldsymbol{D}_{\mathrm{KL}}(\mathrm{p}_{\mathrm{M}_i}(x_{i+1}, x_i)||\mathrm{p}_{\mathrm{M}_0}(x_{i+1}, x_i)) \\
&= \sum_{x_{i-1}, x_i, x_{i+1}} \mathrm{p}_{\mathrm{M}_i}(x_{i-1}, x_i, x_{i+1}) \log \frac{\mathrm{p}_{\mathrm{M}_i}(x_{i-1}, x_i, x_{i+1})}{\mathrm{p}_{\mathrm{M}_0}(x_{i-1}, x_i, x_{i+1})} \\
&\quad - \sum_{x_i, x_{i+1}} \mathrm{p}_{\mathrm{M}_i}(x_i, x_{i+1}) \log \frac{\mathrm{p}_{\mathrm{M}_i}(x_i, x_{i+1})}{\mathrm{p}_{\mathrm{M}_0}(x_i, x_{i+1})} \\
&= \sum_{x_{i-1}, x_i, x_{i+1}} \mathrm{p}_{\mathrm{M}_i}(x_{i-1}, x_i, x_{i+1}) \log \frac{\mathrm{p}_{\mathrm{M}_i}(x_{i+1}|x_{i-1})}{\mathrm{p}_{\mathrm{M}_0}(x_{i+1}|x_i)} \\
&\quad - \sum_{x_{i-1}, x_i, x_{i+1}} \mathrm{p}_{\mathrm{M}_i}(x_{i-1}, x_i, x_{i+1}) \log \frac{\mathrm{p}_{\mathrm{M}_i}(x_i, x_{i+1})}{\mathrm{p}_{\mathrm{M}_0}(x_i, x_{i+1})} \\
&= \sum_{x_{i-1}, x_i, x_{i+1}} \mathrm{p}_{\mathrm{M}_i}(x_{i-1}, x_i, x_{i+1}) \log \frac{\mathrm{p}_{\mathrm{M}_i}(x_{i+1}, x_{i-1})}{\mathrm{p}_{\mathrm{M}_0}(x_{i+1}, x_i)} \\
&\quad - \sum_{x_{i-1}, x_i, x_{i+1}} \mathrm{p}_{\mathrm{M}_i}(x_{i-1}, x_i, x_{i+1}) \log \frac{\mathrm{p}_{\mathrm{M}_i}(x_i, x_{i+1})}{\mathrm{p}_{\mathrm{M}_0}(x_i, x_{i+1})} \\
&= \sum_{x_{i-1}, x_i, x_{i+1}} \mathrm{p}_{\mathrm{M}_i}(x_{i-1}, x_i, x_{i+1}) \log \mathrm{p}_{\mathrm{M}_i}(x_{i+1}, x_{i-1}) \\
&\quad - \sum_{x_{i-1}, x_i, x_{i+1}} \mathrm{p}_{\mathrm{M}_i}(x_{i-1}, x_i, x_{i+1}) \log \mathrm{p}_{\mathrm{M}_i}(x_i, x_{i+1}) \\
&= \sum_{x_{i-1}, x_i, x_{i+1}} \mathrm{p}_{\mathrm{M}_i}(x_{i-1}, x_i, x_{i+1}) \log \frac{\mathrm{p}_{\mathrm{M}_i}(x_{i+1}, x_{i-1})}{\frac{1}{|\mathcal{X}|}\frac{1}{|\mathcal{X}|}} \\
&\quad - \sum_{x_{i-1}, x_i, x_{i+1}} \mathrm{p}_{\mathrm{M}_i}(x_{i-1}, x_i, x_{i+1}) \log \frac{\mathrm{p}_{\mathrm{M}_i}(x_i, x_{i+1})}{\frac{1}{|\mathcal{X}|}\frac{1}{|\mathcal{X}|}} \\
&= 2\mathbf{I}^o_{\mathrm{M}_i}. && (37)
\end{aligned}$$

Next, note that for the collection of models in Figure 2, if for model M_0 it holds that $X_i \perp\!\!\!\perp X_{i+1}$, $i \in \mathcal{I}_{p-1}$ or if there is an injective mapping $G(\cdot)$ such that $X_{i-1} = G(X_i)$, $i \in \mathcal{I}_{p-1}$, then $\boldsymbol{D}_{\mathrm{KL}}(\mathrm{p}_{\mathrm{M}^l_0}(\mathbf{X})||\mathrm{p}_{\mathrm{M}^l_i}(\mathbf{X})) = 0$. We denote the two cases by (A_1) and (A_2):

(A_1) $\mathrm{p}_{\mathrm{M}_0}(x_i, x_{i+1}) = \mathrm{p}_{\mathrm{M}_0}(x_i)\mathrm{p}_{\mathrm{M}_0}(x_{i+1})$, $\forall i \in \mathcal{I}_{p-1}$.

(A_2) $\mathrm{p}_{\mathrm{M}_0}(x_{i-1}, x_i) = \mathrm{p}_{\mathrm{M}_0}(x_i) \times \mathbf{1}_{x_{i-1}=G(x_i)}$, $\forall i \in \mathcal{I}_{p-1}$.

We study the marginal case of (A_1) by considering a sequence of models M^ℓ_0 distributions $\mathrm{p}^\ell_{\mathrm{M}_0}(x_i, x_{i+1})$, such that

$$\lim_{\ell \to \infty} \mathrm{p}_{\mathrm{M}^\ell_0}(x_i, x_{i+1}) \to \mathrm{p}_{\mathrm{M}_0}(x_i)\mathrm{p}_{\mathrm{M}_0}(x_{i+1}), \tag{38}$$

while models M_i^ℓ, $i \in \mathcal{I}_{p-1}$ are constructed according to Figure 2, (28) and (29). Case (A_2) would work as well; we explain such an approach with a remark right after the end of the proof. Observe that $\boldsymbol{D}_{\mathrm{KL}}(\mathrm{p}_{\mathrm{M}_0^l}(\mathbf{X})||\mathrm{p}_{\mathrm{M}_i^l}(\mathbf{X})) \to 0$ and $\boldsymbol{D}_{\mathrm{KL}}(\mathrm{p}_{\mathrm{M}_i^l}(\mathbf{X})||\mathrm{p}_{\mathrm{M}_0^l}(\mathbf{X})) \to 0$ as $\ell \to \infty$, for all $i \in \mathcal{I}_{p-1}$. Also, by our discussion above, for every $\ell \in \mathbb{N}$, it is true that

$$\boldsymbol{D}_{\mathrm{KL}}(\mathrm{p}_{\mathrm{M}_0^l}(\mathbf{X})||\mathrm{p}_{\mathrm{M}_i^l}(\mathbf{X})) = 2\mathbf{I}^o_{\mathrm{M}_0^l}, \tag{39}$$

and

$$\begin{aligned}&\boldsymbol{D}_{\mathrm{KL}}(\mathrm{p}_{\mathrm{M}_i^l}(\mathbf{X})||\mathrm{p}_{\mathrm{M}_0^l}(\mathbf{X}))\\ &= 2\mathbf{I}^o_{\mathrm{M}_i^l} + \boldsymbol{D}_{\mathrm{KL}}(\mathrm{p}_{\mathrm{M}_i^l}(x_{i+1},x_i)||\mathrm{p}_{\mathrm{M}_0^l}(x_{i+1},x_i)).\end{aligned} \tag{40}$$

From the existence of the limit of the KL divergence under consideration at $\ell \equiv \infty$, it follows that, for any choice of $\eta > 0$, there exist $\ell_1^\star \in \mathbb{N}$ and $\ell_2^\star \in \mathbb{N}$, such that, for every $\ell \geq \max\{\ell_1^\star, \ell_2^\star\} \triangleq \ell^\star$, the inequalities

$$\boldsymbol{D}_{\mathrm{KL}}(\mathrm{p}_{\mathrm{M}_0^l}(\mathbf{X})||\mathrm{p}_{\mathrm{M}_i^l}(\mathbf{X})) < 2\eta \quad \text{and} \tag{41}$$

$$\boldsymbol{D}_{\mathrm{KL}}(\mathrm{p}_{\mathrm{M}_i^l}(\mathbf{X})||\mathrm{p}_{\mathrm{M}_0^l}(\mathbf{X})) < 2\eta \tag{42}$$

hold *simultaneously*. Hereafter, let $\eta > 0$ be associated with the class of tree-structured models $\mathcal{C}_\eta^{\mathcal{T}}$, as defined in the beginning of this section. Then, due to (39) and (40), we also obtain that

$$\begin{aligned}&0 < 2\mathbf{I}^o_{\mathrm{M}_0^{l^\star}} < 2\eta \quad \text{and}\\ &0 < 2\mathbf{I}^o_{\mathrm{M}_i^{l^\star}} < 2\eta,\end{aligned} \tag{43}$$

implying that $\mathrm{M}_i^{l^\star} \in \mathcal{C}_\eta^{\mathcal{T}}$, for all $i \in \mathcal{I}_{p-1} \cup \{0\}$.

Next we apply Fano's inequality: For fixed $\delta \in (0,1)$, as long as

$$n < (1-\delta)\frac{\log((p+1)/2)}{2\eta}, \tag{44}$$

it is true that (note that all constructed models are in $\mathcal{C}_\eta^{\mathcal{T}}$)

$$\begin{aligned}&\inf_{\Phi} \sup_{\mathrm{M}\in\mathcal{C}_\eta^{\mathcal{T}}} \mathbb{P}\big(\Phi(\mathbf{X}^{1:n}_{\mathrm{T_M}}) \neq \mathrm{T_M}\big)\\ &\geq \inf_{\Phi} \max_{i\in\mathcal{I}_{p-1}\cup\{0\}} \mathbb{P}\big(\Phi(\mathbf{X}^{1:n}_{\mathrm{T}_{\mathrm{M}_i^{l^\star}}}) \neq \mathrm{T}_{\mathrm{M}_i^{l^\star}}\big)\\ &\geq \delta - \frac{1}{\log((p+1)/2)}.\end{aligned} \tag{45}$$

By choosing δ appropriately as (see Lemma 1 in Appendix B)

$$\delta \equiv \frac{1}{2} + \frac{1}{\log((p+1)/2)}, \tag{46}$$

we get that, as long as

$$n < \frac{\log((p+1)/2) - 2}{4\eta}, \tag{47}$$

the minimax lower bound

$$\inf_{\Phi} \sup_{\mathrm{M}\in\mathcal{C}_\eta^{\mathcal{T}}} \mathbb{P}\big(\Phi(\mathbf{X}^{1:n}_{\mathrm{T_M}}) \neq \mathrm{T_M}\big) \geq \frac{1}{2} \tag{48}$$

holds. In fact, for every fixed $\eta > 0$, as long as (47) holds, it is also true that

$$\inf_{\Phi} \sup_{\mathrm{M}\in\mathcal{C}_{\eta'}^{\mathcal{T}}} \mathbb{P}\big(\Phi(\mathbf{X}^{1:n}_{\mathrm{T_M}}) \neq \mathrm{T_M}\big) \geq \frac{1}{2}, \quad \forall \eta' \in (0,\eta], \tag{49}$$

where (47) constitutes a *conservative but uniform* upper bound on n, for all $\eta' \in (0,\eta]$. Additionally, it must be the case that

$$\begin{aligned}\frac{1}{2} &\leq \inf_{\eta'\in(0,\eta]} \inf_{\Phi} \sup_{\mathrm{M}\in\mathcal{C}_{\eta'}^{\mathcal{T}}} \mathbb{P}\big(\Phi(\mathbf{X}^{1:n}_{\mathrm{T_M}}) \neq \mathrm{T_M}\big)\\ &\equiv \inf_{\Phi} \inf_{\eta'\in(0,\eta]} \sup_{\mathrm{M}\in\mathcal{C}_{\eta'}^{\mathcal{T}}} \mathbb{P}\big(\Phi(\mathbf{X}^{1:n}_{\mathrm{T_M}}) \neq \mathrm{T_M}\big)\\ &\equiv \inf_{\Phi} \inf_{\eta>0} \sup_{\mathrm{M}\in\mathcal{C}_{\eta}^{\mathcal{T}}} \mathbb{P}\big(\Phi(\mathbf{X}^{1:n}_{\mathrm{T_M}}) \neq \mathrm{T_M}\big)\\ &= \inf_{\Phi} \lim_{\eta\downarrow 0} \sup_{\mathrm{M}\in\mathcal{C}_{\eta}^{\mathcal{T}}} \mathbb{P}\big(\Phi(\mathbf{X}^{1:n}_{\mathrm{T_M}}) \neq \mathrm{T_M}\big),\end{aligned} \tag{50}$$

