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

A. Proof of Theorem 1

We will prove the result by judiciously choosing a sufficiently large subset of tree-structured graphs whose corresponding distributions are close enough (with respect to the KL-divergence "metric"), and then applying Fano's lemma (Cover & Thomas, 2006).

We assume, for simplicity, that d is odd, with $d = 2t+1$ for some $t \in \mathbb{N}$. Let the edge set $\mathcal{E}_0$ consist of $2t$ edges $\big\{\{X_j, X_{2t+1}\}\big\}_{j=1}^{2t}$, and let the corresponding tree (resp. distribution) be denoted T_0 (resp. P_0). From (2) and (3), we know that a tree distribution is uniquely defined by the edge correlation values. Let the edge correlations, under distribution P_0, be given by

$$\rho_{j,2t+1} = \begin{cases} \rho_{\min}, \text{ if } 1 \le j \le t \\ \rho_{\max}, \text{ if } t+1 \le j \le 2t. \end{cases} \tag{18}$$

Let k be a positive integer satisfying $1 \le k \le t^2$, and define

$$k_{\mathrm{a}} \triangleq 1 + \left\lfloor \frac{k-1}{t} \right\rfloor, \quad k_{\mathrm{b}} \triangleq k - (k_{\mathrm{a}} - 1)t.$$

It is seen that $1 \le k_{\mathrm{a}}, k_{\mathrm{b}} \le t$, and the pair $(k_{\mathrm{a}}, k_{\mathrm{b}})$ is unique for every k. Let $\mathcal{E}_k$ denote the edge set of tree structure T_k, where

$$\mathcal{E}_k = \{k_{\mathrm{a}}, k_{\mathrm{b}} + t\} \cup \mathcal{E}_0 \setminus \{k_{\mathrm{a}}, 2t+1\}.$$

Hence, for $1 \le k \le t^2$, the edge set $\mathcal{E}_k$ differs from $\mathcal{E}_0$ in only one edge. Let P_k denote the tree distribution corresponding to T_k. For $\{k_{\mathrm{a}}, k_{\mathrm{b}} + t\} \in \mathcal{E}_k$, let $\rho_{k_{\mathrm{a}}, k_{\mathrm{b}}+t} = \rho_{\min}$, and let the edge correlation for the remaining edges in $\mathcal{E}_k$ be given by (18).

Now consider the noise model in Sec. 2.1, with q_i denoting the crossover probability for a sample corresponding to the ith node. For $d = 2t+1$ nodes, we fix these values as follows

$$q_i = \begin{cases} q_{\max}, & \text{if } 1 \le i \le t \\ 0, & \text{else.} \end{cases} \tag{19}$$

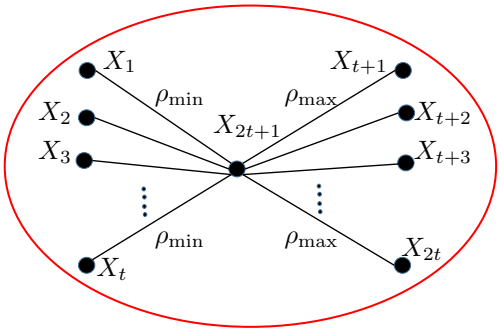
Let the above parameters be applicable to the noisy samples obtained from tree structure T_k, where $0 \le k \le t^2$. See Fig. 6 for diagrams of the trees constructed.

Let Y_i denote the noisy sample corresponding to the ith node. Then, we have $\tilde{\rho}_{i,j} \triangleq \mathbb{E}[Y_i Y_j] = (1-2q_i)(1-2q_j)\rho_{i,j}$ (Nikolakakis et al., 2019b). Let $\tilde{P}_k$ denote the distribution for the noisy vector $(Y_1, Y_2, \ldots, Y_{2t+1})$. As noise is only applied to leaf nodes (19), we have $Y_j = X_j$ for $t+1 \le j \le 2t+1$, and the conditional independence among Y_i, $1 \le i \le 2t+1$ continues to remain encoded via the tree structure T_k.[6] See Fig. 7 for the construction of the distribution of the noisy samples.

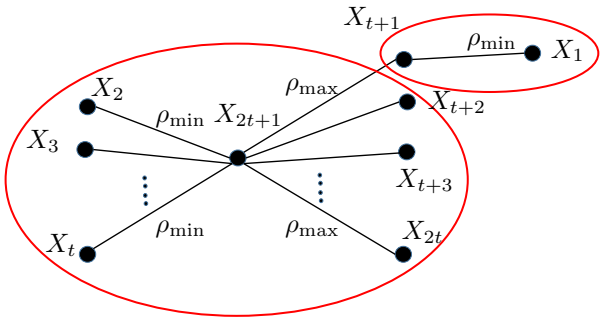
Let $M = t^2$, and let the tree structure T be chosen uniformly from the set $\{\mathrm{T}_0, \mathrm{T}_1, \ldots, \mathrm{T}_M\}$. Then we have the Markov chain $\mathrm{T} \longrightarrow \mathbf{X}_1^n \longrightarrow \mathbf{Y}_1^n \longrightarrow \hat{\mathrm{T}}$, and Fano's inequality gives a lower bound on the error probability $\mathbb{P}\big(\Psi(\mathbf{Y}_1^n) \neq \mathrm{T}\big)$ for any estimator Ψ using the multiple hypothesis testing framework. A key observation from our construction of the $M+1$ tree structures T_k, $0 \le k \le M$, is that their corresponding equivalence classes (see Sec. 2.2) are disjoint, i.e., $[\mathrm{T}_i] \cap [\mathrm{T}_j] = \varnothing$ for $i \neq j$. When learning the underlying tree structure using the multiple hypothesis framework, this observation has the important consequence that $\mathbb{P}\big(\Psi(\mathbf{Y}_1^n) \notin [\mathrm{T}]\big) = \mathbb{P}\big(\Psi(\mathbf{Y}_1^n) \neq \mathrm{T}\big)$, and we have the following result.

Lemma 1 (Fano's Inequality, Lemma 6.2 in (Bresler & Karzand, 2020))**.** *For $k \in \{0, 1, \ldots, M\}$, let $\tilde{P}_k$ be the probability law of the noisy observation $\mathbf{Y}$, with the models satisfying the properties given in Sec. 2.1. Let $\Psi : \{+1, -1\}^{d \times n} \to \mathcal{T}_d$ denote an estimator using n i.i.d. samples $\mathbf{Y}_1^n$. Let the KL-divergence be $D(\tilde{P}_k \| \tilde{P}_0) \triangleq$*

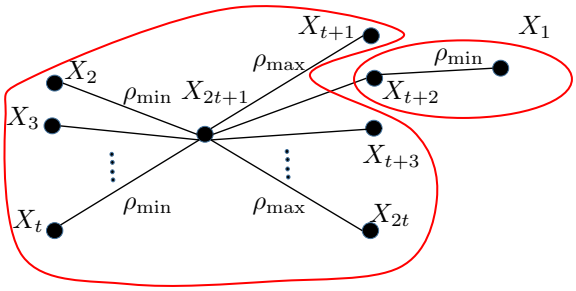
[6] In general, the noisy samples do not satisfy the Ising model (Bresler et al., 2013; Nikolakakis et al., 2019a). However, noisy samples from an underlying tree-structured Ising model retain the tree-structure when noise is only applied to leaf nodes.



