SGA: A Robust Algorithm for Partial Recovery of Tree-Structured Graphical Models with Noisy Samples

Anshoo Tandon [1] **Aldric J. Y. Han** [2] **Vincent Y. F. Tan** [1 2]

Abstract

We consider learning Ising tree models when the observations from the nodes are corrupted by independent but non-identically distributed noise with unknown statistics. Katiyar et al. (2020) showed that although the *exact* tree structure cannot be recovered, one can recover a *partial* tree structure; that is, a structure belonging to the equivalence class containing the true tree. This paper presents a systematic improvement of Katiyar et al. (2020). First, we present a novel impossibility result by deriving a bound on the necessary number of samples for partial recovery. Second, we derive a significantly improved sample complexity result in which the dependence on the minimum correlation $\rho_{\min}$ is $\rho_{\min}^{-8}$ instead of $\rho_{\min}^{-24}$. Finally, we propose Symmetrized Geometric Averaging (SGA), a more statistically robust algorithm for partial tree recovery. We provide error exponent analyses and extensive numerical results on a variety of trees to show that the sample complexity of SGA is significantly better than the algorithm of Katiyar et al. (2020). SGA can be readily extended to Gaussian models and is shown via numerical experiments to be similarly superior.

1. Introduction

Graphical models provide a succinct diagrammatic representation of the dependencies among a set of random variables. The vertices of the graph are in one-to-one correspondence with the random variables, while the edges encode conditional independence relationships. Graphical models have found applications in domains from biology (Friedman, 2004), to coding theory (Kschischang & Frey, 1998), social networks (Lauritzen, 1996) and computer vision (Besag, 1986). For a detailed description of graphical models, the reader is referred to Wainwright & Jordan (2008).

This paper concerns the learning of the structure of a certain class of graphical models from data (Johnson et al., 2007). In particular, we consider undirected Ising graphical models that are Markov on trees. While this is a classical problem that has been studied extensively (Chow & Liu, 1968; Bresler & Karzand, 2020), here we focus on a relatively unexplored problem in which the vector-valued samples presented to the learner are *corrupted by independent but non-identically distributed noise*. This scenario is motivated by two different real-world scenarios. Firstly, suppose each component is obtained at spatially distributed locations of a large sensor network. The scalar-valued components need to be transmitted to a fusion center for the reconstruction of the graphical structure of the data. Due to the presence of disturbances (e.g., pollution, ambient noise), the components will inevitably be corrupted and the corruption levels are different as the distances of the spatially distributed devices to the fusion center are different. Secondly, imagine that a drug company would like to understand the inter-dependencies between different chemicals to design an effective vaccine given training samples obtained from human subjects. To thwart the company's attempts, a competitor corrupts certain measurements or features of the human subjects (Wang & Gu, 2017), where these corruptions may be non-identically distributed. In these scenarios, we would like to design robust algorithms to recover, as best as possible, a certain structure that is "close" to the true one.

Motivated by these examples, we consider a tree learning problem in which each component of the vector-valued samples may be corrupted in a non-identical but independent (across components) manner. Because the corruption noises are non-identical, the ordering of the observed correlations are distorted, and hence the tree, in general, cannot be identified using the maximum likelihood Chow-Liu algroithm (Chow & Liu, 1968). More precisely Katiyar et al. (2020) showed that in such a case, even in the infinite sample limit, the structure can only be identified up to its equivalent class, a notion that will be defined precisely in the sequel. Building on their previous work for the robust learning of

[1]Department of Electrical & Computer Engineering, National University of Singapore, Singapore. [2]Department of Mathematics, National University of Singapore, Singapore. Correspondence to: Anshoo Tandon <anshoo.tandon@gmail.com>, Vincent Y. F. Tan <vtan@nus.edu.sg>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

Gaussian graphical models (Katiyar et al., 2019), Katiyar et al. (2020) proposed a certain algorithm for the learning of Ising tree-structured models in which the samples are observed in noise. They provided a sample complexity result to recover the true tree up to its equivalence class. They showed that in this adversarial scenario, the go-to algorithm for learning tree-structure models – the Chow-Liu algorithm (Chow & Liu, 1968) – fails miserably but their algorithm is able to learn the model up to its equivalence class if the number of samples is sufficiently large.

We significantly improve on theoretical and algorithmic results in Katiyar et al. (2020) and Nikolakakis et al. (2019a).

- Firstly, we provide an information-theoretic impossibility result for tree structure recovery under non-identically distributed noise, that extends the result by Nikolakakis et al. (2019a) which assumes i.i.d. noise. Our impossibility result, which involves the construction and analysis of sufficiently many "nearby" trees, elucidates the effect of noise on the sample complexity even as the number of samples and the number of nodes grow without bound; this desirable feature is not present in previous works on learning graphical models with noisy samples Nikolakakis et al. (2019a).
- Secondly, by a careful analysis of the error events in the algorithm proposed by Katiyar et al. (2020), we significantly improve on the sample complexity result contained therein. In particular, the dependence of the sample complexity on the minimum correlation along the tree edges, $\rho_{\min}$, is improved from $\rho_{\min}^{-24}$ to $\rho_{\min}^{-8}$.
- Finally, and most importantly, we propose an improvement to the IS_NON_STAR subroutine in Katiyar et al. (2020); we call our subroutine *Symmetrized Geometric Averaging* (SGA). This improvement is motivated by symmetry considerations in the error events in distinguishing a star versus non-star structure and the folklore theorem that "more averaging generally helps". Indeed, we show through error exponent analyses (Tan et al., 2011) using Sanov's theorem (Cover & Thomas, 2006) that the error exponents of SGA are higher (and hence better) than that in Katiyar et al. (2020) for all but a small fraction of trees. This is corroborated by extensive numerical experiments for a variety of trees of different structures, correlations, and corruption noises. SGA is also shown to be amenable to improving the structure learning of *Gaussian* in addition to Ising trees.

1.1. Related Work

The learning of tree-structured graphical models dates back to the seminal work of Chow & Liu (1968) who showed that the maximum likelihood estimation of the tree structure is equivalent to that of a maximum weight spanning tree problem with edge weights given by the empirical mutual information. Chow & Wagner (1973) showed that structure learning is consistent and Tan et al. (2011) derived the error exponent. The estimates of the error probability of learning the tree structure was further refined by Tandon et al. (2020). Bresler & Karzand (2020) considered a variant of the problem in which a tree was learned to make predictions instead of the traditional objective of inferring the structure.

Tandon et al. (2020) showed that the Chow-Liu algorithm is error exponent-optimal if the noises that corrupt the observations are independent and identically distributed. Nikolakakis et al. (2019b) considered the learning of tree structures when the noise is possibly non-identically distributed, and derived conditions under which structure learning can be achieved using the Chow-Liu algorithm. We note that this setting is in contrast to recent work on robust tree learning under *adversarial* noise (Cheng et al., 2018); in our work, *random* noise is added to clean samples.

Katiyar et al. (2019) showed that it is, in general, not possible to learn the exact tree structure for Gaussian graphical models when one is given independent but non-identically distributed noisy samples with unknown noise distribution. This work was followed by Katiyar et al. (2020), who considered Ising tree models and derived an algorithm based on so-called *proximal sets* to learn the equivalence class (or clusters) of the true tree.

Finally, we remark that there is a large body of literature on learning latent tree models (e.g., Choi et al. (2011); Parikh et al. (2011)) in which one observes a subset of nodes from a tree. The marginal distribution of those observed nodes is, in general, not a tree. A formal connection between learning latent tree models and learning *noisy* tree models was recently established by Casanellas et al. (2021), and is briefly discussed in Sec. 2.4.