where the last equality follows as a consequence of Lemma 2 in Appendix B, after exploiting the fact that for every $\eta \geq \eta' > 0$, the inclusion $\mathcal{C}_\eta^{\mathcal{T}} \supseteq \mathcal{C}_{\eta'}^{\mathcal{T}}$ holds. But since η is arbitrary and the right-hand side of (50) is independent of η, the inequality above holds as long as

$$n < \sup_{\eta>0} \frac{\log((p+1)/2) - 2}{4\eta} \equiv \infty. \tag{51}$$

Therefore, we have shown that for every finite number of samples $n < \infty$, the minimax bound

$$\inf_{\Phi} \lim_{\eta\downarrow 0} \sup_{\mathrm{M}\in\mathcal{C}_{\eta}^{\mathcal{T}}} \mathbb{P}\big(\Phi(\mathbf{X}^{1:n}_{\mathrm{T_M}}) \neq \mathrm{T_M}\big) \geq \frac{1}{2} \tag{52}$$

is true. The proof is now complete. □

Remark 1. *The marginal case of (A_2) can be considered as follows. Instead of using the sequence of distributions in* (38)*, we consider a sequence $\mathrm{p}_{\mathrm{M}_0^\ell}(x_{i-1},x_i)$ such that*

$$\lim_{\ell\to\infty} \mathrm{p}_{\mathrm{M}_0^\ell}(x_{i-1},x_i) \to \mathrm{p}_{\mathrm{M}_0}(x_i) \times \mathbf{1}_{x_{i-1}=G(x_i)} \tag{53}$$

for all $i \in \mathcal{I}_{p-1}$. This alternative condition also guarantees that, as $\ell \to \infty$, $\boldsymbol{D}_{\mathrm{KL}}(\mathrm{p}_{\mathrm{M}_0^l}(\mathbf{X})||\mathrm{p}_{\mathrm{M}_i^l}(\mathbf{X})) \to 0$ and $\boldsymbol{D}_{\mathrm{KL}}(\mathrm{p}_{\mathrm{M}_i^l}(\mathbf{X})||\mathrm{p}_{\mathrm{M}_0^l}(\mathbf{X})) \to 0$, for all $i \in \mathcal{I}_{p-1}$. In fact, the sequences in (38) *and* (53) *can be replaced by any other sequence of distributions related to the models under consideration, provided that the divergences above converge to zero, as above. We explicitly refer to cases (A_1), (A_2) to explicitly clarify that such sequences exist.*

V. Recovering the Structure from Noisy Data

In this section we study the problem of learning hidden tree structures from noise-corrupted observations. We consider Algorithm 1 with input $\mathcal{D} = \mathbf{Y}^{1:n}$. Recall that the hidden model follows the general construction of Section II-A. Notice that the distribution $\mathrm{p}_\dagger(\cdot)$ of $\mathbf{Y}$ does not factorize according to any tree structure [11], [32]. Specifically, the MRF of $\mathbf{Y}$ is a complete graph. Therefore the problem of learning hidden tree-structures from the noisy observable $\mathbf{Y}$ is not trivial. Nevertheless, if $\mathbf{I}^o_\dagger > 0$, then we are still able to learn the tree T of the hidden variables $\mathbf{X}$. Recall that we defined the noisy information threshold $\mathbf{I}^o_\dagger$ in (10), that appears in the condition of exact structure recovery of CL algorithm (9), similarly to the noiseless setting. The latter shows that structure recovery is feasible if $\mathbf{I}^o_\dagger > 0$. To be more precise, for every $\mathbf{I}^o_\dagger > 0$ there exists $N \in \mathbb{N}$ such that if $n > N$ then (9) holds with high probability. Under the assumption of $\mathbf{I}^o > 0$ (see Assumption 2), it is not guaranteed that $\mathbf{I}^o_\dagger > 0$. In fact, if $\mathbf{I}^o_\dagger < 0$ then the CL is not consistent in the sense that

$$\mathbb{P}\left(\lim_{n\to\infty} \mathcal{E}_{\mathrm{T}^{\mathrm{CL}}_\dagger(\mathcal{D}=\mathbf{Y}^{1:n})} \neq \mathcal{E}_{\mathrm{T}}\right) = 1, \tag{54}$$

(see also Section V-B). On the other hand if $\mathbf{I}^o_\dagger = 0$, then ties are broken arbitrarily and the probability of missing an edge does not decrease as n increases.

In what follows, we provide finite sample complexity bounds for the CL algorithm with noisy input data. Also, we provide sufficient conditions which ensure that $\mathbf{I}^o_\dagger > 0$; in particular, whenever $\mathbf{I}^o_\dagger < 0$, we show that pre-processing on the the noisy data can be used as an extra step to overcome the inconsistency of the CL algorithm on the original noisy data. The importance of the results lies in the fact that we can simply run the CL algorithm on the noisy observations in the same way as we would do if noiseless observations $\mathbf{X}^{1:n}$ were given. Indeed, *the model of the noise might be unknown, or we might be unaware of the existence of the noise altogether; still, we can learn the hidden structure efficiently.*

A. Sample Complexity of Hidden Tree-Structure Learning

The next result characterizes the sufficient number of samples for exact structure structure recovery $\mathrm{T}^{\mathrm{CL}}_\dagger = \mathrm{T}$ with probability at least $1 - \delta \in (0, 1)$.

Theorem 4. *Assume that $\mathbf{X} \sim \mathrm{p}(\cdot) \in \mathcal{P}_{\mathrm{T}}(c_1, c_2)$. Assume that noisy data $\mathbf{Y} \sim \mathrm{p}_\dagger(\cdot)$ are generated by a randomized set of mappings $\mathcal{F} = \{F_i : i \in [p]\}$, and $\mathrm{p}_\dagger(\cdot)$ satisfies the Assumption 1 for some $c' \in (1, 2), c'_1 > c'_2 > 0$. Fix $\delta \in (0, 1)$. There exists a constant $C' > 0$ independent of δ such that, if $\mathbf{I}^o_\dagger > 0$ and the number of samples n of $\mathbf{Y}$ satisfies the inequalities*

$$\frac{n}{\log_2^2 n} \geq \frac{72 \log\left(\frac{p}{\delta}\right)}{\left(\mathbf{I}^o_\dagger - C' n^{\frac{1-c'}{c'}}\right)^2} \quad \text{and} \quad \mathbf{I}^o_\dagger > C' n^{\frac{1-c'}{c'}}, \tag{55}$$

then Algorithm 1 with input the noisy data $\mathcal{D} = \mathbf{Y}^{1:n}$ returns $\mathrm{T}^{CL}_\dagger = \mathrm{T}$ with probability at least $1 - \delta$. The relationship between C' and c', c'_1, c'_2 is identical to that between C and c, c_1, c_2 in Theorem 1.

Proof of Theorem 4. The proof is similar to the proof of Theorem 1. The difference is introduced by the convergence condition in (9). That is, the error on the mutual information estimates should be less than the noisy information threshold $\mathbf{I}^o_\dagger$. Note that for the entropy estimates of $\mathbf{Y}$, equations (12) up to (15) hold with possibly different constants c', c'_1, c'_2, C' (see Assumption 1). Here we consider the case where $c' \in (1, 2)$ (the case $c' \geq 2$ is similar, see also the proof of Theorem 1), and the corresponding bound on the estimation error ϵ has to be at most $\mathbf{I}^o_\dagger/3$. Thus, (16) becomes

$$\mathbf{I}^o_\dagger > 3C' n^{\frac{1-c'}{c'}}, \tag{56}$$

and (17) now is written as

$$\mathbb{P}\left[\left|\hat{I}\left(Y_\ell; Y_{\bar{\ell}}\right) - I\left(Y_\ell; Y_{\bar{\ell}}\right)\right| > \mathbf{I}^o_\dagger\right] \leq 6 e^{-n\left(\frac{\mathbf{I}^o_\dagger}{3} - C' n^{\frac{1-c'}{c'}}\right)^2 / 2\log_2^2 n}. \tag{57}$$

Finally, by applying union bound over the pairs $\ell, \bar{\ell} \in \mathcal{V}$ we derive the statement of Theorem 4, by following the equivalent steps of (18) and (19). The latter completes the proof. □

First we observe that Theorem 4 requires $\mathbf{I}^o_\dagger > 0$. Secondly, note that the sample complexity bounds in Theorem 4 are similar to those of Theorem 1, but the noisy threshold $\mathbf{I}^o_\dagger$ takes the place of $\mathbf{I}^o$ and the constants c', c'_1, c'_2 are associated with the distribution $\mathrm{p}_\dagger(\cdot)$ (Assumption 1). Thirdly, to compare the sample complexity of the noiseless and noisy setting, let n and $n_\dagger$ denote the sufficient number of samples of $\mathbf{X}$ and $\mathbf{Y}$ respectively and consider p, δ fixed, then the ratio $n/n_\dagger$ is $\mathcal{O}\left((\mathbf{I}^o/\mathbf{I}^o_\dagger)^{-2(1+\zeta)}\right)$ for all $\zeta > 0$. The latter shows how $n_\dagger$ changes relative to n under the same probability of success for both settings (noiseless and noisy). Lastly, the inequality $\mathbf{I}^o > \mathbf{I}^o_\dagger$ should hold for a large class of hidden models, since learning from noisy data requires (in general) more samples than learning directly from noiseless data (see upcoming Sections V-C & VII).

Next, we provide conditions which guarantee that $\mathbf{I}^o_\dagger > 0$ and, therefore, feasibility of noisy structure learning, by applying the CL algorithm with input $\mathcal{D} = \mathbf{Y}^{1:n}$, that is, $\mathrm{T}^{\mathrm{CL}}_\dagger \to \mathrm{T}$.

B. IOP and the Importance of Pre-Processing

Prior to running our estimation algorithm, we would like to know if recovering the hidden tree structure is possible given the noise model. Unfortunately, the definition of $\mathbf{I}^o_\dagger$ involves the structure T that we want to estimate. We first find a condition under which we can guarantee $\mathbf{I}^o_\dagger > 0$ without any knowledge of T beforehand. We state this in terms of the randomized mapping $\mathcal{F}$ (see Section II-A). The next property guarantees that $\mathbf{I}^o > 0$ if and only if $\mathbf{I}^o_\dagger > 0$, for any $\mathrm{T} \in \mathcal{T}$.

Definition 4 (Information Order Preservation (IOP)). *Let $\mathbf{X} \in \mathcal{X}^p$ and $\mathbf{Y} \in \mathcal{Y}^p$ be random vectors such that $\mathbf{Y} = \mathcal{F}(\mathbf{X})$, with $\mathcal{F}$ defined as in Section II-A. We say that the randomized mapping $\mathcal{F}$ is information order-preserving (IOP) relative to $\mathbf{X}$ if and only if, for every tuple $\big((k, l), (m, r)\big) \in \mathcal{V}^2 \times \mathcal{V}^2$, such that $k \neq \ell$, $m \neq r$, $\{k, l\} \neq \{m, r\}$, it is true that*

$$I\left(X_k; X_\ell\right) < I\left(X_m; X_r\right) \iff I\left(F_k(X_k); F_\ell(X_\ell)\right) < I\left(F_m(X_m); F_r(X_r)\right). \tag{58}$$

Note that if the randomized mapping $\mathcal{F}$ is IOP and $\mathbf{I}^o > 0$ then we can guarantee that $\mathbf{I}^o_\dagger > 0$, without any knowledge of the hidden structure T. On the other hand, if the IOP does not hold it is still possible that $\mathbf{I}^o_\dagger > 0$. Thus, in general the condition (9) characterizes the error event of the CL algorithm (see Section V-C).