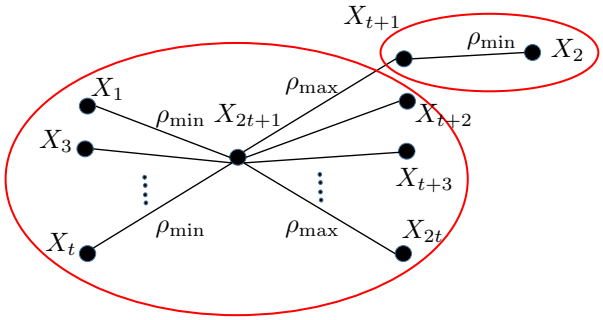
Tree Structure T_0



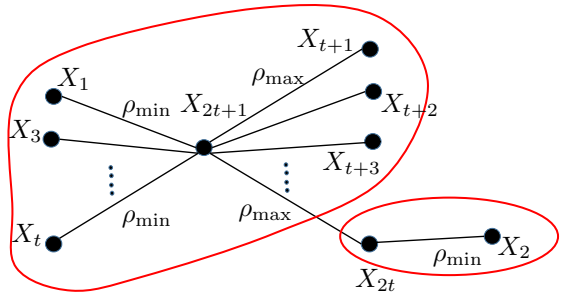
Tree Structure T_1 ($k = 1 \Rightarrow k_{\mathrm{a}} = k_{\mathrm{b}} = 1$)



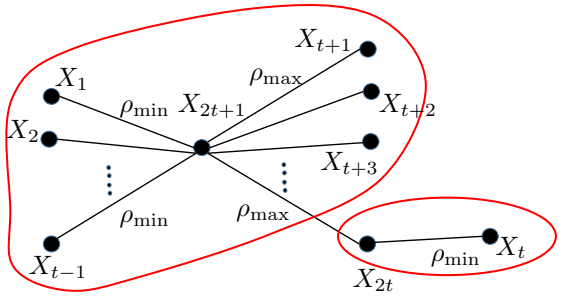
Tree Structure T_2 ($k = 2 \Rightarrow k_{\mathrm{a}} = 1, k_{\mathrm{b}} = 2$)



Tree Structure T_{t+1} ($k = t+1 \Rightarrow k_{\mathrm{a}} = 2, k_{\mathrm{b}} = 1$)



Tree Structure T_{2t} ($k = 2t \Rightarrow k_{\mathrm{a}} = 2, k_{\mathrm{b}} = t$)



Tree Structure T_{t^2} ($k = t^2 \Rightarrow k_{\mathrm{a}} = t, k_{\mathrm{b}} = t$)

Figure 6. Tree structures constructed in the proof of Theorem 1. Equivalence clusters are circled in red, where an equivalence cluster is defined as a set containing a non-leaf (internal) node and all the leaf nodes connected to it (Katiyar et al., 2020).

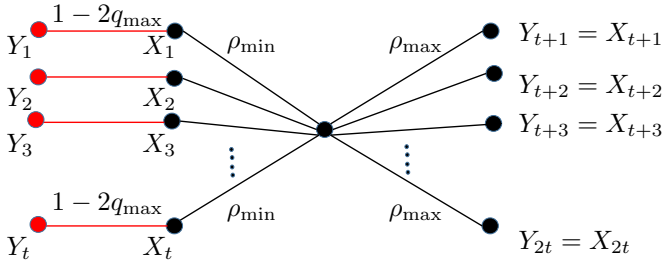


Figure 7. Construction of the distribution of the noisy samples $\{Y_i\}$.

$\sum_{\mathbf{y}\in\{+1,-1\}^d} \tilde{P}_k(\mathbf{y}) \log\left(\tilde{P}_k(\mathbf{y})/\tilde{P}_0(\mathbf{y})\right)$, *and define the symmetric KL-divergence* $J(\tilde{P}_k, \tilde{P}_0) \triangleq D(\tilde{P}_k\|\tilde{P}_0) + D(\tilde{P}_0\|\tilde{P}_k)$. *If the number of samples satisfy*

$$n \;<\; (1-\delta)\frac{\log M}{\frac{1}{M+1}\sum_{k=1}^{M} J(\tilde{P}_k, \tilde{P}_0)},$$

then the minimax error $\mathcal{M}_n(q_{\max}, \rho_{\min}, \rho_{\max})$ *in* (5) *is lower bounded as*

$$\mathcal{M}_n(q_{\max}, \rho_{\min}, \rho_{\max}) \geq \delta - \frac{1}{\log M}.$$

We remark that Lemma 1 continues to hold if the symmetric KL-divergence $J(\tilde{P}_k, \tilde{P}_0)$ is replaced by $D(\tilde{P}_k\|\tilde{P}_0)$; however, we use the given form as it is easier to quantify $J(\tilde{P}_k, \tilde{P}_0)$ for Ising models. We proceed to prove Theorem 1 by quantifying $J(\tilde{P}_k, \tilde{P}_0)$ and applying Lemma 1.

As discussed previously, if Y_j denotes a noisy sample corresponding to the jth node, then the conditional independence among Y_j, $1 \leq j \leq 2t+1$, continues to be encoded by the tree structure T_k for distribution $\tilde{P}_k$. Further, for $1 \leq k \leq M$, the edge set $\mathcal{E}_k$ differs from $\mathcal{E}_0$ in only one edge, and we have $\mathcal{E}_k \setminus \mathcal{E}_0 = \{k_{\mathrm{a}}, k_{\mathrm{b}}+t\}$ and $\mathcal{E}_0 \setminus \mathcal{E}_k = \{k_{\mathrm{a}}, 2t+1\}$. Let $\rho^{(k)}_{j_1,j_2}$ (resp. $\rho^{(0)}_{j_1,j_2}$) denote the correlation $\mathbb{E}[X_{j_1}X_{j_2}]$ with respect to the distribution P_k (resp. P_0). Then, from the construction of P_k and P_0 we have

$$\rho^{(k)}_{k_{\mathrm{a}},k_{\mathrm{b}}+t} = \rho_{\min}, \quad \rho^{(k)}_{k_{\mathrm{b}}+t,2t+1} = \rho_{\max}, \tag{20}$$

$$\rho^{(0)}_{k_{\mathrm{a}},2t+1} = \rho_{\min}, \quad \rho^{(0)}_{k_{\mathrm{b}}+t,2t+1} = \rho_{\max}. \tag{21}$$