2. Preliminaries and Problem Statement

An *undirected graphical model* is a multivariate probability distribution that factorizes according to the structure an undirected graph (Lauritzen, 1996). Specifically, a d-dimensional random vector $\mathbf{X} \triangleq (X_1, X_2, \ldots, X_d)$ is said to be *Markov* on $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex (or node) set $\mathcal{V} = \{1, 2, \ldots, d\}$ and edge set $\mathcal{E} \subset \binom{\mathcal{V}}{2}$ if its distribution satisfies the (local) Markov property $P(x_i|x_{\mathcal{V}\setminus i}) = P(x_i|x_{\mathrm{nbd}(i)})$ where $\mathrm{nbd}(i) := \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}\}$ is the *neighborhood* of node i. In this work, we focus on tree-structured graphical models P, where the underlying graph of P is an acyclic and connected (tree) graph, denoted by $\mathrm{T} = \mathrm{T}_P = (\mathcal{V}, \mathcal{E}_P)$ with $|\mathcal{V}| = d$ and $|\mathcal{E}_P| = d - 1$. For an undirected tree, we may assume, without loss of generality, that node 1 is the *root* node and we arrange all the other nodes at different levels on a plane, with node 1 at level-0. Then, the tree-structured graphical model P can be

alternatively factored as (Chow & Liu, 1968)

$$P(\mathbf{x}) = P_1(x_1)\prod_{i=2}^{d} P_{i|\mathrm{pa}(i)}(x_i|x_{\mathrm{pa}(i)}), \tag{1}$$

where $\mathrm{pa}(i)$ (with $|\mathrm{pa}(i)| = 1$) is the parent node of node i.

We denote the KL-divergence between distributions Q and P as $D(Q\|P) = \sum_{\mathbf{x}} Q(\mathbf{x}) \log \frac{Q(\mathbf{x})}{P(\mathbf{x})}$. The set of distributions supported on alphabet $\mathcal{X}$ is denoted as $\mathcal{P}(\mathcal{X})$. Proofs of the theorems and propositions are provided in the appendices, as part of the supplementary material. For the sake of brevity, the presentation in the main body of the paper is focused on Ising tree models; the extension to the Gaussian case with experiments is deferred to App. J.

2.1. System Model

We consider binary random variables X_i with alphabet $\mathcal{X} = \{+1, -1\}$, where $1 \le i \le d$ and the joint distribution of $(X_1, \ldots, X_d)$ is represented by a tree-structured Ising model (Bresler & Karzand, 2020). The observation for the ith node is represented by random variable $Y_i = X_i N_i$, where N_i is multiplicative binary noise with $\Pr(N_i = -1) = q_i$ and $\Pr(N_i = +1) = 1 - q_i$. Thus, the observations are corrupted by independent but non-identical noise; that is, q_i may differ for different values of i. We assume that our model satisfies the following properties:

- P1: (*Zero external field*): The marginals for the hidden variables are uniform, i.e., $\Pr(X_i = 1) = \Pr(X_i = -1) = 0.5$, for $1 \le i \le d$.
- P2: (*Bounded Correlation*): If $\mathcal{E}$ denotes the edge set of tree T, and $\{i, j\} \in \mathcal{E}$, then the correlations $\rho_{i,j} \triangleq \mathbb{E}[X_i X_j]$ are uniformly bounded as follows: $\rho_{\min} \le |\rho_{i,j}| \le \rho_{\max}$, where $0 < \rho_{\min} \le \rho_{\max} < 1$.
- P3: (*Bounded Noise*): The noise crossover probability q_i, for $1 \le i \le d$, satisfies $0 \le q_i \le q_{\max} < 0.5$.

Properties P1 and P2 are common assumptions in the literature on learning Ising models (Tandon et al., 2014; Scarlett & Cevher, 2016; Nikolakakis et al., 2019a; Bresler & Karzand, 2020), while P3 ensures that no node is independent of any other node due to noise (Katiyar et al., 2020). For an Ising model with zero external field, the joint distribution given by (1) can be expressed as (Bresler & Karzand, 2020)

$$P(\mathbf{x}) = \frac{1}{Z}\exp\left(\sum_{\{i,j\}\in\mathcal{E}} \theta_{i,j} x_i x_j\right), \tag{2}$$

where $\mathcal{E}$ is the edge set of the graph T, and Z is the normalization factor. For an Ising tree, if $\{i, j\} \in \mathcal{E}$ then the interaction (exponential) parameter $\theta_{i,j}$ is related to the correlation $\rho_{i,j} = \mathbb{E}[X_i X_j]$ as (Nikolakakis et al., 2019c, Lem. A.3)

$$\theta_{i,j} = \mathrm{atanh}(\rho_{i,j}). \tag{3}$$

The bounded correlation property P2 then implies that $\mathrm{atanh}(\rho_{\min}) \le |\theta_{i,j}| \le \mathrm{atanh}(\rho_{\max})$. Note that although property P1 is a common assumption that simplifies the presentation (Tandon et al., 2014; Scarlett & Cevher, 2016; Nikolakakis et al., 2019a; Bresler & Karzand, 2020), the extension of our results to the case where the marginals are not necessarily uniform can be readily obtained by following the approach outlined by Katiyar et al. (2020).

2.2. Definition of Equivalent Tree Structures

We now define an equivalence relation on d-node trees. Let $\mathcal{T}_d$ denote the set of all distinct trees on d nodes. For a given tree graph, a *leaf* is a node whose degree is one, i.e. a leaf node has only one neighbor. For $\mathrm{T} \in \mathcal{T}_d$, let $\mathcal{L}_{\mathrm{T}}$ denote set of leaf nodes in T, $\mathcal{L}_{\mathrm{T}} \triangleq \{X_i : i \text{ is a leaf node in T}\}$. Let $\mathscr{S}_{\mathrm{T}}$ be the set of subsets of $\mathcal{L}_{\mathrm{T}}$ such that no two nodes in a subset share a common neighbor, i.e., $\mathscr{S}_{\mathrm{T}} \triangleq \{\mathcal{S} \subseteq \mathcal{L}_{\mathrm{T}} : \text{no two nodes in } \mathcal{S} \text{ have the same neighbor}\}$. For a given $\mathcal{S} \in \mathscr{S}_{\mathrm{T}}$, let $\mathrm{T}_{\mathcal{S}}$ denote the tree obtained from T by interchanging each node in $\mathcal{S}$ with its corresponding neighbor in T.[1] Define $[\mathrm{T}]$ to be the set of trees

$$[\mathrm{T}] \triangleq \{T_{\mathcal{S}} : \mathcal{S} \in \mathscr{S}_{\mathrm{T}}\}. \tag{4}$$

For $\hat{\mathrm{T}} \in \mathcal{T}_d$, we say $\hat{\mathrm{T}} \sim \mathrm{T}$ if $\hat{\mathrm{T}} \in [\mathrm{T}]$. The relation $\sim$ on $\mathcal{T}_d$ is *reflexive*, *symmetric*, and *transitive*, and is hence an *equivalence relation* (Herstein, 1975). Therefore, with respect to relation $\sim$, the set $\mathcal{T}_d$ is partitioned into disjoint equivalence classes, where $[\mathrm{T}]$ is the equivalence class of T.

2.3. Problem Statement

Let $\mathbf{y}_1^n = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ denote n independently sampled noisy observations, where the ith noisy sample is a d-dimensional column vector given by $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,d})^T$ with $y_{i,j}$ denoting the ith observation corresponding to the jth node. As discussed in Sec. 2.1, we assume a binary multiplicative non-identical noise model at each node, with $y_{i,j} \in \mathcal{Y} \triangleq \{+1, -1\}$ and $\mathbf{y}_i \in \mathcal{Y}^d$. Given $\mathbf{y}_1^n$, a learning algorithm (or estimator) $\Psi : \mathcal{Y}^{d\times n} \to \mathcal{T}_d$ provides an estimate of the underlying tree structure T. We are interested in partial tree recovery (up to equivalence class $[\mathrm{T}]$), and an error is declared if the event $\{\Psi(\mathbf{Y}_1^n) \notin [\mathrm{T}]\}$ occurs.[2] For this setup, the following questions are of interest. (a) Quantify the number of samples *necessary* to achieve a target error probability (with any possible estimator). (b) Quantify the number of samples *sufficient* to achieve a given error probability (using a particular estimator). We answer (a) in Section 3 and (b) in Sections 4 and 5. The estimator only knows $q_{\max}$, but does not know the individual q_i's, the noise

[1] Note that when $\mathcal{S}$ is the empty set, then $\mathrm{T}_{\mathcal{S}} = \mathrm{T}$.