In certain cases, by knowing only the noise distribution we can find if IOP holds. To make this clear we provide an example of an erasure channel for M-ary alphabets. Specifically, for each hidden variable $X_i \in [M]$ and $i \in \mathcal{V}$, the corresponding observable $Y_i \in [M+1]$ is either $Y_i = X_i$ with probability $1 - q_i$ or $Y_i = 0$ (an erasure occurs) with probability q_i. Further, it is true that $I(Y_i; Y_j) = (1 - q_i)(1 - q_j) I\left(X_i; X_j\right)$ for all $i, j \in \mathcal{V}$, and $q_i, q_j \in [0, 1]$ (see Appendix, Section C-A). As a consequence, for certain values of the pairs (q_i, q_j) (58) holds, for instance consider the case $q_i = q_j$. Additionally, we show

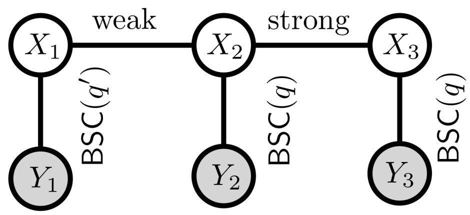

Fig. 3. Representation of the 3 nodes hidden model and non-identically distributed noise. Pre-processing is required to to guarantee that $\mathbf{I}^o_\dagger > 0$.

that the IOP is satisfied for several important cases of hidden graphical models, as we will discuss later in Section V-C. Next we show that CL algorithm is consistent when the IOP holds, while if IOP does not hold then an appropriate processing on the noisy data can help to overcome this issue.

1) Convergence of CL Algorithm under the IOP: If $\mathcal{F}$ is *information order-preserving* (IOP), then Algorithm 1 with input $\mathcal{D} = \mathbf{Y}^{1:n}$ converges to the true tree ($\mathrm{T}^{\mathrm{CL}}_\dagger \to \mathrm{T}$) (also see Theorem 4). For IOP randomized mappings, the following ordering with respect to all pairs of nodes

$$I(X_{i_1};X_{j_1}) < I(X_{i_2};X_{j_2}) < \cdots < I(X_{i_r};X_{j_r}), \tag{59}$$

for $r = \binom{p}{2}$, $i_s, j_s \in \mathcal{V}$, $s \in [r]$, remains unchanged for the observable node variables $\mathbf{Y}$. That is, (59) through (58), implies that

$$I(Y_{i_1};Y_{j_1}) < I(Y_{i_2};Y_{j_2}) < \cdots < I(Y_{i_r};Y_{j_r}), \tag{60}$$

for $r = \binom{p}{2}$, $i_s, j_s \in \mathcal{V}$, $s \in [r]$. Therefore, the maximum spanning tree algorithm with input weights in (59) returns the same tree structure T if the input weights are changed to the corresponding mutual information values of $\mathbf{Y}$ in (60). Therefore, for sufficient large number of samples the order is also preserved for the estimates $\hat{I}(Y_i;Y_j)$, which ensures that Algorithm 1 is consistent; $\mathrm{T}^{\mathrm{CL}}_\dagger \to \mathrm{T}$ for $\mathcal{D} = \mathbf{Y}^{1:n}$. Thus, the information order preservation property is sufficient to guarantee convergence of the CL Algorithm, that is, $\mathbf{I}^o_\dagger > 0$.

2) Enforcing the IOP through pre-processing: Although we are primarily interested in conditions ensuring a positive $\mathbf{I}^o_\dagger$, a negative $\mathbf{I}^o_\dagger$ is also informative. As we will show later, if we can find an appropriate pre-processing $\mathbf{Y} \to \mathbf{Z}$ such that $\mathbf{I}^o_{\dagger,\mathbf{Z}} > 0$, Theorem 4 applies with $\mathbf{I}^o_{\dagger,\mathbf{Z}} > 0$ as the new threshold, and $\mathbf{I}^o_{\dagger,\mathbf{Z}}$ is defined by replacing $\mathbf{Y}$ variable with $\mathbf{Z}$ in (10). To further explain this, we demonstrate the pre-processing procedure by enforcing the IOP in the example that follows. Although we consider a 3-node hidden Markov chain with binary random variables for brevity, the same technique can be applied on larger trees with p nodes. Additionally, the technique can be extended in certain models with larger alphabets.

To illustrate the case of $\mathbf{I}^o_\dagger < 0$ and the effect of the pre-processing (on the input data-set $\mathcal{D}$) we present a simple example of hidden tree-structured models for which Algorithm 1 succeeds with high probability, only if an appropriate pre-processing is being applied. Consider the smallest tree structure; let a three node Markov chain $X_1 - X_2 - X_3$ be the hidden layer and Y_1, Y_2, Y_3 be the variables of the observable layer (Figure 3). Specifically, $X_i, N_i \in \{-1,+1\}$ and $Y_i = N_i X_i$, for all $i \in \{1,2,3\}$. The distribution of the noise is

$$\mathbb{P}(N_1 = -1) = 1 - \mathbb{P}(N_1 = +1) = q' \in [0, 1/2), \tag{61}$$

and for $i \in \{2,3\}$

$$\mathbb{P}(N_i = -1) = 1 - \mathbb{P}(N_i = +1) = q \in (0, 1/2). \tag{62}$$

Thus the noisy variables are generated by a BSC(q_i) and the noise is not identically distributed, because $q \neq q'$, see Figure 3. Recall that the tree structure is $\mathrm{T} = (\mathcal{V}, \mathcal{E})$, $\mathcal{E} = \{(1,2),(2,3)\}$ and $\mathcal{V} = \{1,2,3\}$. Further $|\mathbb{E}[X_1X_2]|, |\mathbb{E}[X_3X_2]| \in (0,1)$, and without loss of generality assume that

$$|\mathbb{E}[X_1X_2]| \leq |\mathbb{E}[X_3X_2]|. \tag{63}$$

The Markov property [5], [24] of $X_1, X_2, X_3 \in \{-1,+1\}$ gives

$$\mathbb{E}[X_1X_3] = \mathbb{E}[X_1X_2]\mathbb{E}[X_3X_2]. \tag{64}$$

The definition of $\mathbf{I}^o$ in (6) together with (63),(64) give

$$\mathbf{I}^o = I(X_1;X_2) - I(X_1;X_3), \tag{65}$$

because $I(X_i;X_j)$ is increasing with respect to $|\mathbb{E}[X_iX_j]|$ (see Appendix, (95)). Additionally, it is true that $I(X_2;X_3) \geq I(X_1;X_2) > I(X_1;X_3)$ because $|\mathbb{E}[X_1X_3]| < |\mathbb{E}[X_1X_2]| \leq |\mathbb{E}[X_3X_2]| < 1$. The latter guarantees that the Chow-Liu algorithm, with input $\mathcal{D} = \mathbf{X}^{1:n}$ returns $\mathrm{T}^{\mathrm{CL}} = \mathrm{T}$ for $n \to \infty$ with probability 1. However, structure recovery is not guaranteed from noisy data. In fact $\mathbb{E}[N_i] = 1 - 2q_i$, $N_i \perp\!\!\!\perp X_i$ for all $i \in \{1,2,3\}$ and

$$\begin{aligned}\mathbb{E}[Y_1Y_2] &= \mathbb{E}[N_1X_1N_2X_2] \\ &= (1-2q')(1-2q)\mathbb{E}[X_1X_2],\end{aligned} \tag{66}$$

$$\begin{aligned}\mathbb{E}[Y_2Y_3] &= \mathbb{E}[N_2X_2N_3X_3] = (1-2q)^2\mathbb{E}[X_2X_3] \\ \mathbb{E}[Y_1Y_3] &= \mathbb{E}[N_1X_1N_3X_3] = (1-2q)\mathbb{E}[X_1X_3] \\ &= (1-2q')(1-2q)\mathbb{E}[X_1X_2]\mathbb{E}[X_2X_3].\end{aligned} \tag{67}$$

Under a possible error event, CL algorithm replaces $\{2,3\}$ by $\{1,3\}$ and returns $\mathcal{E}_{\mathrm{T}^{\mathrm{CL}}_\dagger} = \{(1,2),\{1,3\}\}$. The latter occurs even for zero estimation error; $I(Y_1;Y_2) = \hat{I}(Y_1;Y_2)$, $I(Y_1;Y_3) = \hat{I}(Y_1;Y_3)$, $I(Y_2;Y_3) = \hat{I}(Y_2;Y_3)$ if and only if

$$I(Y_1;Y_2) > I(Y_1;Y_3) > I(Y_2;Y_3), \tag{68}$$

and the last implies that $\mathbf{I}^o_\dagger = I(Y_2;Y_3) - I(Y_1;Y_3) < 0$ (by the definition (10) of $\mathbf{I}^o_\dagger$). Specifically, for $n \to \infty$ the edge $(2,3)$ will be replaced by $(1,3)$ (w.p. 1) if and only if $I(Y_1;Y_3) > I(Y_2;Y_3)$. The latter gives the locus of q, q' that yields an error in the structure estimation process as follows,

$$\begin{aligned}&I(Y_1;Y_3) > I(Y_2;Y_3) \iff \\ &|\mathbb{E}[Y_1Y_3]| > |\mathbb{E}[Y_2Y_3]| \iff \\ &(1-2q')(1-2q)|\mathbb{E}[X_1X_2]||\mathbb{E}[X_2X_3]| > \\ &\qquad (1-2q)^2|\mathbb{E}[X_2X_3]| \\ &\iff |\mathbb{E}[X_1X_2]| > \frac{1-2q}{1-2q'}.\end{aligned} \tag{69}$$

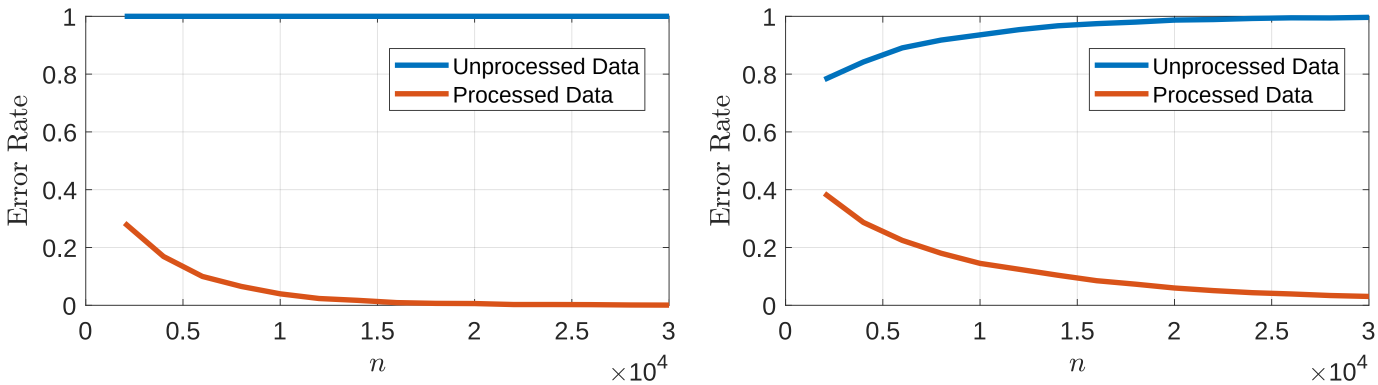

Fig. 4. Estimate of the probability of incorrect recovery for a hidden Markov model with 3 nodes, and $\mathbf{I}^o_\dagger < 0$ (originally), for different values of $n \in [10^3 \times 10^4]$, before and after pre-processing. Right: $q_1 = 0.01$, $q_2 = q_3 = 0.3$, Left: $q_1 = 0.2$, $q_2 = q_3 = 0.25$.