Let $\tilde{\rho}^{(k)}_{j_1,j_2}$ (resp. $\tilde{\rho}^{(0)}_{j_1,j_2}$) denote the correlation $\mathbb{E}[Y_{j_1}Y_{j_2}]$ with respect to the distribution $\tilde{P}_k$ (resp. $\tilde{P}_0$). Then, using (19), (20), and (21), and applying the correlation decay property for tree-structured Ising models (Nikolakakis et al., 2019c, Lemma A.2), we obtain

$$\tilde{\rho}^{(k)}_{k_{\mathrm{a}},k_{\mathrm{b}}+t} = (1-q_{\max})\rho_{\min}, \qquad \tilde{\rho}^{(k)}_{k_{\mathrm{a}},2t+1} = (1-q_{\max})\rho_{\min}\rho_{\max}, \tag{22}$$

$$\tilde{\rho}^{(0)}_{k_{\mathrm{a}},k_{\mathrm{b}}+t} = (1-q_{\max})\rho_{\min}\rho_{\max}, \quad \tilde{\rho}^{(0)}_{k_{\mathrm{a}},2t+1} = (1-q_{\max})\rho_{\min}. \tag{23}$$

Finally, using (2), (3), (22), (23), and Eqn. (6.3) in Bresler & Karzand (2020), we obtain

$$J(\tilde{P}_k, \tilde{P}_0) \;=\; 2\,\mathrm{atanh}(\rho_q)\,\rho_q\,(1-\rho_{\max})\,, \tag{24}$$

where $\rho_q = (1-q_{\max})\rho_{\min}$. Now, using (24) and Lemma 1, we observe that if the number of samples satisfy

$$n \;<\; (1-\delta)\frac{\log M}{2\,\mathrm{atanh}(\rho_q)\,\rho_q\,(1-\rho_{\max})},$$

then we have $\mathcal{M}_n(q_{\max}, \rho_{\min}, \rho_{\max}) \geq \delta - (1/\log M)$. Setting $\delta = 1/2 + (1/\log M)$, and using the fact $M = t^2$, we get that $\mathcal{M}_n(q_{\max}, \rho_{\min}, \rho_{\max}) \geq 1/2$ if n satisfies

$$n \;<\; \frac{\log(t) - 1}{2\,\mathrm{atanh}(\rho_q)\,\rho_q\,(1-\rho_{\max})}. \tag{25}$$

The proof of Theorem 1 is complete by using (25) and observing that $\log(t) - 1 = \log(\frac{d-1}{2}) - 1 > \frac{1}{2}\log d$ for $d > 32$. □

B. Minimum and Maximum Possible Size of $[\mathrm{T}]$

The following proposition quantifies the minimum and maximum size of the equivalence class $[\mathrm{T}]$, for a given number of nodes d.

Proposition 4. *Let* $d \geq 4$. *Then, we have*

$$\min_{T\in\mathcal{T}_d} |[\mathrm{T}]| = 4, \tag{26}$$

$$\max_{T\in\mathcal{T}_d} |[\mathrm{T}]| \leq 3^{(d/3)}, \tag{27}$$

where the minimum is achieved in (26) *for a chain tree structure, and the inequality in* (27) *becomes tight when* d *is a multiple of* 3.

Algorithm 2 IS_NON_STAR

Let the set of 4 nodes be $\{X_1, X_2, X_3, X_4\}$
Input: Empirical correlations $\widehat{\rho}_{i,j},\ 1 \le i < j \le 4,$ Threshold $\alpha = \frac{1+\rho_{\max}^2}{2}$
if $\frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} < \alpha$ and $\frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}} > \alpha$ **then**
  Declare Non-star where $\{X_1, X_2\}$ forms a pair
else if $\frac{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}}{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}} < \alpha$ and $\frac{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}}{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}} > \alpha$ **then**
  Declare Non-star where $\{X_1, X_3\}$ forms a pair
else if $\frac{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}}{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}} < \alpha$ and $\frac{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}}{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}} > \alpha$ **then**
  Declare Non-star where $\{X_1, X_4\}$ forms a pair
else
  Declare Star
end if

Proof. For a given $d \ge 4$, and $\mathrm{T} \in \mathcal{T}_d$, recall that $\mathcal{L}_\mathrm{T}$ is the set of leaf nodes of T. Now, partition $\mathcal{L}_\mathrm{T}$ into smaller subsets, such that all elements in the same subset share a common neighbor in the tree structure T. Let κ denote the number of such distinct subsets, and let $\mathcal{L}_\mathrm{T}^{(i)}$ denote the ith subset. Then, $\mathcal{L}_\mathrm{T}$ can be expressed as a disjoint union of $\mathcal{L}_\mathrm{T}^{(i)}$ as

$$\mathcal{L}_\mathrm{T} = \uplus_{i=1}^{\kappa} \mathcal{L}_\mathrm{T}^{(i)}.$$

Now, if $\ell_i \triangleq |\mathcal{L}_\mathrm{T}^{(i)}|$, then it follows from (4) that the size of the equivalence class is given by

$$|\,[\mathrm{T}]\,| = \prod_{i=1}^{\kappa}(1+\ell_i). \tag{28}$$

When $\kappa = 1$, then the tree structure is a *star* (Tan et al., 2011), and we have $|\,[\mathrm{T}]\,| = d \ge 4$. As $\ell_i \ge 1$, it follows from (28) that for $\kappa \ge 2$, we have $|\,[\mathrm{T}]\,| \ge 4$. In particular, for a chain tree structure, we have $\kappa = 2$ with $\ell_1 = \ell_2 = 1$, and hence $|\,[\mathrm{T}]\,| = 4$ for a chain.