[2] Exact tree-structure identification cannot be guaranteed, in general, when the noise distribution across nodes is unknown and non-identical across nodes (Katiyar et al., 2020, Thm. 2).

parameters corresponding to each node.

2.4. Relation with Latent Tree Recovery

The recovery of partial tree structure (up to equivalence class $[\mathrm{T}]$), when node samples are subjected to non-identical noise, is also possible using latent tree recovery algorithms by suppressing the degree two nodes in the extended tree obtained by adding extra nodes representing the noisy variables (Casanellas et al., 2021). This approach outlined in Casanellas et al. (2021) generalizes the identifiability result by Katiyar et al. (2020) to the setting where the random variables corresponding to the nodes belong to an arbitrary discrete alphabet. Comparison of the sample complexity for our proposed SGA algorithm to that of latent tree learning algorithms in discussed in Sec. 5.2.

3. Impossibility Result

For a given tree $\mathrm{T} = (\mathcal{V}, \mathcal{E})$, and edge correlation bounds $0 < \rho_{\min} \leq \rho_{\max} < 1$, let $\mathcal{P}_{\mathrm{T}}(\rho_{\min}, \rho_{\max})$ denote the set of all tree-structured Ising models satisfying properties P1 and P2. Let the noise crossover probability at each node satisfy property P3, and hence be upper bounded by $q_{\max}$. Now, given n independent noisy samples, the *minimax error probability* for partial tree structure recovery up to equivalence class $[\mathrm{T}]$, denoted $\mathcal{M}_n(q_{\max}, \rho_{\min}, \rho_{\max})$, is

$$\inf_{\Psi: \mathcal{Y}^{d\times n} \to \mathcal{T}_d} \sup_{\substack{\mathrm{T}\in\mathcal{T}_d, P\in\mathcal{P}_{\mathrm{T}}(\rho_{\min},\rho_{\max}),\\ 0\leq q_i \leq q_{\max}<0.5}} \mathbb{P}_P\big(\Psi(\mathbf{Y}_1^n) \notin [\mathrm{T}]\big), \quad (5)$$

where $\mathbb{P}_P(\cdot)$ denotes the probability measure of the samples when the underlying tree distribution is P.

Theorem 1 (Necessary Samples for Partial Tree Recovery)**.** *Let $\rho_q \triangleq (1-2q_{\max})\rho_{\min}$. If $d > 32$, and n satisfies*

$$n < \frac{\log d}{4\,(1-\rho_{\max})\,\rho_q\,\operatorname{atanh}(\rho_q)}, \quad (6)$$

then the minimax error $\mathcal{M}_n(q_{\max}, \rho_{\min}, \rho_{\max}) \geq 1/2$. In other words, the optimal sample complexity satisfies $\Omega\left((\log d)/[(1-\rho_{\max})(1-2q_{\max})^2\rho_{\min}^2]\right)$.

Theorem 1 is proved by combining two key ingredients. The first is the choice of a sufficiently large number of tree structures that are relatively close each other and their respective equivalence classes are disjoint; see Fig. 6 in App. A. The second key ingredient is the choice of noise parameters for different nodes that have a high impact on the error probability, while ensuring that the corresponding KL-divergence is approximated by a closed-form expression; see Fig. 7 in App. A. Theorem 1 also provides an impossibility result for *exact* tree structure recovery under non-identically distributed noise.

In a related work, a bound on the necessary number of samples required for *exact* tree structure recovery for the *noiseless* setting was presented in Bresler & Karzand (2020, Thm. 3.1). Our result in Theorem 1, specialized to the case where $q_{\max} = 0$, gives the same bound on the number of necessary samples as Bresler & Karzand (2020, Thm. 3.1). Therefore, in the noiseless setting, the partial tree structure recovery (up to equivalence class $[\mathrm{T}]$) is not easier than *exact* structure recovery, in the minimax sense.[3]

In another related work, Nikolakakis et al. (2019a) provided a bound on the number of samples necessary for learning the *exact* tree structure, under the assumption that the noise distribution for all the nodes is *identical*. In particular, for a tree-structured Ising model, it was shown (Nikolakakis et al., 2019a, Thm. 2) that the impact of noise gets manifested as a multiplicative factor $[1-(4q(1-q))^d]^{-1}$, where q denotes the noise crossover probability. This result implies that for any $q \in (0, 0.5)$, the impact of noise becomes negligible, i.e. the multiplicative factor tends to 1, *as the number of nodes d tends to infinity*. In contrast, our result in Theorem 1 for *non-identical* noise, with $0 \leq q_i \leq q_{\max} < 0.5$ shows that the necessary number of samples for $q_{\max} > 0$ is strictly and uniformly greater than the number of samples for the noiseless setting by a multiplicative factor of *at least* $(1-2q_{\max})^{-2}$ *irrespective of how large the value of d is*.

For a given $P \in \mathcal{P}_{\mathrm{T}}(\rho_{\min}, \rho_{\max})$, the error probability $\mathbb{P}_P\big(\Psi(\mathbf{Y}_1^n) \notin [\mathrm{T}]\big)$ will depend on the specific tree structure (Tan et al., 2011; Tandon et al., 2020), and the size of the equivalence class $[\mathrm{T}]$. The minimum and maximum possible size of the equivalence class, for a given number of nodes d, is quantified in App. B.

4. Algorithm by Katiyar et al. (2020) and our Improved Sufficiency Result

Let $\mathrm{T} = (\mathcal{V}, \mathcal{E})$ be a given tree, and let the underlying sample distribution and noise satisfy the properties in Sec. 2.1. Then, even if the noise is potentially large, and the noise statistics are unknown, the tree structure can be partially recovered (up to equivalence class $[\mathrm{T}]$) using the algorithm by Katiyar et al. (2020). This algorithm for partial structure recovery of Ising tree models extends a previous method for partial recovery of Gaussian tree models using noisy samples (Katiyar et al., 2019). The cornerstone of the partial tree structure learning algorithm for Ising tree models (Katiyar et al., 2020) is the classification of any set of 4 distinct nodes in $\mathcal{V}$ as *non-star* or *star*.[4]

Definition 1 (Non-star and star (Katiyar et al., 2020))**.** *Any set of 4 distinct nodes in $\mathcal{V}$ forms a* non-star *if there exists*

[3] A related observation was made by Scarlett & Cevher (2016) in terms of comparison of the number of necessary samples for *partial* recovery (to within a given *edit distance*), with the number of necessary samples for *exact* recovery, in the minimax sense.

[4] An overview of the algorithm classifying a set of 4 nodes as non-star or star is presented in App. C.

at least one edge in $\mathcal{E}$ which, when removed, splits the tree into two sub-trees such that exactly 2 of the 4 nodes lie in one sub-tree and the other 2 nodes lie in the other sub-tree. The nodes in the same sub-tree form a pair. *If the set is not a* non-star, *it is categorized as a* star.

A salient feature of the partial tree structure learning algorithm described by Katiyar et al. (2020) is that of "proximal sets". If we let $t_1 \triangleq (1-2q_{\max})^2\rho_{\min}^4$, and $t_2 \triangleq \min\left\{t_1, \frac{t_1(1-2q_{\max})}{\rho_{\max}}\right\}$, then the proximal set of node i is defined as the set of all nodes j that satisfy $|\widehat{\rho}_{i,j}| \geq 0.5t_2$, where $\widehat{\rho}_{i,j}$ denotes the empirical correlation between nodes i and j. For making a star/non-star categorization, the algorithm by Katiyar et al. (2020) only considers nodes that lie within each others' proximal sets, and they use this property to prove the following theorem on the *sufficient* number of noisy samples required to achieve a given error probability. The result is stated in terms t_2 and $\alpha \triangleq (1+\rho_{\max}^2)/2$.