Further, note that (66) and (67) and guarantee that the left hand-side (first inequality) of (68) holds, thus

$$\begin{aligned} &|\mathbb{E}[X_1X_2]| > \frac{1-2q}{1-2q'} \\ &\overset{(66),(67)}{\iff} I(Y_1;Y_2) > I(Y_1;Y_3) > I(Y_2;Y_3) \\ &\overset{(66),(67)}{\iff} \mathbb{P}\left(\lim_{n\to\infty} \mathcal{E}_{\mathrm{T}^{\mathrm{CL}}_\dagger(\mathcal{D}=\mathbf{Y}^{1:n})} \neq \mathcal{E}_{\mathrm{T}}\right) = 1. \end{aligned} \tag{70}$$

As a consequence the structure learning from raw data is not guaranteed (with high probability) for all the values of pairs (q, q') that satisfy (70) (even for $n \to \infty$). To overcome this we have enforce the IOP (Definition 4) by considering the following pre-processing for each sample of the variables Y_1, Y_2, Y_3

$$\begin{aligned} Z_1 &\triangleq Y_1/(1-2q') \\ Z_2 &\triangleq Y_2/(1-2q) \\ Z_3 &\triangleq Y_3/(1-2q), \end{aligned} \tag{71}$$

then it is true that $I(Z_2;Z_3) \geq I(Z_1;Z_2) > I(Z_1;Z_3)$. The latter guarantees that IOP holds and

$$\mathbb{P}\left(\lim_{n\to\infty} \mathcal{E}_{\mathrm{T}^{\mathrm{CL}}_\dagger(\mathcal{D}=\mathbf{Z}^{1:n})} = \mathcal{E}_{\mathrm{T}}\right) = 1, \tag{72}$$

for any $q \in (0, 1/2)$, $q' \in [0, 1/2)$. Simulations on synthetic data verify our analysis (see Figure 4). As a final observation, even for $q' = 0$, (69) may hold as $|\mathbb{E}[X_1X_2]| > 1 - 2q$ and structure learning is infeasible by running the Chow-Liu on raw data. However if $q = q' \neq 0$ then (69) does not hold and structure learning from raw data is feasible which yields to the counterintuitive fact that introducing more noise (turning q' from 0 to $q > 0$) can make structure learning feasible in certain scenarios.

C. *Examples and Applications on Parametric Models*

Theorem 4 can be applied on a wide class models that satisfies the general Assumptions 1 and 2. To illustrate the effect of noise on the structure learning complexity, we consider two classical noisy channels in the hidden model: the M-ary erasure and the (generalized) symmetric channel. We show that a simple comparison of $\mathbf{I}^o$ and $\mathbf{I}^o_\dagger$ determines the impact of noise on the sample complexity, and we present the relationship between the two. As we explained, the CL algorithm can fail when noise is not i.i.d.. In the next examples, we present conditions for accurate structure estimation for certain model scenarios, while the number of nodes p is arbitrary and the noise is non-identically distributed.

Example 1: M-ary Erasure Channel. Assume that the randomized mappings $F_i(\cdot)$ "erase" each variable independently with probability q, so for all $i \in [p]$, we have $Y_i = X_i$ with probability $1 - q$ and $Y_i = M + 1$ (an erasure) with probability q. Then $I(Y_i;Y_j) = (1-q)^2 I(X_i;X_j)$ for all $i, j \in \mathcal{V}$ (see Appendix C-A). Therefore, $\mathbf{I}^o_\dagger = (1-q)^2\mathbf{I}^o \leq \mathbf{I}^o$ and the information order preservation (58) holds for any $q \in [0, 1)$. The latter guarantees that if $\mathbf{I}^o > 0$ then $\mathbf{I}^o_\dagger > 0$. Given the values of p, δ, q and $\mathbf{I}^o$, Theorem 4 provides the sample complexity for exact structure recovery from noisy observations. For fixed values of p and δ, the ratio of sufficient number of samples in the noiseless and noisy settings is $\mathcal{O}\big((1-q)^{4(1+\zeta)}\big)$, for any $\zeta > 0$. In contrast, consider the scenario where the erasure probability is not the same for every node (non-identically distributed noise). Each F_i erases the i^{th} node value with probability $q_i \in [0, 1)$, so $I(Y_i;Y_j) = (1-q_i)(1-q_j)I(X_i;X_j)$ for all $i, j \in \mathcal{V}$, and the condition (9) shows that Algorithm 1 with input $\mathcal{D} = \mathbf{Y}^{1:n}$ converges; $\mathrm{T}^{\mathrm{CL}}_\dagger \to \mathrm{T}$, if for all tuples $(w, w), u, \bar{u} \in \mathcal{EV}^2$

$$\frac{(1-q_w)(1-q_{\bar{w}})}{(1-q_u)(1-q_{\bar{u}})} > \frac{I(X_u;X_{\bar{u}})}{I(X_w;X_{\bar{w}})}. \tag{73}$$

Define $\mathbf{RI} \triangleq \max_{(w,\bar{w}),u,\bar{u}\in\mathcal{EV}^2} I(X_u;X_{\bar{u}})/I(X_w;X_{\bar{w}})$. Inequality (73) provides the following simplified sufficient condition for convergence of CL Algorithm with input noisy data; $\mathrm{T}^{\mathrm{CL}}_\dagger \to \mathrm{T}$ if for all $i, j \in \mathcal{V}$

$$\frac{1-q_i}{1-q_j} \in \left(\mathbf{RI}^{1/2}, \mathbf{RI}^{-1/2}\right). \tag{74}$$

Given the values q_i, q_j and $\mathbf{RI}$ or estimates of them, (74) provides rule for testing if structure estimation is possible directly from raw noisy data. If direct structure estimation is not possible then we can consider a pre-processing that enforces the IOP similarly to the example of Section V-B2. We continue by providing a feasibility condition for the binary symmetric channel with non-identically distributed noise. The identically distributed noise case was studied in our prior work [11].

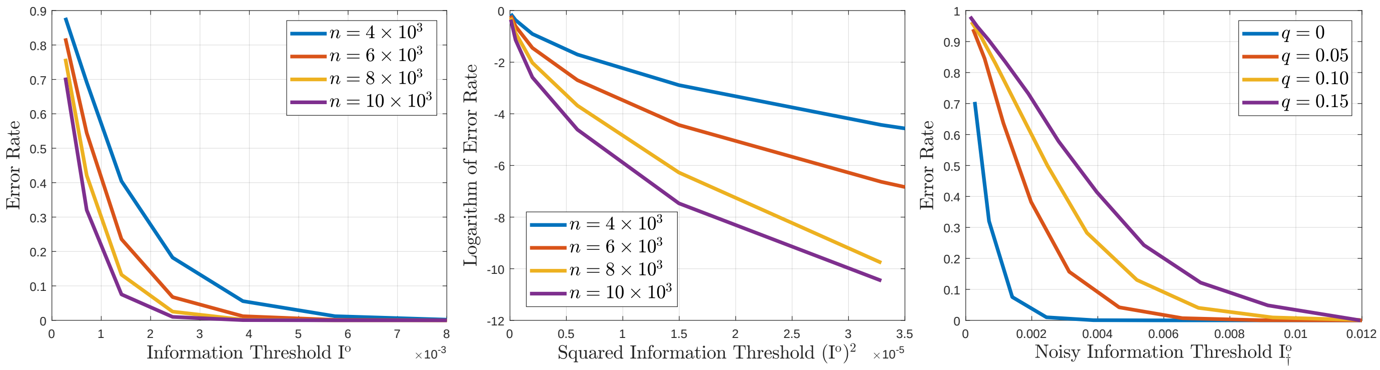

Fig. 5. Left, center: estimating the probability $\mathbb{P}\left(\mathrm{T}^{\mathrm{CL}} \neq \mathrm{T}\right)$ for different values of $\mathbf{I}^o$ and n through 5×10^3 independent runs with noiseless data. Right: estimating the probability $\mathbb{P}\big(\mathrm{T}^{\mathrm{CL}}_{\dagger} \neq \mathrm{T}\big)$ through 10^4 samples and 5×10^3 independent runs with noisy data.

Example 2: Binary Symmetric Channel. Assume that the hidden variables $\mathbf{X} \in \{-1,+1\}^p$ follow a tree-structured Ising model, the noisy observable variables $\mathbf{Y} \in \{-1,+1\}^p$ are generated by setting $Y_i = X_i$ with probability $1-q_i$ and $Y_i = -X_i$ with probability $q_i \in [0,1/2)$, $i \in [p]$, and q_i is the probability of a value to change sign. Structure recovery directly from noisy observations $\mathbf{Y}^{1:n}$ is feasible, because the information order preservation property holds when the crossover probability values q_i are equal for all $i \in \mathcal{V}$ [11]. To guarantee convergence of Algorithm 1 ($\mathcal{D} = \mathbf{Y}^{1:n}$) for more general cases, the condition $\mathbf{I}^o_\dagger > 0$ should hold. As an example, consider the case of non identically distributed noise, that is, the probability q_i of a flip may differ for each node, In this case, the information order preservation property does not hold for all the possible sequences $\{q_1, q_2 \dots, q_p\} \in [0,1/2)^p$ and we would like to know for which values of q_i, $i \in [p]$, Algorithm 1 with $\mathcal{D} = \mathbf{Y}^{1:n}$ learns the hidden structure. The condition (9) implies that if for all $i,j \in \mathcal{V}$

$$\frac{(1-2q_i)}{(1-2q_j)} \in \left(\max_{(i,j)\in\mathcal{E}_{\mathrm{T}}} |\mathbb{E}[X_i X_j]|, \frac{1}{\max\limits_{(i,j)\in\mathcal{E}_{\mathrm{T}}} |\mathbb{E}[X_i X_j]|} \right) \tag{75}$$

then $\mathrm{T} \to \mathrm{T}^{\mathrm{CL}}_\dagger$ and the proof of (75) is given in Appendix C-B. The condition (75) provides a testing rule for tree-structure estimation directly from raw noisy data, given the model parameters q_i, q_j and $\max_{(i,j)\in\mathcal{E}_{\mathrm{T}}}$, or estimates of these parameters. Note that for identical noise $q_i = q_j = q$ it is true that $(1-2q_i)/(1-2q_j) = 1$ for all $i,j \in \mathcal{V}$, thus (75) is always satisfied because $|\mathbb{E}[X_i X_j]| \in (0,1)$, and structure learning is always feasible for this regime. If the condition (75) is not satisfied then structure recovery is still feasible by applying an appropriate pre-processing on the data $\mathbf{Y}^{1:n}$. The pre-processing procedure requires the values q_i (or estimates of them) to be known, similarly to the example in Section V-B2. In contrast, Algorithm 1 does not require any information related to the values q_i, and its convergence is guaranteed under the condition (75). In the next example we study an extension of the binary symmetric channel to alphabets of size M.

Example 3: Generalized Symmetric Channel. We define the generalized symmetric channel as follows, assume $\mathbf{X} \sim \mathrm{p}(\cdot) \in \mathcal{P}_{\mathcal{T}}$, $\mathbf{X} \in [M]^p$ and let Z_i for $i \in [p]$ be i.i.d uniform random variables, such that $\mathbb{P}(Z_i = k) = 1/M$, for all $k \in [M]$. Also assume that $\boldsymbol{Z}$ and $\mathbf{X}$ are independent, then the i^{th} variable of the channel output $\mathbf{Y} \in [M]^p$ is defined as

$$Y_i = F_i(X_i) = \begin{cases} X_i, & \text{with probability } 1-q \\ Z_i, & \text{with probability } q, \end{cases} \tag{76}$$

and $q \in [0,1)$. Note that the probability of a symbol to remain unchanged is $\mathbb{P}(Y_i = X_i) = (1-q) + q/M$, and for $M = 2$ it reduces to the binary symmetric channel. Theorem 4 can be directly applied for given values of p, δ and $\mathbf{I}^o_\dagger$. However, closed form expressions of $\mathbf{I}^o_\dagger$ relative to $\mathbf{I}^o$ are unknown. As a consequence, an explicit comparison between the bounds of Theorems 1 and 4 is hard to evaluate in a closed-form expression for any $q \in [0,1)$. Nevertheless, we are able to derive approximations of the relationship of $\mathbf{I}^o$ and $\mathbf{I}^o_\dagger$ by considering sufficiently small q. Lemma 6 (Section C-C, Appendix) shows that in the small noise regime it is true that

$$\begin{aligned} &\mathbf{I}^o_\dagger = (1-q)^2 \mathbf{I}^o - (1-(1-q)^2)\Delta_{\mathrm{KL}} + \mathcal{O}(\epsilon^2), \\ &\Delta_{\mathrm{KL}} \triangleq \mathrm{KL}(U \| \mathrm{p}(x_{w^*}, x_{\bar{w}^*})) - \mathrm{KL}(U \| \mathrm{p}(x_{u^*}, x_{\bar{u}^*})), \end{aligned} \tag{77}$$

U is the uniform distribution on the alphabet $[M]^2$, the tuple $((w^*, \bar{w}^*), u^*, \bar{u}^*) \in \mathcal{E}\mathcal{V}^2$ is the argument of the minimum in (6), and $\epsilon = \left(1-(1-q)^2\right)/M^2 < q$. Whenever $\mathbf{I}^o_\dagger > 0$ perfect reconstruction is possible and Theorem 4 provides the sufficient number of samples.