We proceed to prove the upper bound in (27). From (28), it follows that the size $|\,[\mathrm{T}]\,|$ can be upper bounded by the solution of the following constrained maximization problem,

$$|\,[\mathrm{T}]\,| \le \max_{m_1+\cdots+m_\kappa \le d} \prod_{i=1}^{\kappa} m_i.$$

The expression on the right side is upper bounded by $3^{d/3}$, and this bound is tight when $d \mod 3 = 0$ (Krause, 1996). Thus, we have

$$|\,[\mathrm{T}]\,| \le 3^{d/3},$$

with equality if d is a multiple of 3. Note that when $d \mod 3 = 0$, and $\kappa = d/3$ with $\ell_1 = \cdots = \ell_\kappa = 2$, then we have $|\,[\mathrm{T}]\,| = 3^{d/3}$. □

C. Overview of the Algorithm by Katiyar et al. (2020) for Declaring Star/Non-star

Let $\mathbf{y}_1^n = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ denote n independently sampled noisy observations, where the kth noisy sample is a d-dimensional column vector $\mathbf{y}_k = (y_{k,1}, \ldots, y_{k,d})^T$. The estimator in Katiyar et al. (2020) proceeds by first calculating the pairwise empirical correlations,

$$\widehat{\rho}_{i,j} \triangleq \frac{1}{n}\sum_{k=1}^{n} y_{k,i}\, y_{k,j}, \tag{29}$$

where $1 \le i < j \le d$. The procedure used in Katiyar et al. (2020) to declare a set of 4 nodes as star or non-star, based on the knowledge of empirical correlations $\widehat{\rho}_{i,j}$, is described in Algorithm 2. The intuition behind Algorithm 2 can be roughly outlined by considering an example where the 4 nodes form a Markov-chain $X_1 - X_2 - X_3 - X_4$. If the noisy correlations

are denoted $\tilde{\rho}_{i,j} \triangleq \mathbb{E}[Y_i Y_j]$,[7] then we have $\frac{\tilde{\rho}_{1,3}\,\tilde{\rho}_{2,4}}{\tilde{\rho}_{1,2}\,\tilde{\rho}_{3,4}} \leq \rho_{\max}^2$ and $\frac{\tilde{\rho}_{1,3}\,\tilde{\rho}_{2,4}}{\tilde{\rho}_{1,4}\,\tilde{\rho}_{2,3}} = 1$, and hence we would expect the empirical correlations to satisfy the conditions $\frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} < \alpha$ and $\frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}} > \alpha$, where $\alpha = (1+\rho_{\max}^2)/2$.

The partial tree structure learning algorithm detailed in Katiyar et al. (2020) ensures that if Algorithm 2 correctly declares any set of 4 nodes as star or non-star (with appropriate pairing of nodes), then the equivalence class $[\mathrm{T}]$ is successfully detected, i.e. $\Psi(\mathbf{y}_1^n) \in [\mathrm{T}]$. Therefore, the performance of the estimator critically depends on the accuracy of Algorithm 2.

D. Proof of Theorem 3

The algorithm by Katiyar et al. (2020) correctly estimates the equivalence class $[\mathrm{T}]$ if any set of 4 nodes within each others *proximal sets* are declared correctly as *star* or *non-star*. The algorithm used for declaring 4 nodes as star or non-star is described in Alg. 2 in App. C. 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 $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,d})^T$. The algorithm proceeds by first calculating the pairwise empirical correlations, $\widehat{\rho}_{j,k} \triangleq \frac{1}{n}\sum_{i=1}^n y_{i,j}\, y_{i,k}$, where $1 \leq j < k \leq d$.

Without loss of generality, consider the set of 4 nodes $\{X_1, X_2, X_3, X_4\}$ with the corresponding noisy variables $\{Y_1, Y_2, Y_3, Y_4\}$, and let $\tilde{\rho}_{j,k} = \mathbb{E}[Y_j\, Y_k]$. Let these nodes form a non-star with $\{X_1, X_2\}$ as a pair. From the procedure in Alg. 2 in App. C, it follows that a correct decision is made if

$$\frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} \;<\; \alpha \quad \text{and} \quad \frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}} \;>\; \alpha, \tag{30}$$

where $\alpha = (1+\rho_{\max}^2)/2$. Now, as $\{X_1, X_2\}$ forms a pair, we have

$$\frac{\tilde{\rho}_{1,3}\,\tilde{\rho}_{2,4}}{\tilde{\rho}_{1,2}\,\tilde{\rho}_{3,4}} \;\leq\; \rho_{\max}^2 \quad \text{and} \quad \frac{\tilde{\rho}_{1,3}\,\tilde{\rho}_{2,4}}{\tilde{\rho}_{1,4}\,\tilde{\rho}_{2,3}} \;=\; 1. \tag{31}$$

Define $\Delta_{j,k} \triangleq \tilde{\rho}_{j,k} - \widehat{\rho}_{j,k}$, and $\Delta = \max_{1\leq j<k\leq 4} |\Delta_{j,k}|$. We will show that the inequalities in (30) are satisfied if

$$\Delta < \tilde{\delta} \triangleq \frac{t_2(1-\alpha)}{20}, \tag{32}$$

where $t_2 = \min\left\{t_1, \frac{t_1(1-2q_{\max})}{\rho_{\max}}\right\}$ and $t_1 = (1-2q_{\max})^2\rho_{\min}^4$. Assume the inequality in (32) to be true and define $\beta \triangleq 0.1(1-\alpha)$. Then, for $1 \leq j < k \leq 4$, we have

$$\left|\frac{\Delta_{j,k}}{\widehat{\rho}_{i,j}}\right| \overset{(a)}{\leq} \frac{\Delta}{0.5t_2} \overset{(b)}{<} \frac{\tilde{\delta}}{0.5t_2} \overset{(c)}{=} \beta, \tag{33}$$

where (a) follows from the fact that proximal sets are chosen to satisfy $|\widehat{\rho}_{i,j}| \geq 0.5t_2$, (b) follows from (32), and (c) follows from the definitions of $\tilde{\delta}$ and β. Now, we have

$$\begin{aligned}
\rho_{\max}^2 &\overset{(d)}{\geq} \frac{\tilde{\rho}_{1,3}\,\tilde{\rho}_{2,4}}{\tilde{\rho}_{1,2}\,\tilde{\rho}_{3,4}} \\
&= \frac{\big(\widehat{\rho}_{1,3}+\Delta_{1,3}\big)\big(\widehat{\rho}_{2,4}+\Delta_{2,4}\big)}{\big(\widehat{\rho}_{1,2}+\Delta_{1,2}\big)\big(\widehat{\rho}_{3,4}+\Delta_{3,4}\big)} \\
&= \frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}}\,\frac{\big(1+\Delta_{1,3}/\widehat{\rho}_{1,3}\big)\big(1+\Delta_{2,4}/\widehat{\rho}_{2,4}\big)}{\big(1+\Delta_{1,2}/\widehat{\rho}_{1,2}\big)\big(1+\Delta_{3,4}/\widehat{\rho}_{3,4}\big)} \\
&\overset{(e)}{>} \frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}}\,\frac{(1-\beta)^2}{(1+\beta)^2},
\end{aligned} \tag{34}$$

[7] Note that $\tilde{\rho}_{i,j} = (1-2q_i)(1-2q_j)\,\rho_{i,j}$.