Theorem 2 (Sufficient Sample Complexity Bound (Katiyar et al., 2020))**.** *The equivalence class* $[\mathrm{T}]$ *can be correctly recovered with probability at least* $1-\tau$ *using the algorithm by Katiyar et al. (2020) if the number of noisy samples* n *satisfy* $n \geq \frac{128}{\delta^2}\log\left(\frac{6d^2}{\tau}\right)$*, where* $\delta \triangleq \frac{t_2^3(1-\alpha)}{128}$.

From the statement of the above theorem, we observe that $\delta = \frac{t_2^3(1-\alpha)}{128} \propto \rho_{\min}^{12}$ because t_2 scales linearly with t_1 while $t_1 \propto \rho_{\min}^4$, thereby implying that n scales as $\rho_{\min}^{-24}$. We show that Theorem 2 can be significantly improved. Indeed, n only needs to scale as $\rho_{\min}^{-8}$.

Theorem 3 (Improved Sample Complexity Bound)**.** *The equivalence class* $[\mathrm{T}]$ *can be correctly recovered with probability at least* $1-\tau$ *using the algorithm by Katiyar et al. (2020) if the number of samples satisfies*

$$n \geq \frac{2}{\tilde{\delta}^2}\log\left(\frac{d^2}{\tau}\right) \quad \textit{where} \quad \tilde{\delta} \triangleq \frac{t_2(1-\alpha)}{20}. \tag{7}$$

Compared to Theorem 2, this significantly improved result (because the right-hand-side is $O(\tilde{\delta}^{-2})$ instead of $O(\delta^{-2})$) is obtained by refining probability bounds for events such as $\frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} < \alpha$ and $\frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}} > \alpha$. In contrast to Theorem 3, the impossibility result in Theorem 1 was derived assuming the knowledge of the noise statistics $\{q_i\}_{i=1}^d$, and hence the sample complexity differs in how it scales with different parameters $(\rho_{\min}, \rho_{\max}, q_{\max})$. The scenario in which the estimator knows the noise statistics and further comparisons of various sample complexities are discussed in App. D.1.

5. Symmetrized Geometric Averaging (SGA)

We now present a modified procedure for declaring a 4-node sub-tree as a star or non-star. This algorithm is denoted SGA_IS_NON_STAR (see Algorithm 1), and is a symmetrized variant of the corresponding algorithm by Katiyar et al. (2020) with additional geometric averaging. The motivation behind SGA can be seen by considering an example where $\{X_1, X_2, X_3, X_4\}$ forms a non-star with pair $\{X_1, X_2\}$. If the noisy correlations are denoted $\tilde{\rho}_{i,j} \triangleq \mathbb{E}[Y_iY_j] = (1-2q_i)(1-2q_j)\rho_{i,j}$, we have $\frac{\tilde{\rho}_{1,3}\,\tilde{\rho}_{2,4}}{\tilde{\rho}_{1,2}\,\tilde{\rho}_{3,4}} \leq \rho_{\max}^2$ and $\frac{\tilde{\rho}_{1,4}\,\tilde{\rho}_{2,3}}{\tilde{\rho}_{1,2}\,\tilde{\rho}_{3,4}} \leq \rho_{\max}^2$. Hence, we would expect the following metrics, based on empirical correlations, to satisfy

$$\text{(i)}\ \frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} < \alpha \quad \text{and} \quad \text{(ii)}\ \frac{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} < \alpha. \tag{8}$$

In contrast to Katiyar et al. (2020), who checks condition (i) in (8) but ignores (ii), the SGA variant compares the geometric average of the metrics in (i) and (ii) against the threshold α for checking if nodes $\{X_1, X_2\}$ form a pair.

Algorithm 1 SGA_IS_NON_STAR

Let the set of 4 nodes be $\{X_1, X_2, X_3, X_4\}$
Input: Empirical correlations $\widehat{\rho}_{i,j}$, $1 \leq i < j \leq 4$, Threshold $\alpha = (1+\rho_{\max}^2)/2$.
Let $v_2 = \frac{\sqrt{|\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}\,\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}|}}{|\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}|}$, $v_3 = \frac{\sqrt{|\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}\,\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}|}}{|\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}|}$, $v_4 = \frac{\sqrt{|\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}\,\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}|}}{|\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}|}$
Let $v = \min_{2\leq i\leq 4} v_i$ and $i^* = \arg\min_{2\leq i\leq 4} v_i$
if $v < \alpha$ **then**
  Declare Non-star where $\{X_1, X_{i^*}\}$ forms a pair
else
  Declare Star
end if

Proposition 1 (Sufficient Sample Complexity Bound)**.** *The equivalence class* $[\mathrm{T}]$ *can be correctly recovered with probability at least* $1-\tau$ *using the SGA_IS_NON_STAR procedure in Algorithm 1 if the number of samples satisfies* (7).

Intuitively, we expect that taking the geometric average of the metrics (i) and (ii) in (8) reduces the effect of noise, hence improving robustness. Although we obtain the same sample complexity bound in Prop. 1 as Theorem 3 because the Hoeffding's inequality used is rather loose (and not distribution dependent), the error exponent analysis in Sec. 6 highlights the advantage of SGA. Furthermore, Monte Carlo simulations with a variety of tree structures and parameters demonstrate SGA's superior robustness in Sec. 7.

5.1. Extension to Gaussian Models

SGA and the algorithm by Katiyar et al. (2020) are applicable to wider classes of models such as Gaussian graphical models in which node observations are corrupted by independent but non-identically distributed Gaussian noise (Katiyar et al., 2019). That is $\mathbf{X} = (X_1, \ldots, X_d)$ follows a zero-mean Gaussian with covariance matrix $\mathbf{\Sigma}^*$ and $(\mathbf{\Sigma}^*)^{-1}$

has sparsity pattern that corresponds to a tree T. However, we observe $\mathbf{Y} = (Y_1, \ldots, Y_d)$ with covariance matrix $\mathbf{\Sigma}^* + \mathbf{D}^*$ where $\mathbf{D}^*$ is an unknown non-negative diagonal matrix. If $\mathbf{D}^*$ is non-zero, the structure of T cannot be identified in general. However, by computing the empirical correlations in an analogous fashion, we show in App. J that SGA is similarly robust vis-à-vis algorithms proposed in Katiyar et al. (2019) and Katiyar et al. (2020).

5.2. Comparing the sample complexity of SGA to that of latent tree learning algorithms

As mentioned in Sec. 2.4, Casanellas et al. (2021) recently showed that noisy tree models can be recovered up to equivalence class using latent tree learning algorithms. In this subsection, we briefly discuss the sample complexity of latent tree learning algorithms and compare these results with the sample complexity for the SGA algorithm.

Popular latent tree recovery algorithms include the *recursive grouping* (RG) method by Choi et al. (2011), and the *neighbor joining* (NJ) method by Saitou & Nei (1987). The sample complexity for the vanilla RG method is known to be $O\big(\log(d^3/\tau)\big)$ (Choi et al., 2011, Thm. 11), with d and τ representing the number of nodes and the target error probability, respectively. But this result does not capture the sample complexity scaling as a function of the parameters of the latent tree $\rho_{\min}$ and $\rho_{\max}$. In contrast, our result in Theorem 3 highlights the explicit dependence of the required number of samples on $\rho_{\min}$, $\rho_{\max}$, and $q_{\max}$, and significantly improves the dependence on $\rho_{\min}$ from $\rho_{\min}^{-24}$ (Katiyar et al., 2020, Thm. 3) to $\rho_{\min}^{-8}$.

Anandkumar et al. (2011) proposed a *spectral RG* algorithm to learn latent trees, and showed that the sample complexity scales with $\rho_{\max}$ as $(1-\rho_{\max})^{-2}$ (Anandkumar et al., 2011, Thm. 1). This dependence on $\rho_{\max}$ is same as that of the SGA algorithm (see Prop. 1 and App. D.1). The explicit dependence of the spectral RG algorithm on $\rho_{\min}$ was not derived by Anandkumar et al. (2011).