VI. Simulations

To demonstrate the relationships between δ, n and $\mathbf{I}^o$, $\mathbf{I}^o_\dagger$, we estimate $\mathrm{T}^{\mathrm{CL}}, \mathrm{T}^{\mathrm{CL}}_\dagger$ and δ through 5×10^3 independent runs on tree-structured synthetic data, for different values of $n \in [10^3, 10^4]$ and $p = 10$. The variable nodes are binary and take values in the set $\{-1,+1\}$. Figure 5 (left) illustrates the relationship between the probability of incorrect reconstruction and $\mathbf{I}^o$ and Figure 5 (center) the relationship between the log-probability of error as a function of the squared information threshold. We observe that probability of incorrect reconstruction decays exponentially with respect to $(\mathbf{I}^o)^2$ as Theorem 1 suggests.

Lastly, Figure 5 (right) presents the effect of noise of a BSC for different values of the crossover q and number of samples $n = 10^4$. Notice that the probability of incorrect reconstruction also decays exponentially with respect to $(\mathbf{I}^o_\dagger)^2$, as Theorem 4 suggests, but with a significantly smaller rate than the noiseless case ($q = 0$).

VII. Conclusion & Future Directions

In this paper we showed how the *information threshold* $\mathbf{I}^o$ characterizes the problem of recovering the structure of a tree-shaped graphical model. This quantity arises naturally in the error of Chow-Liu algorithm. As our main contribution, we introduced the first finite sample complexity bound on the performance of the CL algorithm for learning tree structured models. Additionally, we proved that as long as $\mathbf{I}^o \to 0$, no algorithm can estimate efficiently the tree-structure with large probability. Further, we introduced the noisy information threshold $\mathbf{I}^o_\dagger$, and based on that, we presented finite sample complexity bounds for learning hidden tree-structured models. More specifically, our sample complexity bounds show how the number of nodes p, the probability of failure δ, and $\mathbf{I}^o_\dagger$ are related for the problem of structure recovery. Our results demonstrate how noise affects the sample complexity of learning for a variety of standard models, including models for which the noise is not identically distributed.

Although we strictly consider the class of tree-structured models in this paper, our approaches of Theorem 1 and Theorem 3 can be extended to the class of forests. For that purpose, we should consider an generalization of $\mathbf{I}^o$ to forests and a modified version of CL the CLThres algorithm [7]. We leave this part for future as it is out of scope of this paper.

Additionally, our approach is more generally applicable to the analysis of δ-PAC Maximum Spanning Tree (MST) algorithms. At its root, our work shows how the error probability of MST algorithms (for example, Kruskal's algorithm or Prim's algorithm) behaves when edge weights are uncertain, i.e., when only (random) estimates of the true edge weights are known.

To conclude, the non-parametric graphical model setting presents interesting theoretical challenges that are connected with other statistical problems, out of the focus of this paper. The relationship between $\mathbf{I}^o$ and $\mathbf{I}^o_\dagger$ is connected with open problems in information theory related to Strong Data Processing Inequalities [35], [36], for which tight characterizations are only known for a few channels. In our situation, a general analytical relationship may be similarly challenging. From a practical standpoint, we may wish to estimate the sample size needed to guarantee recovery with a pre-specified error probability. To do so would require knowing $\mathbf{I}^o_\dagger$ before collecting the full data; since $\mathbf{I}^o_\dagger$ depends on the noise *model*, we could find such a bound by considering a reasonable class of underlying models and taking the worst case. An interesting open question for future work is how to effectively estimate $\mathbf{I}^o_\dagger$ from (auxiliary) training data rather than relying on such a priori modeling assumptions. This may help design pre-processing methods that can make structure learning algorithms more robust against noise or adversarial attacks.

Appendix A
Proofs & Results for the Noiseless Case

A. Proof of Proposition 3

We consider the case $u^* \equiv w^*$ and $\bar{u} \in \mathcal{N}_{\mathrm{T}}(\bar{w})$, while the other three cases that are given by the locality property can be identically proved. The case $u^* \equiv w^*$ and $\bar{u} \in \mathcal{N}_{\mathrm{T}}(\bar{w})$ implies that $w^* - \bar{w}^* - \bar{u}^*$ is a subgraph of T. Assume for sake of contradiction that $I(X_{w^*}; X_{\bar{w}^*}) = I(X_{w^*}; X_{\bar{u}^*})$ then $I(X_{w^*}; X_{\bar{w}^*} | X_{\bar{u}^*}) = 0$. The latter implies that $w^* - \bar{u}^* - \bar{w}^*$ is also a subgraph of T and it contradicts with the uniqueness of the structure (Assumption 2). □

B. Proof of Proposition 4

Assume for sake of contradiction that $u^* \neq w^*$ and $u^* \neq \tilde{w}^*$ or $\bar{u} \notin \mathcal{N}_{\mathrm{T}}(w)$ and $\bar{u} \notin \mathcal{N}_{\mathrm{T}}(\bar{w})$ and let ν be a node such that $\nu \in \mathcal{N}_{\mathrm{T}}(w) \cup \mathcal{N}_{\mathrm{T}}(\bar{w})$, then the data processing inequality [9] and Assumption 2 give

$$I(X_w; X_{\bar{w}}) - I(X_{\bar{w}}; X_\nu) < I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}}) \quad (78)$$

and

$$I(X_w; X_{\bar{w}}) - I(X_w; X_\nu) < I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}}). \quad (79)$$

The last two inequalities contradict the assumption (8). □

C. Sample Complexity for the Noiseless Setting when $c \geq 2$

The corresponding variation of Theorem 1 for the case where $c \geq 2$ follows. Note that, for any $c \geq 2$ (c is free), the constants c_1, c_2 can be evaluated according to Assumption 1.

Theorem 5. *Assume that $\mathbf{X} \sim \mathrm{p}(\cdot) \in \mathcal{P}_{\mathrm{T}}(c_1, c_2)$ for some $c \geq 2$. Fix a number $\delta \in (0,1)$. If the number of samples of $\mathbf{X}$ satisfies the inequalities*

$$\frac{n}{\log_2^2 n} \geq \frac{72 \log\left(\frac{p}{\delta}\right)}{\left(\mathbf{I}^o - C\frac{\log n}{\sqrt{n}}\right)^2} \quad \textit{and} \quad \mathbf{I}^o > C\frac{\log n}{\sqrt{n}}, \quad (80)$$

for a constant $C > 0$, then Algorithm 1 with input $\mathcal{D} = \mathbf{X}^{1:n}$ returns $\mathrm{T}^{\mathrm{CL}} = \mathrm{T}$ with probability at least $1 - \delta$.

Appendix B
Converse: Technical Lemmata

Lemma 1 (Fano's Inequality [37]). *Fix $M \geq 2$ and let Θ be a family of models $\theta^0, \theta^1, \ldots, \theta^M$. Let $\mathbb{P}_{\theta^j}$ denote the probability law of $\mathbf{X}$ under model θ^j, and consider n i.i.d. observations $\mathbf{X}^{1:n}$. If*

$$n < (1-\delta)\frac{\log M}{\frac{1}{M+1}\sum_{j=1}^M \boldsymbol{D}_{KL}(\mathbb{P}_{\theta^j} || \mathbb{P}_{\theta^0})}, \quad (81)$$

then it is true that

$$\inf_{\Phi} \max_{0 \leq j \leq M} \mathbb{P}_{\theta^j}\left[\Phi(\mathbf{X}^{1:n}) \neq j\right] \geq \delta - \frac{1}{\log(M)}, \quad (82)$$

where the infimum is relative to all estimators (statistical tests) $\Phi : \mathcal{X}^{p \times n} \to \{0, 1, \ldots, M\}$.

Lemma 2. *Let $f : (0, \infty] \to \mathbb{R}$ be nondecreasing. Then, it is true that*

$$\inf_{\eta > 0} f(\eta) \equiv \lim_{\eta \downarrow 0} f(\eta). \quad (83)$$

Proof of Lemma 2. If f is unbounded from below (on $(0, \infty]$), then $\inf_{\eta>0} f(\eta) \equiv -\infty \equiv \lim_{\eta \downarrow 0} f(\eta)$. So, for the rest of the proof we assume that f is bounded from below. We follow a standard sequential argument. Consider *any* sequence $\{\eta_n > 0\}_{n\in\mathbb{N}}$, such that $\eta_n \underset{n\to\infty}{\to} 0$. We want to show that

$$f(\eta_n) \underset{n\to\infty}{\to} \inf_{\eta>0} f(\eta), \tag{84}$$

or that, for every $\varepsilon > 0$, there exists $N(\varepsilon) > 0$ such that, for $n \geq N(\varepsilon)$, it holds that

$$f(\eta_n) - \inf_{\eta>0} f(\eta) < \varepsilon. \tag{85}$$

Consider a *nonincreasing sequence* defined as $\widetilde{\eta}_n \triangleq \sup_{k\geq n} \eta_k \geq \eta_n$. Then, by construction, it is true that

$$\begin{aligned}\lim_{n\to\infty} \widetilde{\eta}_n &\equiv \lim_{n\to\infty} \sup_{k\geq n} \eta_k \\ &\equiv \limsup_{n\to\infty} \eta_n \\ &\equiv \lim_{n\to\infty} \eta_n \equiv 0. \end{aligned}\tag{86}$$

By monotone convergence (note that f is bounded from below), we get that, for every $\varepsilon/2 > 0$, there exists $N(\varepsilon) > 0$ such that, for $n \geq N(\varepsilon)$,

$$f(\widetilde{\eta}_n) - \inf_n f(\widetilde{\eta}_n) < \varepsilon/2. \tag{87}$$

Note, though, that

$$\inf_n f(\widetilde{\eta}_n) \leq f(\widetilde{\eta}_n), \quad \forall n, \tag{88}$$

and also that we can always find a sufficiently small $\widetilde{\eta}_{n^o} \equiv \widetilde{\eta}_{n^o}(\varepsilon)$, such that

$$f(\widetilde{\eta}_{n^o}) < \inf_{\eta>0} f(\eta) + \varepsilon/2. \tag{89}$$

Therefore, it follows that

$$\inf_n f(\widetilde{\eta}_n) < \inf_{\eta>0} f(\eta) + \varepsilon/2. \tag{90}$$

Additionally, $f(\eta_n) \leq f(\widetilde{\eta}_n)$ by monotonicity. Consequently, it is true that

$$f(\eta_n) - \inf_{\eta>0} f(\eta) - \varepsilon/2 < f(\widetilde{\eta}_n) - \inf_n f(\widetilde{\eta}_n) < \varepsilon/2, \tag{91}$$

which implies that $f(\eta_n) - \inf_{\eta>0} f(\eta) < \varepsilon$, thus proving our claim. □

APPENDIX C
PROOFS FOR THE NOISY CASE

A. M-ary Erasure Channel

For the M-ary erasure channel, it is true that

$$I(Y_i;Y_j) = (1-q_i)(1-q_j) I\left(X_i;X_j\right). \tag{92}$$

for all $i,j \in \mathcal{V}$ and $q_i, q_j \in [0,1)$. To prove this, we start by expanding the mutual information from the definition and pulling out the erasure event as follows

$$\begin{aligned}
&I(Y_i;Y_j)\\
&= \sum_{y_i,y_j\in[M+1]^2} \mathrm{p}_\dagger(y_i,y_j) \log \frac{\mathrm{p}_\dagger(y_i,y_j)}{\mathrm{p}_\dagger(y_i)\mathrm{p}_\dagger(y_j)}\\
&= \sum_{y_i,y_j\in[M]^2} \mathrm{p}_\dagger(y_i,y_j) \log \frac{\mathrm{p}_\dagger(y_i,y_j)}{\mathrm{p}_\dagger(y_i)\mathrm{p}_\dagger(y_j)}\\
&\quad + \sum_{y_i\in[M]} \mathrm{p}_\dagger(y_i,M+1) \log \frac{\mathrm{p}_\dagger(y_i,M+1)}{\mathrm{p}_\dagger(y_i)\mathrm{p}_\dagger(M+1)}\\
&\quad + \sum_{y_j\in[M]} \mathrm{p}_\dagger(M+1,y_j) \log \frac{\mathrm{p}_\dagger(M+1,y_j)}{\mathrm{p}_\dagger(M+1)\mathrm{p}_\dagger(y_j)}\\
&\quad + \mathrm{p}_\dagger(M+1,M+1) \log \frac{\mathrm{p}_\dagger(M+1,M+1)}{\mathrm{p}_\dagger(M+1)\mathrm{p}_\dagger(M+1)}\\
&= \sum_{y_i,y_j\in[M]^2} \mathrm{p}_\dagger(y_i,y_j) \log \frac{\mathrm{p}_\dagger(y_i,y_j)}{\mathrm{p}_\dagger(y_i)\mathrm{p}_\dagger(y_j)}\\
&= \sum_{x_i,x_j\in[M]^2} (1-q_i)(1-q_j)\mathrm{p}(x_i,x_j) \log \frac{\mathrm{p}(x_i,x_j)}{\mathrm{p}(x_i)\mathrm{p}(x_j)}\\
&= (1-q_i)(1-q_j) I\left(X_i;X_j\right).
\end{aligned}\tag{93}$$