where (d) follows from (31), and (e) follows from (33). We can equivalently express (34) as

$$\begin{aligned}\frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} &< \rho_{\max}^2\frac{(1+\beta)^2}{(1-\beta)^2}\\ &< \rho_{\max}^2\frac{1+2.1\beta}{1-2\beta}\\ &< \rho_{\max}^2(1+2.1\beta)(1+3\beta)\\ &< \rho_{\max}^2(1+6\beta),\end{aligned} \tag{35}$$

where we have applied the fact that $\beta < 0.1$, and hence $\beta^2 < 0.1\beta$. As $\beta = 0.1(1-\alpha) = (1-\rho_{\max}^2)/20$, it follows from (35) that

$$\begin{aligned}\frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} &< \rho_{\max}^2 + 0.3\rho_{\max}^2(1-\rho_{\max}^2)\\ &< \rho_{\max}^2 + 0.3(1-\rho_{\max}^2)\\ &= 0.3 + 0.7\rho_{\max}^2\\ &< 0.5 + 0.5\rho_{\max}^2\\ &= \alpha,\end{aligned}$$

and this proves the first inequality in (30). To prove the second inequality in (30), we note that

$$\begin{aligned}1 &= \frac{\tilde{\rho}_{1,3}\,\tilde{\rho}_{2,4}}{\tilde{\rho}_{1,4}\,\tilde{\rho}_{2,3}}\\ &= \frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}}\,\frac{\big(1+\Delta_{1,3}/\widehat{\rho}_{1,3}\big)\big(1+\Delta_{2,4}/\widehat{\rho}_{2,4}\big)}{\big(1+\Delta_{1,4}/\widehat{\rho}_{1,4}\big)\big(1+\Delta_{2,3}/\widehat{\rho}_{2,3}\big)}\\ &< \frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}}\,\frac{(1+\beta)^2}{(1-\beta)^2},\end{aligned} \tag{36}$$

where the last inequality follows from (33). We can equivalently express (36) as

$$\begin{aligned}\frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}} &> \frac{(1-\beta)^2}{(1+\beta)^2}\\ &> \frac{1-2\beta}{1+2.1\beta}\\ &> (1-2\beta)(1-2.1\beta)\\ &> 1-5\beta\\ &= 1-0.25(1-\rho_{\max}^2)\\ &= 0.75+0.25\rho_{\max}^2\\ &> 0.5+0.5\rho_{\max}^2\\ &= \alpha,\end{aligned}$$

thereby proving the second inequality in (30). Thus, we have shown that if $\{X_1, X_2, X_3, X_4\}$ form a non-star with pair $\{X_1, X_2\}$, then the condition in (32) is sufficient for the algorithm to make a correct decision. In a similar fashion, it can be shown that (32) provides a sufficient condition for making the correct decision even when $\{X_1, X_2, X_3, X_4\}$ form a star or a non-star with a different pairing.

Define the event $\mathcal{B}_{j,k}$ for $1 \le j < k \le d$ as

$$\mathcal{B}_{j,k} \triangleq \Big\{|\widehat{\rho}_{j,k} - \tilde{\rho}_{j,k}| \ge \tilde{\delta}\Big\}. \tag{37}$$

Then, as (32) is a sufficient condition for correct declaration as star/non-star for any set 4 nodes that are within the proximal sets of each other, it follows that for any $P \in \mathcal{P}_{\mathrm{T}}(\rho_{\min}, \rho_{\max})$, the error probability $\mathbb{P}_P\big(\Psi(\mathbf{Y}_1^n) \notin [\mathrm{T}]\big)$ can be upper bounded as

$$\mathbb{P}_P\big(\Psi(\mathbf{Y}_1^n) \notin [\mathrm{T}]\big) \;\le\; \Pr\big(\bigcup_{1\le j<k\le d}\mathcal{B}_{j,k}\big). \tag{38}$$

From the definition of the event $\mathcal{B}_{j,k}$ in (37), it follows using Hoeffding's inequality that

$$\Pr\left(\mathcal{B}_{j,k}\right) \leq 2\exp\left(-\frac{n\,\tilde{\delta}^2}{2}\right). \tag{39}$$

Now, using (38), (39), and applying the union bound over $\binom{d}{2}$ pairs of nodes, we obtain

$$\mathbb{P}_P\big(\Psi(\mathbf{Y}_1^n) \notin [\mathrm{T}]\big) \leq d^2\exp\left(-\frac{n\,\tilde{\delta}^2}{2}\right). \tag{40}$$

Therefore, for the error probability to be upper bounded by τ, it is sufficient for the number of samples n to satisfy

$$n \geq \frac{2}{\tilde{\delta}^2}\log\left(\frac{d^2}{\tau}\right). \tag{41}$$

This completes the proof. □

D.1. Discussion

We note that this is significantly improved over Katiyar et al. (2020, Theorem 3) as the right-hand-side in (41) is $O(1/\tilde{\delta}^2)$ instead of $O(1/\delta^2)$ (see Theorem 2). Recall that $\tilde{\delta} = \Theta(t_2)$ while $\delta = \Theta(t_2^3)$ and $t_2 = \min\left\{t_1, \frac{t_1(1-2q_{\max})}{\rho_{\max}}\right\}$ and $t_1 = (1-2q_{\max})^2\rho_{\min}^4$.

The result in Theorem 3 was derived for the scenario where the noise statistics are unknown. However, if the noise distribution at each node is *known*, the Chow-Liu algorithm (Chow & Liu, 1968) can be applied via appropriate pre-processing where the input to Chow-Liu is a complete weighted graph with *scaled* weights $\widehat{\rho}_{i,j}/\big((1-2q_i)(1-2q_j)\big)$ applied to node pair $\{i,j\}$. In this case, it can be shown (by a careful application of the Hoeffding's inequality) that for a given target error probability $\tau \in (0,1)$, the optimal sample complexity satisfies

$$n^*(\rho_{\min},\rho_{\max},q_{\max},d) \leq n^*_{\mathrm{CL}}(\rho_{\min},\rho_{\max},q_{\max},d) = O\left(\frac{\log(d/\tau)}{(1-\rho_{\max})^2(1-2q_{\max})^4\rho_{\min}^2}\right). \tag{42}$$

Contrast this to the algorithm by Katiyar et al. (2020) (Theorem 3) and SGA (Proposition 1) which *do not assume knowledge of the q_i's*. By noting that $\tilde{\delta} \propto (1-\rho_{\max})(1-2q_{\max})^3\rho_{\min}^4$ (assuming $\rho_{\max} \to 1$), Theorem 3 and Proposition 1 respectively say that

$$n^*(\rho_{\min},\rho_{\max},q_{\max},d) \leq n^*_{\mathrm{KA}}(\rho_{\min},\rho_{\max},q_{\max},d) = O\left(\frac{\log(d/\tau)}{(1-\rho_{\max})^2(1-2q_{\max})^6\rho_{\min}^8}\right), \quad \text{and} \tag{43}$$

$$n^*(\rho_{\min},\rho_{\max},q_{\max},d) \leq n^*_{\mathrm{SGA}}(\rho_{\min},\rho_{\max},q_{\max},d) = O\left(\frac{\log(d/\tau)}{(1-\rho_{\max})^2(1-2q_{\max})^6\rho_{\min}^8}\right). \tag{44}$$

We note that (43) pertains to the *improved* analysis of the algorithm by Katiyar et al. (2020) and not Theorem 2. The impossibility result in Theorem 1 (derived assuming the knowledge of the noise statistics) says that the optimal sample complexity satisfies

$$n^*(\rho_{\min},\rho_{\max},q_{\max},d) = \Omega\left(\frac{\log(d/\tau)}{(1-\rho_{\max})(1-2q_{\max})^2\rho_{\min}^2}\right). \tag{45}$$

We make the following observations based on the above sample complexity bounds.