Recently, Jaffe et al. (2021) proposed a *spectral NJ* algorithm for recovering latent trees, and derived the dependence of its sample complexity on $\rho_{\min}$ and $\rho_{\max}$ (Jaffe et al., 2021, Thm. 4.10). It is shown that number of samples scale with $\rho_{\max}$ as $(1-\rho_{\max})^{-2}$, while the scaling with respect to $\rho_{\min}$ is given by $\rho_{\min}^{-(2+4\log_2(d/2))}$, where d is the number of nodes. Note that the scaling with respect to $\rho_{\max}$ matches that of our proposed SGA algorithm, while the scaling with respect to $\rho_{\min}$ is strictly better for SGA when $d \geq 8$ (see Prop. 1 and App. D.1). Moreover, for spectral NJ, the sample complexity has a quadratic scaling $\tilde{O}(d^2)$ (Jaffe et al., 2021, Thm. 4.10). In contrast, for the SGA algorithm, the sample complexity only increases as $\log d$. This analytically superior sample complexity result using SGA (over the spectral NJ algorithm) is promising, especially because numerical results demonstrate that spectral NJ method outperforms the classical NJ and RG methods (Jaffe et al., 2021, Sec. 7).

6. Error Exponent Analyses

The *error exponent*, also called the *inaccuracy rate* (Kester & Kallenberg, 1986), captures the exponential decay of the error probability with n as a function of the distribution. For a given tree with d nodes $\mathrm{T} \in \mathcal{T}_d$ and graphical model $P \in \mathcal{P}_{\mathrm{T}}(\rho_{\min}, \rho_{\max})$, let $\tilde{P}$ denote the joint distribution of a noisy sample vector $\mathbf{Y}$, where the noise crossover probability at the ith node satisfies $0 \leq q_i \leq q_{\max} < 0.5$. Then, the error exponent of a given algorithm Ψ is[5]

$$E(\Psi, \tilde{P}) = E(\Psi, \tilde{P}, q_{\max}, \rho_{\min}, \rho_{\max}) \tag{9}$$

$$\triangleq \liminf_{n\to\infty} -\frac{1}{n} \log \mathbb{P}_{\tilde{P}}\big(\Psi(\mathbf{Y}_1^n) \notin [\mathrm{T}]\big). \tag{10}$$

We label the estimator by Katiyar et al. (2020) as Ψ_{KA} which uses Algorithm 2 in App. C for declaring a set of 4 nodes as star or non-star. We also label the estimator employing the SGA algorithm described in Algorithm 1 as Ψ_{SGA}. In the following, we use Sanov's theorem (Sanov, 1957) to quantify the error exponents of Ψ_{KA} and Ψ_{SGA}, and demonstrate that, in general, Ψ_{SGA} provides a better (i.e., higher) error exponent compared to Ψ_{KA}.

6.1. Error exponent using Ψ_{KA}

The performance of Ψ_{KA} depends on its ability to correctly declare a set of 4 nodes as star or non-star (with the appropriate pairing of nodes). The following proposition characterizes the error exponent for a 4-node tree.

Proposition 2. *Let P be a tree-structured graphical model for $\{X_1, X_2, X_3, X_4\}$, and let $\tilde{P}$ denote the joint distribution of the noisy samples. Let $\rho_{j,k}^{(Q)} \triangleq \mathbb{E}_Q[Y_j Y_k]$ and $\mathcal{Y} \triangleq \{+1, -1\}$.*

(a) If the tree distribution P corresponds to the Markov chain $X_1 - X_2 - X_3 - X_4$, and we define

$$e_1 \triangleq \min_{Q\in\mathcal{P}(\mathcal{Y}^4)} \Big\{ D(Q\|\tilde{P}) \,:\, \frac{\rho_{1,3}^{(Q)}\,\rho_{2,4}^{(Q)}}{\rho_{1,2}^{(Q)}\,\rho_{3,4}^{(Q)}} \geq \alpha \Big\}, \tag{11}$$

$$e_2 \triangleq \min_{Q\in\mathcal{P}(\mathcal{Y}^4)} \Big\{ D(Q\|\tilde{P}) \,:\, \frac{\rho_{1,3}^{(Q)}\,\rho_{2,4}^{(Q)}}{\rho_{1,4}^{(Q)}\,\rho_{2,3}^{(Q)}} \leq \alpha \Big\}, \tag{12}$$

then we have $E(\Psi_{\mathrm{KA}}, \tilde{P}) = \min\{e_1, e_2\}$.

(b) If P corresponds to a star tree structure, then $E(\Psi_{\mathrm{KA}}, \tilde{P})$ can be expressed as

$$\min_{Q\in\mathcal{P}(\mathcal{Y}^4)} \Big\{ D(Q\|\tilde{P}) : \frac{\rho_{1,3}^{(Q)}\,\rho_{2,4}^{(Q)}}{\rho_{1,2}^{(Q)}\,\rho_{3,4}^{(Q)}} \leq \alpha,\ \frac{\rho_{1,3}^{(Q)}\,\rho_{2,4}^{(Q)}}{\rho_{1,4}^{(Q)}\,\rho_{2,3}^{(Q)}} \geq \alpha \Big\}. \tag{13}$$

[5] For the estimators considered in this paper, it can be shown that the limit of the expression on the right side of (10) exists.

For a tree T with $d \geq 4$ nodes, if $\{X_{i_j}\}_{j=1}^4$ are 4 nodes in T that form a star structure (resp. non-star structure with pair $\{X_{i_1}, X_{i_2}\}$), and $\tilde{P}$ denotes the distribution of the noisy variables $\{Y_{i_j}\}_{j=1}^4$, then the exponent corresponding to an incorrect decision on the structure of these nodes by the procedure in Algorithm 2 is equal to expression on the right side of (13) (resp. equal to $\min\{e_1, e_2\}$ with e_1, e_2 defined in (11), (12)), where $\rho_{j,k}^{(Q)} = \mathbb{E}_Q[Y_{i_j} Y_{i_k}]$.

6.2. Error exponent using Ψ_{SGA}

The following proposition characterizes the error exponent using Ψ_{SGA} for a 4-node tree.

Proposition 3. *Let P be a tree-structured graphical model for $\{X_1, X_2, X_3, X_4\}$, and let $\tilde{P}$ denote the joint distribution of the noisy variables. Let $\rho_{j,k}^{(Q)} \triangleq \mathbb{E}_Q[Y_j Y_k]$ and $\mathcal{Y} \triangleq \{+1, -1\}$.*

(a) If the tree distribution P corresponds to the Markov chain $X_1 - X_2 - X_3 - X_4$, and we define

$$e_3 \triangleq \min_{Q \in \mathcal{P}(\mathcal{Y}^4)} \left\{ D(Q\|\tilde{P}) : \frac{\sqrt{|\rho_{1,3}^{(Q)} \rho_{2,4}^{(Q)} \rho_{1,4}^{(Q)} \rho_{2,3}^{(Q)}|}}{|\rho_{1,2}^{(Q)} \rho_{3,4}^{(Q)}|} \geq \alpha \right\}, \quad (14)$$

$$e_4 \triangleq \min_{Q \in \mathcal{P}(\mathcal{Y}^4)} \left\{ D(Q\|\tilde{P}) : |\rho_{1,3}^{(Q)} \rho_{2,4}^{(Q)}| \geq |\rho_{1,2}^{(Q)} \rho_{3,4}^{(Q)}| \right\}, \quad (15)$$

$$e_5 \triangleq \min_{Q \in \mathcal{P}(\mathcal{Y}^4)} \left\{ D(Q\|\tilde{P}) : |\rho_{1,4}^{(Q)} \rho_{2,3}^{(Q)}| \geq |\rho_{1,2}^{(Q)} \rho_{3,4}^{(Q)}| \right\}, \quad (16)$$

then we have $E(\Psi_{\mathrm{SGA}}, \tilde{P}) = \min\{e_3, e_4, e_5\}$.