An erasure occurs independently on each node variable observable and independently with respect to the $\mathbf{X}$, thus $\mathrm{p}_\dagger(y_i, M+1) = \mathrm{p}_\dagger(y_i)\mathrm{p}_\dagger(M+1)$, for any $y_i \in [M+1]$ and $\mathrm{p}_\dagger(M+1, y_j) = \mathrm{p}_\dagger(M+1)\mathrm{p}_\dagger(y_j)$ for any $y_j \in [M+1]$. The latter gives (93). □

B. Binary Symmetric Channel with Non-Identically Distributed Noise

Under the assumption of $\mathbf{I}^o > 0$, we wish to show that if

$$\frac{(1-2q_i)}{(1-2q_j)} \in \left(\max_{(i,j)\in\mathcal{E}_{\mathrm{T}}} |\mathbb{E}[X_iX_j]|, \frac{1}{\max_{(i,j)\in\mathcal{E}_{\mathrm{T}}} |\mathbb{E}[X_iX_j]|} \right), \tag{94}$$

for all $i,j \in \mathcal{V}$ then $\mathbf{I}^o_\dagger > 0$. We start by finding the values of the sequence of crossover probabilities $q_1, q_2, \ldots, q_k \in [0, 1/2)$ which guarantee that $\mathbf{I}^o_\dagger > 0$. The mutual information of two binary random variables $Y_i, Y_j \in \{-1, +1\}$ (see [24]) is

$$\begin{aligned}
&I\left(Y_i, Y_j\right)\\
&= \frac{1}{2}\log_2 \left(\left(1 - \mathbb{E}\left[Y_iY_j\right]\right)^{1-\mathbb{E}[Y_iY_j]} \left(1 + \mathbb{E}\left[Y_iY_j\right]\right)^{1+\mathbb{E}[Y_iY_j]} \right).
\end{aligned}\tag{95}$$

The definition of $\mathbf{I}^o_\dagger$ (Definition 3) and (95) give

$$\begin{aligned}
\mathbf{I}^o_\dagger &= \frac{1}{2}\left\{I\left(Y_w;Y_{\bar{w}}\right) - I\left(Y_u;Y_{\bar{u}}\right)\right\}\\
&= \frac{1}{2}\log_2 \frac{\left(1-\mathbb{E}\left[Y_wY_{\bar{w}}\right]\right)^{1-\mathbb{E}[Y_wY_{\bar{w}}]}\left(1+\mathbb{E}\left[Y_wY_{\bar{w}}\right]\right)^{1+\mathbb{E}[Y_wY_{\bar{w}}]}}{\left(1-\mathbb{E}\left[Y_uY_{\bar{u}}\right]\right)^{1-\mathbb{E}[Y_uY_{\bar{u}}]}\left(1+\mathbb{E}\left[Y_uY_{\bar{u}}\right]\right)^{1+\mathbb{E}[Y_uY_{\bar{u}}]}}.
\end{aligned}\tag{96}$$

Define the function $f(\cdot)$ as

$$f(x) \triangleq (1-x)^{1-x}(1+x)^{1+x} \equiv f(|x|), \tag{97}$$

then

$$\mathbf{I}^o_\dagger = \frac{1}{2}\log_2 \frac{f(|\mathbb{E}\left[Y_wY_{\bar{w}}\right]|)}{f(|\mathbb{E}\left[Y_uY_{\bar{u}}\right]|)} \tag{98}$$

and

$$\mathbb{E}\left[Y_uY_{\bar{u}}\right] = (1-2q_w)(1-2q_{\bar{w}})\mathbb{E}\left[X_wX_{\bar{w}}\right], \tag{99}$$

$$\begin{aligned}
\mathbb{E}\left[Y_uY_{\bar{u}}\right] = (1-2q_w)(1-2q_{\bar{u}})\mathbb{E}\left[X_wX_{\bar{w}}\right]\\
\times \prod_{(i,j)\in\mathrm{path}_{\mathrm{T}}(u.\bar{u})\setminus(w,\bar{w})} \mathbb{E}\left[X_iX_j\right])
\end{aligned}\tag{100}$$

for the last equality we used the correlation decay property [5], [11]) and the fact that for ±1-valued variables the binary symmetric channel can be consider as multiplicative binary noise [11]. Note that $f(x)$ is increasing for $x > 0$. To guarantee that $\mathbf{I}^o_\dagger > 0$ we need

$$\frac{(1-2q_w)(1-2q_{\bar{w}})}{(1-2q_u)(1-2q_{\bar{u}})} > \prod_{(i,j)\in \mathrm{path}_{\mathrm{T}}(u.\bar{u})\setminus(w,\bar{w})} |\mathbb{E}\left[X_i X_j\right]|. \quad (101)$$

Recall that (101) should hold for all $(w,\bar{w}) \in \mathcal{E}$ and for all $u, \bar{u} \in \mathcal{V}$ such that $(w,\bar{w}) \in \mathrm{path}_{\mathrm{T}}(u,\bar{u})$.
In addition,

$$\prod_{(i,j)\in \mathrm{path}_{\mathrm{T}}(u.\bar{u})\setminus(w,\bar{w})} |\mathbb{E}\left[X_i X_j\right]| \leq \left(\max_{(i,j)\in\mathcal{E}_{\mathrm{T}}} |\mathbb{E}[X_i X_j]|\right).$$

The last two inequalities give the sufficient condition

$$\frac{(1-2q_w)(1-2q_{\bar{w}})}{(1-2q_u)(1-2q_{\bar{u}})} > \left(\max_{(i,j)\in\mathcal{E}_{\mathrm{T}}} |\mathbb{E}[X_i X_j]|\right)^{|\mathrm{path}_{\mathrm{T}}(u,\bar{u})|-1}.$$

As a consequence, if

$$\frac{(1-2q_i)}{(1-2q_j)} \in \left(\max_{(i,j)\in\mathcal{E}_{\mathrm{T}}} |\mathbb{E}[X_i X_j]|, \frac{1}{\max_{(i,j)\in\mathcal{E}_{\mathrm{T}}} |\mathbb{E}[X_i X_j]|}\right),$$

for all $i, j \in \mathcal{V}$ then $\mathbf{I}^o_\dagger > 0$. Note that for the case of i.i.d. noise ($q_{\bar{i}} = q_{\bar{j}}$ for all $i, j \in \mathcal{V}$) the inequality always holds because $\max_{(i,j)\in\mathcal{E}_{\mathrm{T}}} |\mathbb{E}[X_i X_j]| \in (0,1)$. □

C. M-ary Symmetric Channel

Lemma 3. *Let $A, B \in [L]$ be two discrete random variables, such that $A \sim \mathrm{p}_A(\cdot)$, $B \sim \mathrm{p}_B(\cdot)$ and $H(A) < H(B)$. Assume $A', B' \in [L]$ and $A' \sim \mathrm{p}_{A'}(\cdot)$ and $B' \sim \mathrm{p}_{B'}(\cdot)$ and $q \in [0.1/2)$ such that*

$$\mathrm{p}_{A'}(\ell) = (1-q)^2 \mathrm{p}_A(\ell) + \frac{1-(1-q)^2}{L}, \quad \textit{for all } \ell \in [L],$$
$$\mathrm{p}_{B'}(\ell) = (1-q)^2 \mathrm{p}_B(\ell) + \frac{1-(1-q)^2}{L}, \quad \textit{for all } \ell \in [L].$$

Then $H(A') < H(B')$ for sufficiently small values of $q > 0$.

Proof. Define $\epsilon \triangleq \left(1-(1-q)^2\right)/L$, so $0 < \epsilon < q$ for any $L \geq 2$.

$$\begin{aligned}
&\mathrm{p}_{A'}(\ell)\log_2 \mathrm{p}_{A'}(\ell) \\
&= \left((1-q)^2\mathrm{p}_A(\ell)+\epsilon\right)\log_2\left((1-q)^2\mathrm{p}_A(\ell)+\epsilon\right) \\
&= (1-q)^2\mathrm{p}_A(\ell)\log_2\left((1-q)^2\mathrm{p}_A(\ell)+\epsilon\right)+ \\
&\qquad \epsilon\log_2\left((1-q)^2\mathrm{p}_A(\ell)+\epsilon\right) \qquad (102)\\
&= (1-q)^2\mathrm{p}_A(\ell)\log_2\left((1-q)^2\mathrm{p}_A(\ell)\right)+\epsilon+\mathcal{O}(\epsilon^2) \\
&\qquad +\epsilon\log_2\left((1-q)^2\mathrm{p}_A(\ell)\right)+\mathcal{O}(\epsilon^2).
\end{aligned}$$

To derive (102) recall that $\log(1+x) = x + \mathcal{O}(x^2)$ for $x < 1$, and set $x = \epsilon/(1-q)^2\mathrm{p}_A(\ell)$ then the latter gives

$$\begin{aligned}
&\mathrm{p}_{A'}(\ell)\log_2 \mathrm{p}_{A'}(\ell) \\
&= (1-q)^2\mathrm{p}_A(\ell)\log_2\left((1-q)^2\mathrm{p}_A(\ell)\right) \\
&\qquad +\epsilon\left(1+\log_2\left((1-q)^2\mathrm{p}_A(\ell)\right)\right)+\mathcal{O}(\epsilon^2). \qquad (103)
\end{aligned}$$

Also,

$$\begin{aligned}
&\mathrm{p}_{B'}(\ell)\log_2 \mathrm{p}_{B'}(\ell) \\
&= (1-q)^2\mathrm{p}_B(\ell)\log_2\left((1-q)^2\mathrm{p}_B(\ell)\right) \\
&\qquad +\epsilon\left(1+\log_2\left((1-q)^2\mathrm{p}_B(\ell)\right)\right)+\mathcal{O}(\epsilon^2). \qquad (104)
\end{aligned}$$

Expanding both sides of the inequality $H(A) < H(B)$ we obtain the following:

$$\begin{aligned}
&-\sum_{\ell=1}^{L}\mathrm{p}_A(\ell)\log_2\mathrm{p}_A(\ell) < -\sum_{\ell}\mathrm{p}_B(\ell)\log_2\mathrm{p}_B(\ell) \implies \\
&-\sum_{\ell=1}^{L}(1-q)^2\mathrm{p}_A(\ell)\log_2\left((1-q)^2\mathrm{p}_A(\ell)\right) \\
&\quad < -\sum_{\ell}(1-q)^2\mathrm{p}_B(\ell)\log_2\left((1-q)^2\mathrm{p}_B(\ell)\right),
\end{aligned}$$

then for sufficiently small q and for any $L > 2, \mathrm{p}_A(\cdot), \mathrm{p}_B(\cdot)$ there exist $\epsilon > 0$ such that

$$\begin{aligned}
&-\sum_{\ell=1}^{L}\Big[(1-q)^2\mathrm{p}_A(\ell)\log_2\left((1-q)^2\mathrm{p}_A(\ell)\right) \\
&\qquad +\epsilon\left(1+\log_2\left((1-q)^2\mathrm{p}_A(\ell)\right)\right)+\mathcal{O}(\epsilon^2)\Big] \\
&< -\sum_{\ell=1}^{L}\Big[(1-q)^2\mathrm{p}_B(\ell)\log_2\left((1-q)^2\mathrm{p}_B(\ell)\right) \\
&\qquad +\epsilon\left(1+\log_2\left((1-q)^2\mathrm{p}_B(\ell)\right)\right)+\mathcal{O}(\epsilon^2)\Big].
\end{aligned}$$