- The bound in (42) coincides with that by Nikolakakis et al. (2019a, Thm. 1). However, the latter is designed for the more restrictive case in which the noise parameters $\{q_i\}_{i=1}^d$ are the same across all nodes. Note that in this case, one can simply run the vanilla Chow-Liu algorithm to learn the tree since the *order* of the correlations or mutual information quantities remains unchanged.
- Assuming the noise statistics are known, by comparing (42) and (45), we see that the dependence on $\rho_{\min}$ is order-optimal, i.e., $\Theta(\rho_{\min}^{-2})$ with all other parameters fixed.
- Comparing (42), (43) and (44), we see that the dependence on $1-\rho_{\max}$ for scaled Chow-Liu, Katiyar et al. (2020) and SGA are the same. Thus, in terms of the parameter $1-\rho_{\max}$, there is no cost our incognizance of the noise statistics q_i.

- There is a gap between the sample complexity bounds in Theorems 1 and 3 and Proposition 1 in terms of $1-\rho_{\max}$, $\rho_{\min}$, and $1-2q_{\max}$ (as can be seen from (43)–(45)) because the noise statistics are assumed to be *unknown* for the algorithm by Katiyar et al. (2020) (as well as SGA). As mentioned in Sec. 8, closing the gaps on the dependencies on the parameters is a promising avenue of future work.

E. Proof of Proposition 1

The proof is similar to the proof of Theorem 3 in App. D, and we focus here only on the important steps. Consider the set of 4 nodes $\{X_1, X_2, X_3, X_4\}$ that forms a non-star with $\{X_1, X_2\}$ as a pair. From the procedure in Alg. 1, it follows that a correct decision is made if

$$\frac{\sqrt{|\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}\,\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}|}}{|\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}|} < \alpha, \quad \frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} < 1, \text{ and } \frac{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} < 1, \tag{46}$$

where $\alpha = (1+\rho_{\max}^2)/2$. Now, as $\{X_1, X_2\}$ forms a pair, we have

$$\frac{\sqrt{\tilde{\rho}_{1,3}\,\tilde{\rho}_{2,4}\,\tilde{\rho}_{1,4}\,\tilde{\rho}_{2,3}}}{\tilde{\rho}_{1,2}\,\tilde{\rho}_{3,4}} \le \rho_{\max}^2, \quad \frac{\tilde{\rho}_{1,3}\,\tilde{\rho}_{2,4}}{\tilde{\rho}_{1,2}\,\tilde{\rho}_{3,4}} \le \rho_{\max}^2, \text{ and } \frac{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} \le \rho_{\max}^2. \tag{47}$$

Define $\Delta_{j,k} \triangleq \tilde{\rho}_{j,k} - \widehat{\rho}_{j,k}$, and $\Delta = \max_{1\le j<k\le 4} |\Delta_{j,k}|$. We will show that the inequalities in (46) are satisfied if

$$\Delta < \tilde{\delta} \triangleq \frac{t_2(1-\alpha)}{20}, \tag{48}$$

where $t_2 = \min\left\{t_1, \frac{t_1(1-2q_{\max})}{\rho_{\max}}\right\}$ and $t_1 = (1-2q_{\max})^2\rho_{\min}^4$. Assume the inequality in (48) to be true and define $\beta \triangleq 0.1(1-\alpha)$. Then, for $1 \le j < k \le 4$, we have

$$\left|\frac{\Delta_{j,k}}{\widehat{\rho}_{i,j}}\right| \overset{(a)}{\le} \frac{\Delta}{0.5t_2} \overset{(b)}{<} \frac{\tilde{\delta}}{0.5t_2} \overset{(c)}{=} \beta, \tag{49}$$

where (a) follows from the fact that proximal sets are chosen to satisfy $|\widehat{\rho}_{i,j}| \ge 0.5t_2$, (b) follows from (48), and (c) follows from the definitions of $\tilde{\delta}$ and β. Now, we have

$$\begin{aligned}
\rho_{\max}^2 &\overset{(d)}{\ge} \frac{\sqrt{\tilde{\rho}_{1,3}\,\tilde{\rho}_{2,4}\,\tilde{\rho}_{1,4}\,\tilde{\rho}_{2,3}}}{\tilde{\rho}_{1,2}\,\tilde{\rho}_{3,4}} \\
&= \frac{\sqrt{\big(\widehat{\rho}_{1,3}+\Delta_{1,3}\big)\big(\widehat{\rho}_{2,4}+\Delta_{2,4}\big)\big(\widehat{\rho}_{1,4}+\Delta_{1,4}\big)\big(\widehat{\rho}_{2,3}+\Delta_{2,3}\big)}}{\big(\widehat{\rho}_{1,2}+\Delta_{1,2}\big)\big(\widehat{\rho}_{3,4}+\Delta_{3,4}\big)} \\
&= \frac{\sqrt{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}\,\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} \frac{\sqrt{\big(1+\Delta_{1,3}/\widehat{\rho}_{1,3}\big)\big(1+\Delta_{2,4}/\widehat{\rho}_{2,4}\big)\big(1+\Delta_{1,4}/\widehat{\rho}_{1,4}\big)\big(1+\Delta_{2,3}/\widehat{\rho}_{2,3}\big)}}{\big(1+\Delta_{1,2}/\widehat{\rho}_{1,2}\big)\big(1+\Delta_{3,4}/\widehat{\rho}_{3,4}\big)} \\
&\overset{(e)}{>} \frac{\sqrt{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}\,\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} \frac{(1-\beta)^2}{(1+\beta)^2},
\end{aligned} \tag{50}$$

where (d) follows from (47), and (e) follows from (49). Now, we can equivalently express (50) as

$$\begin{aligned}
\frac{\sqrt{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}\,\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} &< \rho_{\max}^2 \frac{(1+\beta)^2}{(1-\beta)^2} \\
&\overset{(f)}{<} \alpha,
\end{aligned} \tag{51}$$

where (f) follows by employing relations similar to those used around (35), and thereby establishes the first inequality in (46). The other two inequalities in (46) can be readily proved in a similar way. Further, this approach can be repeated to prove that the condition in (48) is sufficient for making a correct decision even when the nodes $\{X_1, X_2, X_3, X_4\}$ form a star or a non-star with a different pairing. Finally, the steps in (37)–(41) can be repeated to complete the proof. □