(b) If the tree distribution P corresponds to a star tree structure, then $E(\Psi_{\mathrm{SGA}}, \tilde{P})$ can be expressed as

$$\min_{Q \in \mathcal{P}(\mathcal{Y}^4)} \left\{ D(Q\|\tilde{P}) : \frac{\sqrt{|\rho_{1,3}^{(Q)} \rho_{2,4}^{(Q)} \rho_{1,4}^{(Q)} \rho_{2,3}^{(Q)}|}}{|\rho_{1,2}^{(Q)} \rho_{3,4}^{(Q)}|} \leq \alpha \right\}. \quad (17)$$

For a tree T with $d \geq 4$ nodes, if $\{X_{i_j}\}_{j=1}^4$ are 4 nodes in T that form a star (resp. non-star with pair $\{X_{i_1}, X_{i_2}\}$), and $\tilde{P}$ denotes the distribution of the noisy variables $\{Y_{i_j}\}_{j=1}^4$, then the exponent corresponding to an incorrect decision on the structure of these nodes by the procedure in Algorithm 1 is equal to expression on the right side of (17) (resp. equal to $\min\{e_3, e_4, e_5\}$ with e_3, e_4, e_5 defined in (14), (15), and (16)), where $\rho_{j,k}^{(Q)} = \mathbb{E}_Q[Y_{i_j} Y_{i_k}]$.

6.3. Numerical comparison of the error exponents

Because the expressions for the error exponents in Props. 2 and 3 are not easily comparable (since $E(\Psi_{\mathrm{KA}}, \tilde{P})$ and $E(\Psi_{\mathrm{SGA}}, \tilde{P})$ are non-convex optimization problems), we present numerical comparisons of $E(\Psi_{\mathrm{KA}}, \tilde{P})$ and $E(\Psi_{\mathrm{SGA}}, \tilde{P})$ for 4-node homogeneous trees. Fig. 1 compares the error exponents for a 4-node Markov chain $X_1 - X_2 - X_3 - X_4$ where all the edge correlations are same, and are denoted as ρ. Fig. 1(a) considers a noiseless scenario where $q_{\max} = 0$, and it is observed that the error exponent for Ψ_{SGA} is *significantly* higher (hence better) than that for Ψ_{KA} for small values of ρ; e.g., when $\rho < 0.6$. On the other hand, Ψ_{KA} has only *marginally* higher exponent for higher values of ρ (when $\rho > 0.71$). Fig. 1(b) compares the error exponents for the scenario where $\rho = 0.74$ is fixed, and where the noise crossover probabilities for the nodes satisfy $q_1 = q_2 = q_3 = 0$ and $q_4 = q_{\max}$. It is seen that although Ψ_{KA} has better exponent than Ψ_{SGA} in the neighborhood of $q_{\max} = 0$, the performance of Ψ_{SGA} is slightly better for relatively higher values of $q_{\max}$.

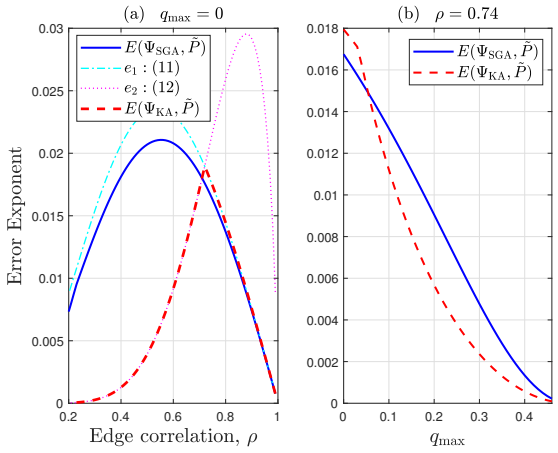

Figure 1. Error exponents for a 4-node homogeneous chain where all tree edges have correlation ρ. (a) Noiseless setting, $q_{\max} = 0$, (b) Error exponent versus $q_{\max}$ for fixed $\rho = 0.74$.

Fig. 2 compares the error exponents with Ψ_{KA} and Ψ_{SGA} for a 4-node star-structured tree where all edge correlations are same and are denoted ρ. Fig. 2(a) considers a noiseless scenario where $q_{\max} = 0$, and it is observed that the error exponent for Ψ_{SGA} is slightly higher than the error exponent for Ψ_{KA} for all values of ρ. Fig. 2(b) compares the error exponents for the scenario where $\rho = 0.4$ is fixed, and where the noise crossover probabilities for the nodes satisfy $q_1 = q_2 = q_3 = 0$ and $q_4 = q_{\max}$. Again, we see that the error exponent for Ψ_{SGA} is better than that for Ψ_{KA}.

Monte Carlo simulations for 4-node homogeneous trees (i.e., trees with equal correlations on the edges) that corroborate the theoretical results in this section, are presented in App. H. Even though Figs. 1 and 2 suggest that Ψ_{KA} sometimes outperforms Ψ_{SGA}, we show for larger trees that the performance of Ψ_{SGA} is almost always better than Ψ_{KA}. We explain why this is so in Sec. 7.1. Since Sanov's theorem is also applicable to random variables with arbitrary alphabets (Deuschel & Stroock, 2000, Ch. 3), we expect similar error exponent performances for Gaussian models.

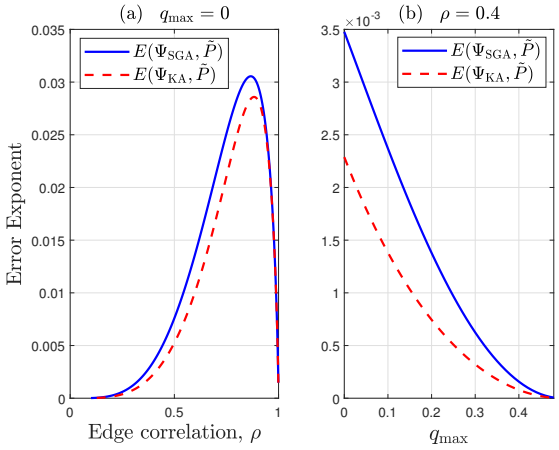

Figure 2. Error exponents for a 4-node star-structured tree where all edges have correlation ρ. (a) Noiseless setting, $q_{\max} = 0$, (b) Error exponent versus $q_{\max}$ for fixed $\rho = 0.4$.

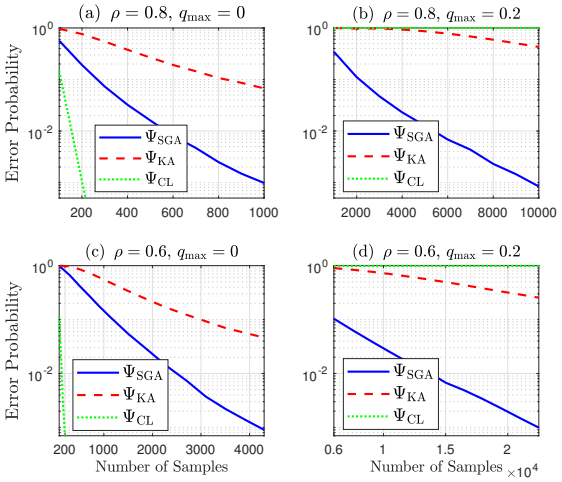

Figure 3. Comparison of error probabilities for a 12-node Markov chain where all edge correlations are equal to ρ.

7. Numerical Results

In this section, we present Monte Carlo simulation results for 12-node trees with three different tree structures: (i) Chain, (ii) Hybrid, (iii) Star (see Fig. 11 in App. I). The chain and the star structures are known to be extremal in terms of the error probability (Tan et al., 2010; Tandon et al., 2020), while the hybrid tree is a combination of the chain and star structures. For a given tree structure T, and n noisy samples $\mathbf{Y}_1^n$, the error probability $\mathbb{P}\big(\Psi(\mathbf{Y}_1^n) \notin [\mathrm{T}]\big)$ for a given learning algorithm Ψ, is estimated using 10^5 iterations in the Monte Carlo simulation framework, where an error is declared if the estimated tree does not belong to the equivalence class $[\mathrm{T}]$. We obtain error probability results for three different algorithms $\Psi_{\rm KA}$, $\Psi_{\rm SGA}$ and $\Psi_{\rm CL}$ – the vanilla Chow-Liu algorithm (Chow & Liu, 1968) applied directly to noisy samples. For $\Psi_{\rm KA}$ and $\Psi_{\rm SGA}$, the knowledge of $\rho_{\min}$ and $\rho_{\max}$ is assumed. The source code to reproduce the experiments can be found here: `https://github.com/AldricHan/SGA-Algorithm-Experiments`.