This together with (103) and (104) give $H(A') < H(B')$. □

We consider the extension an extension of the binary symmetric channel to alphabets of size M as follows. Assume $\mathbf{X} \sim \mathcal{P}_{\mathrm{T}}(c_1, c_2)$, $\mathbf{X} \in [M]$ and let Z_i for $i \in [p]$ be i.i.d uniform random variables, $\mathbb{P}(Z_i = k) = 1/M$, for all $k \in [M]$. Also assume that $\boldsymbol{Z}$ and $\mathbf{X}$ are independent, then the noisy output variable $\mathbf{Y} \in [M]^p$ of the channel is defined for $q \in [0,1)$ as

$$Y_i = F_i(X_i) = \begin{cases} X_i, & \text{with probability } 1-q \\ Z_i, & \text{with probability } q \end{cases}. \qquad (105)$$

Lemma 4. *The distribution of the two output variables Y_i, Y_j of the M-ary symmetric channel can be expressed as*

$$\begin{aligned}
&\mathbb{P}(Y_i = y_i, Y_j = y_j) \\
&\quad = (1-q)^2\mathbb{P}(X_i = y_i, X_j = y_j) + \frac{1-(1-q)^2}{M^2}.
\end{aligned}$$

Proof. This is a straightforward calculation

$$\begin{aligned}
&\mathrm{p}_\dagger(y_i, y_j) \\
&= \mathbb{P}(Y_i = y_i, Y_j = y_j) \\
&= (1-q)^2\mathbb{P}(X_i = y_i, X_j = y_j) \\
&\qquad + q(1-q)\mathbb{P}(Z_i = y_i, X_j = y_j) \\
&\qquad + q(1-q)\mathbb{P}(X_i = y_i, Z_j = y_j) \\
&\qquad + q^2\mathbb{P}(Z_i = y_i, Z_j = y_j) \\
&= (1-q)^2\mathbb{P}(X_i = y_i, X_j = y_j) + 2\frac{q(1-q)}{M^2} + \frac{q^2}{M^2} \\
&= (1-q)^2\mathbb{P}(X_i = y_i, X_j = y_j) + \frac{1-(1-q)^2}{M^2},
\end{aligned}$$

and we are done. □

Lemma 5. *Consider X_k, X_ℓ, X_m, X_r as four distinct inputs variables of the M-ary symmetric channel (defined in Section V-C) with corresponding outputs Y_k, Y_ℓ, Y_m, Y_r. If the crossover probability q is sufficiently small and $I(X_k; X_\ell) < I(X_m; X_r)$ then $I(Y_k; Y_\ell) < I(Y_m; Y_r)$.*

Proof. Note that the assumption of uniform marginal distributions for all four X_k, X_ℓ, X_m, X_r, implies that Y_k, Y_ℓ, Y_m, Y_r also have uniform marginal distributions. Thus, it it sufficient to show that if $H(X_k, X_\ell) > H(X_m, X_r)$ then

$$H(Y_k, Y_\ell) > H(Y_m, Y_r). \tag{106}$$

Lemma 4 shows that

$$\mathrm{p}_\dagger(y_k, y_\ell) = (1-q)^2 \mathrm{p}(x_k, x_\ell) + \frac{1-(1-q)^2}{M^2}, \tag{107}$$

$$\mathrm{p}_\dagger(y_m, y_r) = (1-q)^2 \mathrm{p}(x_m, x_r) + \frac{1-(1-q)^2}{M^2}. \tag{108}$$

Then we consider $M^2 = L$ and Lemma 3 gives (106). □

Lemma 6. *Consider $X_w, X_{\bar{w}}, X_u, X_{\bar{u}}$ as inputs variables of the M-ary symmetric channel (defined in Section V-C) with corresponding outputs $Y_w, Y_{\bar{w}}, Y_u, Y_{\bar{u}}$, $\mathbf{X} \sim \mathrm{p}(\cdot) \in \mathcal{P}_{\mathrm{T}}(c_1, c_2)$ and $((w, \bar{w}), u, \bar{u}) \in \mathcal{EV}^2$. If the crossover probability q is sufficiently small ($q \to 0$) then*

$$\begin{aligned}
&I(Y_w; Y_{\bar{w}}) - I(Y_u; Y_{\bar{u}}) \\
&= (1-q)^2 \left[I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}}) \right] \\
&\quad - \left(1-(1-q)^2\right) \left[\mathit{KL}(U || \mathrm{p}(x_w, x_{\bar{w}})) - \mathit{KL}(U || \mathrm{p}(x_u, x_{\bar{u}})) \right] \\
&\quad + \mathcal{O}(\epsilon^2),
\end{aligned}$$

and $\epsilon = [1-(1-q)^2]/M^2$ and U is the uniform distribution on the alphabet $[M]^2$.

Proof. Recall that the marginal distributions of each node variable $X_w, X_{\bar{w}}, X_u, X_{\bar{u}}$ is uniform and this implies that the marginal distributions of the corresponding outputs $Y_w, Y_{\bar{w}}, Y_u, Y_{\bar{u}}$ are uniform as well. Note that $\mathbf{Y} \in [M]^p$ and $\mathbf{X} \in [M]^p$ In addition, the pairwise joint distributions of $\mathbf{Y}$ in terms of the corresponding joint pairwise distributions of $\mathbf{X}$

$$\begin{aligned}
&\mathbb{P}\left(Y_w = x_w, Y_{\bar{w}} = x_{\bar{w}}\right) \\
&= (1-q)^2 \mathrm{p}(x_w, x_{\bar{w}}) + \frac{1-(1-q)^2}{M^2}, \quad x_w, x_{\bar{w}} \in [M]^2, \\
&\mathbb{P}\left(Y_u = x_u, Y_{\bar{u}} = x_{\bar{u}}\right) \\
&= (1-q)^2 \mathrm{p}(x_u, x_{\bar{u}}) + \frac{1-(1-q)^2}{M^2}, \quad x_u, x_{\bar{u}} \in [M]^2.
\end{aligned}$$

We denote the probability mass function of the pairs $X_w, X_{\bar{w}}$ and $X_u, X_{\bar{u}}$ by $\mathrm{p}(x_w, x_{\bar{w}}) \triangleq \mathbb{P}\left(X_w = x_w, X_{\bar{w}} = x_{\bar{w}}\right)$ and $\mathrm{p}(x_u, x_{\bar{u}}) \triangleq \mathbb{P}\left(X_u = x_u, X_{\bar{u}} = x_{\bar{u}}\right)$ for $(x_w, x_{\bar{w}}), (x_u, x_{\bar{u}}) \in [M]^2$ and similarly, for the noisy versions $Y_w, Y_{\bar{w}}$ and $Y_u, Y_{\bar{u}}$, we use $\mathrm{p}_\dagger(x_w, x_{\bar{w}}) \triangleq \mathbb{P}\left(Y_w = x_w, Y_{\bar{w}} = x_{\bar{w}}\right)$ and $\mathrm{p}_\dagger(x_u, x_{\bar{u}}) \triangleq \mathbb{P}\left(Y_u = x_u, Y_{\bar{u}} = x_{\bar{u}}\right)$ for $(x_w, x_{\bar{w}}), (x_u, x_{\bar{u}}) \in [M]^2$. Thus,

$$\begin{aligned}
&I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}}) \\
&= -H(X_w, X_{\bar{w}}) + H(X_u, X_{\bar{u}}) \qquad (109) \\
&= \sum_{x_w, x_{\bar{w}} \in [M]^2} \mathrm{p}(x_w, x_{\bar{w}}) \log \mathrm{p}(x_w, x_{\bar{w}}) \\
&\quad - \sum_{x_u, x_{\bar{u}} \in [M]^2} \mathrm{p}(x_u, x_{\bar{u}} \log \mathrm{p}(x_u, x_{\bar{u}}) \\
&= \sum_{x_w, x_{\bar{w}} \in [M]^2} \mathrm{p}(x_w, x_{\bar{w}}) \log(1-q)^2 \mathrm{p}(x_w, x_{\bar{w}}) \\
&\quad - \sum_{x_u, x_{\bar{u}} \in [M]^2} \mathrm{p}(x_u, x_{\bar{u}}) \log(1-q)^2 \mathrm{p}(x_u, x_{\bar{u}}) \\
&= \Big[\sum_{x_w, x_{\bar{w}} \in [M]^2} (1-q)^2 \mathrm{p}(x_w, x_{\bar{w}}) \log(1-q)^2 \mathrm{p}(x_w, x_{\bar{w}}) \\
&\quad - \sum_{x_u, x_{\bar{u}} \in [M]^2} (1-q)^2 \mathrm{p}(x_u, x_{\bar{u}}) \log(1-q)^2 \mathrm{p}(x_u, x_{\bar{u}}) \Big] \\
&\quad \times (1-q)^{-2}, \qquad (110)
\end{aligned}$$

and (109) holds because the marginal distributions are uniform. Define $\epsilon \triangleq \left(1-(1-q)^2\right)/M^2$, so $0 < \epsilon < q$ for any $M^2 \geq 2$. Similarly to Lemma 3, for any $x_w, x_{\bar{w}} \in [M]^2$

$$\begin{aligned}
&\mathrm{p}_\dagger(x_w, x_{\bar{w}}) \log_2 \mathrm{p}_\dagger(x_w, x_{\bar{w}}) \\
&= \left((1-q)^2 \mathrm{p}(x_w, x_{\bar{w}}) + \epsilon\right) \log_2 \left((1-q)^2 \mathrm{p}(x_w, x_{\bar{w}}) + \epsilon\right) \\
&= (1-q)^2 \mathrm{p}(x_w, x_{\bar{w}}) \log_2 \left((1-q)^2 \mathrm{p}(x_w, x_{\bar{w}}) + \epsilon\right) \\
&\quad + \epsilon \log_2 \left((1-q)^2 \mathrm{p}(x_w, x_{\bar{w}}) + \epsilon\right) \qquad (111) \\
&= (1-q)^2 \mathrm{p}(x_w, x_{\bar{w}}) \log_2 \left((1-q)^2 \mathrm{p}(x_w, x_{\bar{w}})\right) + \epsilon \\
&\quad + \epsilon \log_2 \left((1-q)^2 \mathrm{p}(x_w, x_{\bar{w}})\right) + \mathcal{O}(\epsilon^2). \qquad (112)
\end{aligned}$$

Here the last equality holds because $\log(1 + \epsilon/(1-q)^2 \mathrm{p}(x_w, x_{\bar{w}})) = \epsilon/(1-q)^2 \mathrm{p}(x_w, x_{\bar{w}}) + \mathcal{O}(\epsilon^2)$ for ϵ sufficiently small, while $\mathrm{p}(x_w, x_{\bar{w}})$ and M are considered fixed. Also,

$$\begin{aligned}
&\mathrm{p}_\dagger(x_u, x_{\bar{u}}) \log_2 \mathrm{p}_\dagger(x_u, x_{\bar{u}}) \\
&= (1-q)^2 \mathrm{p}(x_u, x_{\bar{u}}) \log_2 \left((1-q)^2 \mathrm{p}(x_u, x_{\bar{u}})\right) + \epsilon \\
&\quad + \epsilon \log_2 \left((1-q)^2 \mathrm{p}(x_u, x_{\bar{u}})\right) + \mathcal{O}(\epsilon^2). \qquad (113)
\end{aligned}$$

Now, we add and subtract the terms $(1-q)^{-2} \sum_{x_w, x_{\bar{w}}} \epsilon \log_2 \left(2(1-q)^2 \mathrm{p}(x_w, x_{\bar{w}})\right) + \mathcal{O}(\epsilon^2)$ and $(1-q)^{-2} \sum_{x_u, x_{\bar{u}}} \epsilon \log_2 \left(2(1-q)^2 \mathrm{p}(x_u, x_{\bar{u}})\right) + \mathcal{O}(\epsilon^2)$ in (110), and we get

$$\begin{aligned}
&I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}}) \\
&= \Big[\sum_{x_w, x_{\bar{w}} \in [M]^2} (1-q)^2 \mathrm{p}(x_w, x_{\bar{w}}) \log(1-q)^2 \mathrm{p}(x_w, x_{\bar{w}}) \\
&\quad + \epsilon \log_2 \left(2(1-q)^2 \mathrm{p}(x_w, x_{\bar{w}})\right) + \mathcal{O}(\epsilon^2) \\
&\quad - \sum_{x_u, x_{\bar{u}} \in [M]^2} (1-q)^2 \mathrm{p}(x_u, x_{\bar{u}}) \log(1-q)^2 \mathrm{p}(x_u, x_{\bar{u}}) \\
&\quad + \epsilon \log_2 \left(2(1-q)^2 \mathrm{p}(x_u, x_{\bar{u}})\right) + \mathcal{O}(\epsilon^2) \\
&\quad + \sum_{x_u, x_{\bar{u}} \in [M]^2} \epsilon \log_2 \left(2(1-q)^2 \mathrm{p}(x_u, x_{\bar{u}})\right) \\
&\quad - \sum_{x_w, x_{\bar{w}} \in [M]^2} \epsilon \log_2 \left(2(1-q)^2 \mathrm{p}(x_w, x_{\bar{w}})\right) \Big] (1-2q)^{-2} \\
&\quad + \mathcal{O}(\epsilon^2).
\end{aligned}$$