F. Proof of Proposition 2

- We first prove the claim given by Proposition 2(a) where P corresponds to a Markov chain. We observe from the chain structure that the 4 nodes form a non-star with $\{X_1, X_2\}$ forming a pair. It follows from the procedure in Algorithm 2 (in App. C) that the two events that lead to error using Ψ_{KA} are $\mathscr{E}_1 = \left\{\frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} \geq \alpha\right\}$, and $\mathscr{E}_2 = \left\{\frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}} \leq \alpha\right\}$. The exponents corresponding to these error events are defined as $e_i \triangleq \lim_{n\to\infty} -\frac{1}{n}\log \Pr\left(\mathscr{E}_i\right),\ i \in \{1,2\}$. Using Sanov's theorem (Cover & Thomas, 2006, Chap. 11), it follows that these exponents are given by (11) and (12), respectively. Now, we have $\mathbb{P}_{\tilde{P}}\big(\Psi(\mathbf{Y}_1^n) \notin [\mathrm{T}]\big) = \mathbb{P}_{\tilde{P}}\big(\mathscr{E}_1 \cup \mathscr{E}_2\big)$, and hence it follows from the definition in (10) that $E(\Psi_{\mathrm{KA}}, \tilde{P}) = \min\{e_1, e_2\}$.
- We now prove the claim given by Proposition 2(b) where P corresponds to a star structured tree. When the 4 nodes form a star structure, then an error is made if only if the procedure in Algorithm 2 (in App. C) make any one of the following incorrect declarations:
  - Non-star with pair $\{X_1, X_2\}$.
  - Non-star with pair $\{X_1, X_3\}$.
  - Non-star with pair $\{X_1, X_4\}$.

  By symmetry of the underlying star structure, the probability of each of the three erroneous declarations is same, and hence it is sufficient to analyze the exponent of the probability that a non-star with pair $\{X_1, X_2\}$ is incorrectly declared, in order to characterize $E(\Psi_{\mathrm{KA}}, \tilde{P})$. Now, it follows from Alg. 2 that a non-star with pair $\{X_1, X_2\}$ is declared if $\frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} < \alpha$, and $\frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}} > \alpha$. Finally, using Sanov's theorem (Cover & Thomas, 2006, Ch. 11), it follows that the exponent of the probability that these conditions are satisfied are given by (13).

□

G. Proof of Proposition 3

- We first prove the claim given by Proposition 3(a) where P corresponds to a Markov chain. We observe from the chain structure that the 4 nodes form a non-star with $\{X_1, X_2\}$ forming a pair. It follows from the procedure in Algorithm 1 that a correct decision is made if (i) $v_2 < \alpha$, (ii) $v_2 < v_3$, and (iii) $v_2 < v_4$. Thus, an incorrect decision is made if any of the following events are true.
  - Event $\mathscr{E}_3 \triangleq \left\{\frac{\sqrt{|\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}\,\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}|}}{|\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}|} \geq \alpha\right\}$, implying $v_2 \geq \alpha$.
  - Event $\mathscr{E}_4 \triangleq \{|\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}| \geq |\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}|\}$, implying $v_2 \geq v_3$.[8]
  - Event $\mathscr{E}_5 \triangleq \{|\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}| \geq |\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}|\}$, implying $v_2 \geq v_4$.

  The exponents corresponding to these error events are defined as $e_i \triangleq \lim_{n\to\infty} -\frac{1}{n}\log \Pr\left(\mathscr{E}_i\right),\ i = \{3,4,5\}$. Using Sanov's theorem (Cover & Thomas, 2006, Chap. 11), it follows that these exponents are given by (14), (15) and (16), respectively. Now, we have $\mathbb{P}_{\tilde{P}}\big(\Psi(\mathbf{Y}_1^n) \notin [\mathrm{T}]\big) = \mathbb{P}_{\tilde{P}}\big(\mathscr{E}_3 \cup \mathscr{E}_4 \cup \mathscr{E}_5\big)$, and hence it follows from the definition in (10) that $E(\Psi_{\mathrm{SGA}}, \tilde{P}) = \min\{e_3, e_4, e_5\}$.
- We now prove the claim given by Proposition 3(b) where P corresponds to a star structured tree. When the 4 nodes form a star structure, then an error is made if only if the procedure in Algorithm 1 make any one of the following incorrect declarations:
  - Non-star with pair $\{X_1, X_2\}$.
  - Non-star with pair $\{X_1, X_3\}$.
  - Non-star with pair $\{X_1, X_4\}$.

  By symmetry of the underlying star structure, the probability of each of the three erroneous declarations is same, and hence it follows from Algorithm 1 that $E(\Psi_{\mathrm{SGA}}, \tilde{P})$ is equal to the exponent of the probability that $v_2 = \frac{\sqrt{|\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}\,\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}|}}{|\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}|} < \alpha$. Finally, using Sanov's theorem (Cover & Thomas, 2006, Chap. 11), it follows that the exponent of the probability that this condition is satisfied is given by (17).

□

[8] Note that we have taken a slightly pessimistic approach where an error is declared in case of a tie $v_2 = v_3$. In practice, the ties can be broken by a coin toss, but this does not affect the error exponent.

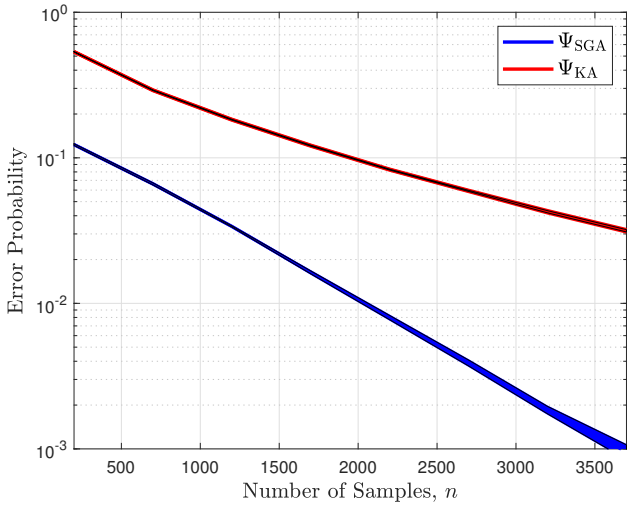


Figure 8. Comparison of error probabilities for 4-node noiseless chains when all the edge correlation are equal to $\rho = 0.4$.