7.1. Error Probabilities for 12-node chain

Fig. 3 compares the error probabilities for a 12-node Markov chain, where all edge correlations are equal to ρ, using three learning algorithms: $\Psi_{\rm KA}$, $\Psi_{\rm SGA}$, and $\Psi_{\rm CL}$. Fig. 3(a) plots the results for the noiseless setting, $q_{\max} = 0$, with $\rho = 0.8$. In this case, it is seen that the Chow-Liu algorithm $\Psi_{\rm CL}$ provides the minimum error probability as it the maximum-likelihood algorithm for the noiseless setting (Chow & Wagner, 1973; Tan et al., 2011). The observation that $\Psi_{\rm SGA}$ has lower error probability than $\Psi_{\rm KA}$ for $\rho = 0.8$, $q_{\max} = 0$ can be intuitively explained as follows. An error event using $\Psi_{\rm KA}$ or $\Psi_{\rm SGA}$ occurs when a set of 4 nodes in the tree is incorrectly declared as star/non-star (see Sec. 4 and Sec. 5, respectively). Now, in the process of building a tree structure, these algorithms may pick a set of 4 non-neighboring nodes (that belong to each others' proximal sets) to characterize them as star or non-star. For instance, consider the set of 4 nodes X_1, X_3, X_5, X_7 that forms a sub-chain (of the 12-node chain) where the effective edge correlation for the sub-chain is $0.8^2 = 0.64$. From Fig. 1(a), we note that $\Psi_{\rm SGA}$ has a significantly higher exponent than $\Psi_{\rm KA}$ when the edge correlation is 0.64, and therefore we would expect $\Psi_{\rm SGA}$ to have a lower error probability compared to $\Psi_{\rm KA}$ when characterizing these 4 nodes as star or non-star.

Fig. 3(b) compares the error probabilities when $\rho = 0.8$ for the noisy case, where noise is added to alternate nodes, i.e., $q_i = q_{\max} = 0.2$ for $i \in O_{12} \triangleq \{1, 3, 5, 7, 9, 11\}$, while $q_j = 0$ for $j \in E_{12} \triangleq \{2, 4, 6, 8, 10, 12\}$. In contrast to the noiseless case, we observe from Fig. 3(b) that $\Psi_{\rm CL}$ performs extremely poorly with error probability roughly equal to 1. Such a poor performance is expected of $\Psi_{\rm CL}$ because $\mathbb{E}[Y_4 Y_6] = 0.64 > 0.48 = \mathbb{E}[Y_4 Y_5] = \mathbb{E}[Y_5 Y_6]$, and hence the tree estimated using $\Psi_{\rm CL}$ is more likely to pick the incorrect edge $\{X_4, X_6\}$ over the correct edges $\{X_4, X_5\}$ and $\{X_5, X_6\}$. Similar to Fig. 3(a), the plots in Fig. 3(b) highlight the clear superiority of $\Psi_{\rm SGA}$ over $\Psi_{\rm KA}$. A similar robust performance of $\Psi_{\rm SGA}$ is observed in the plots in Figs. 3(c) and 3(d) that compare the error probabilities for a 12-node Markov chain where $\rho = 0.6$.

7.2. Error Probabilities for 12-node hybrid tree

Fig. 4 compares the error probabilities for a 12-node hybrid tree where all correlations are equal to ρ. Fig. 4(a) plots the results for the noiseless setting with $\rho = 0.8$, while Fig. 4(b) considers the noisy case where noise is only added to even

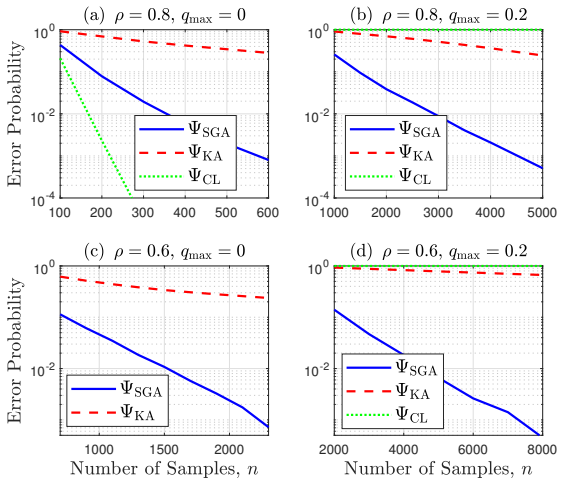

Figure 4. Comparison of error probabilities for a 12-node hybrid tree where all edge correlations are equal to ρ.

nodes, i.e. $q_i = 0$ for $i \in O_{12}$, and $q_j = q_{\max} = 0.2$ for $j \in E_{12}$. For the noiseless case, as expected, Ψ_{CL} provides the minimum error probability. However, for the noisy case, Ψ_{CL} performs poorly with error probability ≈ 1. Again, this is expected because $\mathbb{E}[Y_3Y_5] = 0.64 > 0.48 = \mathbb{E}[Y_3Y_4] = \mathbb{E}[Y_4Y_5]$, and hence the tree estimated using Ψ_{CL} is more likely to pick the incorrect edge $\{X_3, X_5\}$ over the correct edges $\{X_3, X_4\}$ and $\{X_4, X_5\}$ (see Fig. 11(ii) in App. I).

The plots in Fig. 4(c) and (d) compare the error probabilities for a hybrid tree when the edge correlation is $\rho = 0.6$. For the noiseless case in (c), the error probability using Ψ_{CL} is not plotted because it results in zero errors over 10^5 Monte Carlo simulation runs for the given values of n in Fig. 4(c). On the other hand, Fig. 4(d) considers the noisy case where noise is only added to even nodes, i.e. $q_i = 0$ for $i \in O_{12}$, and $q_j = q_{\max} = 0.2$ for $j \in E_{12}$. The error probability using Ψ_{CL} is quite high for the noisy case (note, for instance, that $\mathbb{E}[Y_3Y_5] = \mathbb{E}[Y_3Y_4] = \mathbb{E}[Y_4Y_5] = 0.36$). Fig. 4 clearly highlights the robustness of Ψ_{SGA} over Ψ_{KA} and Ψ_{CL} when the underlying tree is a hybrid of the chain and the star.

7.3. Error Probabilities for 12-node star tree

Fig. 5 compares the error probabilities using Ψ_{KA}, Ψ_{SGA}, and Ψ_{CL}, for a 12-node star where all correlations are equal to $\rho = 0.6$. Fig. 5(a) plots the results for the noiseless setting ($q_{\max} = 0$), while Fig. 5(b) considers the noisy case where noise is only added to odd nodes, i.e. $q_i = q_{\max} = 0.2$ for $i \in O_{12}$, while $q_j = 0$ for $j \in E_{12}$. The Chow-Liu algorithm Ψ_{CL} performs well in the noiseless setting, but fails miserably in the noisy setting. For both the noisy and noiseless settings, Ψ_{SGA} performs slightly better than Ψ_{KA}. This is justified using the error exponent result in Fig. 2(a) where it is observed that $E(\Psi_{\mathrm{SGA}}, \tilde{P})$ is slightly higher than

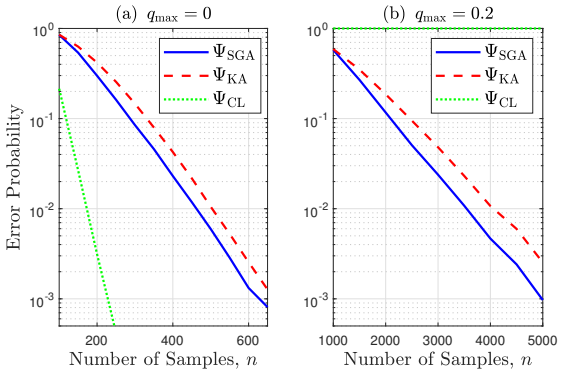

Figure 5. Comparison of error probabilities for a 12-node star, where all edge correlations are equal to $\rho = 0.6$.

that for $E(\Psi_{\mathrm{KA}}, \tilde{P})$ when $\rho = 0.6$.