The latter and (112) and (113) give

$$\begin{aligned}
&I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}}) \\
&= \frac{1}{(1-q)^2} \left[-H(Y_w, Y_{\bar{w}}) + H(Y_u, Y_{\bar{u}}) \right]
\end{aligned}$$

$$+ \frac{\epsilon}{(1-q)^2}\Big[\sum_{x_u, x_{\bar{u}} \in [M]^2} \log_2 \left(\mathrm{p}(x_u, x_{\bar{u}})\right) - \sum_{x_w, x_{\bar{w}} \in [M]^2} \log_2 \left(\mathrm{p}(x_w, x_{\bar{w}})\right) \Big] + \mathcal{O}(\epsilon^2)$$

$$= \frac{1}{(1-q)^2}\left[I(Y_w; Y_{\bar{w}}) - I(Y_u; Y_{\bar{u}})\right] + \frac{\epsilon}{(1-q)^2}\Big[\sum_{x_u, x_{\bar{u}} \in [M]^2} \log_2 \left(\mathrm{p}(x_u, x_{\bar{u}})\right) - \sum_{x_w, x_{\bar{w}} \in [M]^2} \log_2 \left(\mathrm{p}(x_w, x_{\bar{w}})\right) \Big] + \mathcal{O}(\epsilon^2), \quad (114)$$

and the latter holds because the marginal distribution of each Y is uniform. The definition of ϵ, $\epsilon \triangleq \left(1-(1-q)^2\right)/M^2$ together with (114) give

$$\begin{aligned}
&I(Y_w; Y_{\bar{w}}) - I(Y_u; Y_{\bar{u}}) \\
&= (1-q)^2 \left[I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}})\right] \\
&\quad - \frac{1-(1-q)^2}{M^2}\Big[\sum_{x_u, x_{\bar{u}} \in [M]^2} \log_2 \left(\mathrm{p}(x_u, x_{\bar{u}})\right) - \sum_{x_w, x_{\bar{w}} \in [M]^2} \log_2 \left(\mathrm{p}(x_w, x_{\bar{w}})\right) \Big] + \mathcal{O}(\epsilon^2) \\
&= (1-q)^2 \left[I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}})\right] \\
&\quad - \frac{1-(1-q)^2}{M^2}\Big[\sum_{x_u, x_{\bar{u}} \in [M]^2} \log_2 \left(\frac{\mathrm{p}(x_u, x_{\bar{u}})}{1/M^2}\right) - \sum_{x_w, x_{\bar{w}} \in [M]^2} \log_2 \left(\frac{\mathrm{p}(x_w, x_{\bar{w}})}{1/M^2}\right) \Big] + \mathcal{O}(\epsilon^2) \\
&= (1-q)^2 \left[I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}})\right] \\
&\quad - \left(1-(1-q)^2\right)\Big[- \sum_{x_u, x_{\bar{u}} \in [M]^2} \frac{1}{M^2} \log_2 \left(\frac{1/M^2}{\mathrm{p}(x_u, x_{\bar{u}})}\right) + \sum_{x_w, x_{\bar{w}} \in [M]^2} \frac{1}{M^2} \log_2 \left(\frac{1/M^2}{\mathrm{p}(x_w, x_{\bar{w}})}\right) \Big] + \mathcal{O}(\epsilon^2) \\
&= (1-q)^2 \left[I(X_w; X_{\bar{w}}) - I(X_u; X_{\bar{u}})\right] \\
&\quad - \left(1-(1-q)^2\right)\left[\mathrm{KL}(U||\mathrm{p}(x_w, x_{\bar{w}})) - \mathrm{KL}(U||\mathrm{p}(x_u, x_{\bar{u}}))\right] \\
&\quad + \mathcal{O}(\epsilon^2).
\end{aligned}$$

This completes the proof. □

REFERENCES

[1] S. L. Lauritzen, *Graphical Models*, ser. Oxford Statistical Science Series. New York, NY, USA: The Clarendon Press, Oxford University Press, 1996, vol. 17.

[2] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[3] C. Chow and C. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 462–467, May 1968. [Online]. Available: https://ieeexplore.ieee.org/document/1054142

[4] V. Y. F. Tan, A. Anandkumar, and A. S. Willsky, "Learning Gaussian tree models: Analysis of error exponents and extremal structures," *arXiv preprint arXiv:0909.5216*, Jan 2010.

[5] G. Bresler and M. Karzand, "Learning a tree-structured Ising model in order to make predictions," *arXiv preprint arXiv:1604.06749*, Jun 2018. [Online]. Available: https://arxiv.org/abs/1604.06749

[6] V. Y. F. Tan, A. Anandkumar, L. Tong, and A. S. Willsky, "A large-deviation analysis of the maximum-likelihood learning of Markov tree structures," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1714–1735, Feb 2011. [Online]. Available: https://ieeexplore.ieee.org/document/5714274

[7] V. Y. F. Tan, A. Anandkumar, and A. S. Willsky, "Learning high-dimensional Markov forest distributions: Analysis of error rates," *Journal of Machine Learning Research*, vol. 12, pp. 1617–1653, May 2011. [Online]. Available: http://www.jmlr.org/papers/v12/tan11a.html

[8] V. Y. F. Tan, "Large-deviation analysis and applications of learning tree-structured graphical models," Ph.D. dissertation, Massachusetts Institute of Technology, 2011. [Online]. Available: http://hdl.handle.net/1721.1/64486

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: John Wiley & Sons, 2012. [Online]. Available: https://www.wiley.com/en-it/Elements+of+Information+Theory,+2nd+Edition-p-9780471241959

[10] S. Goel, D. M. Kane, and A. R. Klivans, "Learning Ising models with independent failures," *arXiv preprint arXiv:1902.04728*, Feb 2019. [Online]. Available: https://arxiv.org/abs/1902.04728

[11] K. E. Nikolakakis, D. S. Kalogerias, and A. D. Sarwate, "Learning tree structures from noisy data," in *The 22nd International Conference on Artificial Intelligence and Statistics*, Apr 2019, pp. 1771–1782. [Online]. Available: http://proceedings.mlr.press/v89/nikolakakis19a.html

[12] M. Vuffray, S. Misra, A. Lokhov, and M. Chertkov, "Interaction screening: Efficient and sample-optimal learning of ising models," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 2595–2603.

[13] D. Karger and N. Srebro, "Learning Markov networks: Maximum bounded tree-width graphs," in *Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, Jan 2001, pp. 392–401. [Online]. Available: https://dl.acm.org/citation.cfm?id=365486

[14] S.-I. Lee, V. Ganapathi, and D. Koller, "Efficient structure learning of markov networks using ℓ_1-regularization," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J. C. Platt, and T. Hoffman, Eds. MIT Press, 2007, pp. 817–824.

[15] G. Bresler, E. Mossel, and A. Sly, "Reconstruction of Markov random fields from samples: Some observations and algorithms," *Lecture Notes in Computer Science*, vol. 5171, pp. 343–356, 2008. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-85363-3_28

[16] P. Ravikumar, M. J. Wainwright, J. D. Lafferty *et al.*, "High-dimensional Ising model selection using ℓ_1-regularized logistic regression," *The Annals of Statistics*, vol. 38, no. 3, pp. 1287–1319, 2010. [Online]. Available: https://projecteuclid.org/euclid.aos/1268056617

[17] G. Bresler, D. Gamarnik, and D. Shah, "Hardness of parameter estimation in graphical models," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014, pp. 1062–1070.

[18] G. Bresler, "Efficiently learning Ising models on arbitrary graphs," in *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*. ACM, Jun 2015, pp. 771–782. [Online]. Available: https://dl.acm.org/citation.cfm?id=2746631

[19] M. Vuffray, S. Misra, and A. Y. Lokhov, "Efficient learning of discrete graphical models," *arXiv preprint arXiv:1902.00600*, Feb 2019. [Online]. Available: https://arxiv.org/abs/1902.00600

[20] N. P. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4117–4134, Mar 2012. [Online]. Available: https://ieeexplore.ieee.org/document/6172586

[21] R. Tandon, K. Shanmugam, P. K. Ravikumar, and A. G. Dimakis, "On the information theoretic limits of learning ising models," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2303–2311.

[22] L. Hamilton, F. Koehler, and A. Moitra, "Information theoretic properties of Markov random fields, and their algorithmic applications," *arXiv preprint arXiv:1705.11107*, May 2017. [Online]. Available: https://arxiv.org/abs/1705.11107

[23] H. Liu, M. Xu, H. Gu, A. Gupta, J. Lafferty, and L. Wasserman, "Forest density estimation," *Journal of Machine Learning Research*, vol. 12, no. Mar, pp. 907–951, 2011. [Online]. Available: http://www.jmlr.org/papers/v12/liu11a.html

[24] K. E. Nikolakakis, D. S. Kalogerias, and A. D. Sarwate, "Predictive learning on sign-valued hidden Markov trees," *arXiv preprint arXiv:1812.04700*, Feb 2019. [Online]. Available: https://arxiv.org/abs/1812.04700

[25] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of the American Statistical Association*, vol. 60, no. 309, pp. 63–69, Mar 1965. [Online]. Available: https://www.jstor.org/stable/2283137?seq=1#page_scan_tab_contents

[26] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, ser. PODS '03. New York, NY, USA: ACM, 2003, pp. 211–222. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.7574

[27] S. A. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" in *49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, Oct. 2008, pp. 531–540. [Online]. Available: http://dx.doi.org/10.1109/FOCS.2008.27

[28] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM Journal on Computing*, vol. 40, no. 3, pp. 793–826, Jun 2011. [Online]. Available: https://ieeexplore.ieee.org/document/4690986

[29] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, Oct. 2013, pp. 429–438. [Online]. Available: https://ieeexplore.ieee.org/document/6686179

[30] A. Smith, A. Thakurta, and J. Upadhyay, "Is interaction necessary for distributed private learning?" in *Proceedings of the 2017 IEEE Symposium on Security and Privacy*, May 2017, pp. 58–77. [Online]. Available: https://ieeexplore.ieee.org/document/7958571

[31] M. Tavassolipour, S. A. Motahari, and M.-T. M. Shalmani, "Learning of tree-structured Gaussian graphical models on distributed data under communication constraints," *arXiv preprint arXiv:1809.08067*, Sep 2018. [Online]. Available: https://arxiv.org/abs/1809.08067

[32] R. W. Yeung, A. Al-Bashabsheh, C. Chen, Q. Chen, and P. Moulin, "On information-theoretic characterizations of Markov random fields and subfields," *IEEE Transactions on Information Theory*, vol. 65, no. 3, pp. 1493–1511, Aug 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8444473

[33] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Structures & Algorithms*, vol. 19, no. 3-4, pp. 163–193, Nov 2001. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.10019

[34] J. Silva, "Shannon entropy estimation in ∞ alphabets from convergence results: Studying plug-in estimators," *Entropy*, vol. 20, no. 6, p. 397, 2018. [Online]. Available: https://www.mdpi.com/1099-4300/20/6/397

[35] M. Raginsky, "Strong data processing inequalities and phi-sobolev inequalities for discrete channels," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3355–3389, 2016.

[36] Y. Polyanskiy and Y. Wu, "Strong data-processing inequalities for channels and bayesian networks," in *Convexity and Concentration*. Springer, 2017, pp. 211–249.

[37] A. B. Tsybakov, "Introduction to nonparametric estimation. revised and extended from the 2004 french original. translated by vladimir zaiats," 2009.