We notice that for a moderate sample size ($n \approx 1000$), there is a *dichotomy* in the performance of Ψ_{CL}—its error probability is either close to 0 or 1. Hence, if Ψ_{CL} and Ψ_{SGA} output the *same* tree, this implies that the noise does not cause us to learn the wrong tree via Chow-Liu. Conversely, if Ψ_{CL} and Ψ_{SGA} output *different* trees, the one from Ψ_{SGA} should be "trusted" since it is designed to robustly learn trees with noise up to the equivalence class. Additional numerical results for Gaussian trees are presented in App. J.3.

8. Discussion and Future Work

There are several promising avenues for future research. First, SGA and the algorithm by Katiyar et al. (2020) depend on the knowledge of $\rho_{\max}$ through α. Designing algorithms that do not depend on $\rho_{\max}$ would be of practical interest. Second, we may tighten the sample complexity bounds so that the dependencies on $(\rho_{\min}, \rho_{\max}, q_{\max})$ are optimized. Finally, given noisy samples, we can endeavor to define equivalence classes (analogous to $[\mathrm{T}]$ here) and propose algorithms for the learning of various other graph structures such as random graphs (Anandkumar et al., 2012), latent trees (Choi et al., 2011), or forests (Tan et al., 2011).

Acknowledgements

We thank the reviewers for their useful feedback and for highlighting the connection of the present problem to learning latent tree models. The authors are very grateful to Fengzhuo Zhou (ECE, NUS) for his helpful suggestions and insights on latent tree models. This work is partially funded by a Singapore National Research Foundation (NRF) Fellowship (R-263-000-D02-281) and a Singapore Ministry of Education AcRF Tier 1 Grant (R-263-000-E80-114).

References

Anandkumar, A., Chaudhuri, K., Hsu, D., Kakade, S. M., Song, L., and Zhang, T. Spectral methods for learning multivariate latent tree structures. In *Proc. NeurIPS 2011*, pp. 2025–2033, Granada, Spain, 2011.

Anandkumar, A., Tan, V. Y. F., Huang, F., and Willsky, A. S. High-dimensional structure estimation of Ising models: Local separation criterion. *Ann. Statist.*, 40(3): 1346–1375, 2012.

Besag, J. On the statistical analysis of dirty pictures. *J. Roy. Statist. Soc., Ser. B*, 48(3):259–302, 1986.

Bresler, G. and Karzand, M. Learning a tree-structured Ising model in order to make predictions. *Ann. Statist.*, 48(2): 713–737, 2020.

Bresler, G., Mossel, E., and Sly, A. Reconstruction of Markov random fields from samples: Some observations and algorithms. *SIAM Journal on Computing*, 42(2): 563–578, 2013.

Casanellas, M., Garrote-Lopez, M., and Zwiernik, P. Robust estimation of tree structured models. arXiv:2102.05472v1 [stat.ML], Feb. 2021.

Cheng, Y., Diakonikolas, I., Kane, D. M., and Stewart, A. Robust learning of fixed-structure Bayesian networks. In *Proc. NeurIPS 2018*, pp. 10304–10316, Montreal, Canada, 2018.

Choi, M. J., Tan, V. Y. F., Anandkumar, A., and Willsky, A. S. Learning latent tree graphical models. *J. Mach. Learn. Res.*, 12:1771–1812, 2011.

Chow, C. K. and Liu, C. N. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inform. Theory*, 14(3):462–467, May 1968.

Chow, C. K. and Wagner, T. J. Consistency of an estimate of tree-dependent probability distributions. *IEEE Trans. Inform. Theory*, 19(3):369–371, May 1973.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. Wiley-Interscience, Hoboken, N.J., 2nd edition, 2006.

Deuschel, J.-D. and Stroock, D. W. *Large Deviations*. American Mathematical Society, 2000.

Friedman, N. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, 2004.

Herstein, I. *Topics In Algebra*. John Wiley and Sons, New York, 2nd edition, 1975.

Jaffe, A., Amsel, N., Aizenbud, Y., Nadler, B., Chang, J. T., and Kluger, Y. Spectral neighbor joining for reconstruction of latent tree models. *SIAM J. Math. Data Science*, 3 (1):113–141, 2021.

Johnson, J. K., Chandrasekaran, V., and Willsky, A. S. Learning Markov structure by maximum entropy relaxation. In *Proc. AISTATS 2007*, pp. 203–210, San Juan, Puerto Rico, Mar. 2007.

Katiyar, A., Hoffmann, J., and Caramanis, C. Robust estimation of tree structured Gaussian graphical models. In *Proc. ICML 2019*, pp. 3292–3300, Long Beach, California, USA, Jun. 2019.

Katiyar, A., Shah, V., and Caramanis, C. Robust estimation of tree structured Ising models. arXiv:2006.05601 [stat.ML], Jun. 2020.

Kester, A. D. M. and Kallenberg, W. C. M. Large deviations of estimators. *Ann. Statist.*, 14(2):648–664, 1986.

Krause, E. F. Maximizing the product of summands; Minimizing the sum of factors. *Mathematics Magazine*, 69(4): 270–278, Oct. 1996.

Kschischang, F. R. and Frey, B. J. Iterative decoding of compound codes by probability propagation in graphical models. *IEEE J. Sel. Areas Commun.*, 16(2):219–230, 1998.

Lauritzen, S. *Graphical Models*. Oxford Univ. Press, Oxford, U.K., 1996.

Nikolakakis, K. E., Kalogerias, D. S., and Sarwate, A. D. Learning tree structures from noisy data. In *Proc. AISTATS 2019*, pp. 1771–1782, Naha, Okinawa, Japan, 2019a.

Nikolakakis, K. E., Kalogerias, D. S., and Sarwate, A. D. Non-parametric structure learning on hidden tree-shaped distributions. arXiv:1909.09596v1 [stat.ML], Sep. 2019b.

Nikolakakis, K. E., Kalogerias, D. S., and Sarwate, A. D. Predictive learning on hidden tree-structured Ising models. arXiv:1812.04700v2 [stat.ML], Feb. 2019c.

Parikh, A. P., Song, L., and Xing, E. P. A spectral algorithm for latent tree graphical models. In *Proc. ICML 2011*, pp. 1065–1072, Jun. 2011.

Saitou, N. and Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

Sanov, I. N. On the probability of large deviations of random variables. *Mat. Sbornik*, 42(84)(1):11–44, 1957.

Scarlett, J. and Cevher, V. On the difficulty of selecting Ising models with approximate recovery. *IEEE Transactions on Signal and Information Processing over Networks*, 2 (4):625–638, 2016.

Tan, V. Y. F., Anandkumar, A., and Willsky, A. S. Learning Gaussian tree models: Analysis of error exponents and extremal structures. *IEEE Trans. Signal Process.*, 58(5): 2701–2714, May 2010.

Tan, V. Y. F., Anandkumar, A., Tong, L., and Willsky, A. S. A large-deviation analysis of the maximum-likelihood learning of Markov tree structures. *IEEE Trans. Inform. Theory*, 57(3):1714–1735, Mar. 2011.

Tan, V. Y. F., Anandkumar, A., and Willsky, A. S. Learning high-dimensional Markov forest distributions: Analysis of error rates. *J. Mach. Learn. Res.*, 12:1617–1653, 2011.

Tandon, A., Tan, V. Y. F., and Zhu, S. Exact asymptotics for learning tree-structured graphical models with side information: Noiseless and noisy samples. *IEEE Journal on Selected Areas in Information Theory*, 1(3):760–776, Nov 2020.

Tandon, R., Shanmugam, K., Ravikumar, P., and Dimakis, A. G. On the information theoretic limits of learning Ising models. In *Proc. NeurIPS 2014*, pp. 2303–2311, Montreal, Canada, 2014.

Wainwright, M. J. and Jordan, M. I. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, 2008.

Wang, L. and Gu, Q. Robust Gaussian graphical model estimation with arbitrary corruption. In *Proc. ICML 2017*, pp. 3617–3626, Sydney, Australia, Aug. 2017.