H. Simulation results for 4-node homogeneous trees

Sec. 6.3 presented numerical results comparing the error exponents using the Ψ_{SGA} and Ψ_{KA} algorithms, derived using the *large deviation theory* (Cover & Thomas, 2006, Sec. 11.4), for 4-node homogeneous trees. In this appendix, we present Monte Carlo simulation results for 4-node homogeneous trees, that corroborate the results in Sec. 6.3. For any given tree structure, and any value of n (number of samples), the error probability using Ψ_{SGA} or Ψ_{KA} is computed based on 10^5 iterations (or runs) in the simulation setup.

Fig. 8 compares the error probabilities for 4 node noiseless chains, using Ψ_{SGA} and Ψ_{KA}, when all the edge correlation are equal to $\rho = 0.4$. Here, we consider all 12 distinct chain structures using 4 nodes (based on different permutations of the node indices), and for any given n we compute the empirical mean, denoted μ_n, and the empirical standard deviation, denoted σ_n, of the error probabilities using the 12 distinct chain structures. The shaded area in blue (resp. red) in Fig. 8 corresponds to the region between $\mu_n(\Psi_{\mathrm{SGA}}) + \sigma_n(\Psi_{\mathrm{SGA}})$ and $\mu_n(\Psi_{\mathrm{SGA}}) - \sigma_n(\Psi_{\mathrm{SGA}})$ (resp. between $\mu_n(\Psi_{\mathrm{KA}}) + \sigma_n(\Psi_{\mathrm{KA}})$ and $\mu_n(\Psi_{\mathrm{KA}}) - \sigma_n(\Psi_{\mathrm{KA}})$). The negative slope of the error probability curve is indicative of the error exponent, and Fig. 8 demonstrates that the error exponent using Ψ_{SGA} is much higher than that using Ψ_{KA} when $\rho = 0.4$, as shown by the corresponding error exponent values in Fig. 1(a).

Fig. 9 compares the error probabilities for 4 node noiseless chains, using Ψ_{SGA} and Ψ_{KA}, when all the edge correlation are equal to $\rho = 0.8$. Again, the shaded area in blue (resp. red) in Fig. 9 corresponds to the region between $\mu_n(\Psi_{\mathrm{SGA}}) + \sigma_n(\Psi_{\mathrm{SGA}})$ and $\mu_n(\Psi_{\mathrm{SGA}}) - \sigma_n(\Psi_{\mathrm{SGA}})$ (resp. between $\mu_n(\Psi_{\mathrm{KA}}) + \sigma_n(\Psi_{\mathrm{KA}})$ and $\mu_n(\Psi_{\mathrm{KA}}) - \sigma_n(\Psi_{\mathrm{KA}})$), where μ_n denotes the empirical mean and σ_n denotes the empirical standard deviation, for the error probabilities obtained using the 12 distinct chain structures. Fig. 9 shows that the slope of the error probability curves for Ψ_{SGA} and Ψ_{KA} are roughly equal when $\rho = 0.8$, as indicated by the corresponding error exponent values in Fig. 1(a).

Fig. 10 compares the error probabilities for 4 node star structured trees, using Ψ_{SGA} and Ψ_{KA}. The shaded area in blue (resp. red) in Fig. 10 corresponds to the region between $\mu_n(\Psi_{\mathrm{SGA}}) + \sigma_n(\Psi_{\mathrm{SGA}})$ and $\mu_n(\Psi_{\mathrm{SGA}}) - \sigma_n(\Psi_{\mathrm{SGA}})$ (resp. between $\mu_n(\Psi_{\mathrm{KA}}) + \sigma_n(\Psi_{\mathrm{KA}})$ and $\mu_n(\Psi_{\mathrm{KA}}) - \sigma_n(\Psi_{\mathrm{KA}})$), where μ_n denotes the empirical mean and σ_n denotes the empirical standard deviation, for the error probabilities obtained using the 4 distinct star structured trees with 4 nodes (based on different permutations of the node indices). Fig. 10 shows that the slopes of the error probability curves for Ψ_{SGA} and Ψ_{KA} are not very different when $\rho = 0.6$, as suggested by the error exponent values in Fig. 2(a).

Overall, the simulation results in Fig. 8 and Fig. 9 show that the error probability values do not deviate much due to the specific choice of a 4-node chain structure, while Fig. 10 indicates a similar behavior for 4-node star structured trees.

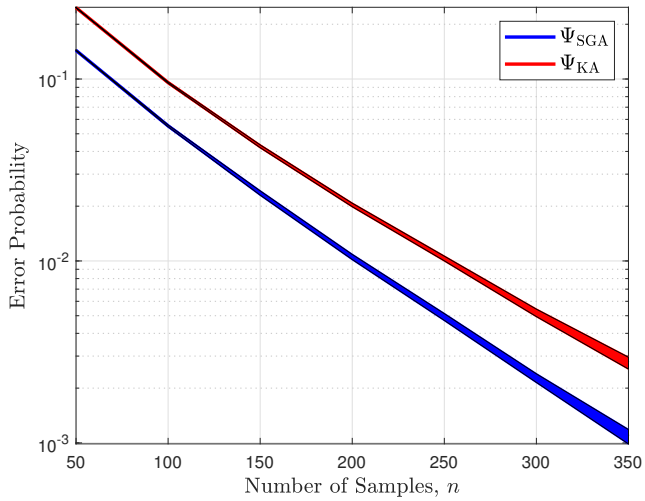


Figure 9. Comparison of error probabilities for 4-node noiseless chains when all the edge correlation are equal to $\rho = 0.8$.

I. Different 12-node tree structures considered in Section 7

Fig. 11 presents 12-node trees with three different tree structures: (i) Chain, (ii) Hybrid, (iii) Star. The chain and the star structures are known to be extremal tree structures in terms of the error probability (Tan et al., 2010; Tandon et al., 2020), while the hybrid tree structure is a combination of the chain and star structures.

J. Extension of Katiyar et al. (2020) and SGA to Gaussian trees with numerical results

We compare the error probabilities of Ψ_{SGA} and Ψ_{KA} in recovering the trees from tree-structured Gaussian graphical models with $d = 10$ nodes. The tree structures used for comparison are similar to those in the Ising model experiments in Sec. 7. Each observation of $X_i \in \mathbb{R}$ is corrupted by independent but non-identically distributed Gaussian noise such that the observed variable $Y_i = X_i + N_i$, where the noise $N_i \sim \mathcal{N}(0, \sigma_i^2)$ for some $\sigma_i > 0$.

J.1. Experiment setup

Generating samples for Gaussian models begins with choosing a tree structure $\mathrm{T}_P = (\mathcal{V}, \mathcal{E}_P)$ where the number of nodes $d = 10$. We then generate the inverse covariance matrix $(\mathbf{\Sigma}^*)^{-1}$, by setting its $(i,j)^{\text{th}}$ entry as

$$[(\mathbf{\Sigma}^*)^{-1}]_{i,j} = \begin{cases} w, & \text{if } (i,j) \in \mathcal{E}_P; \\ 1, & \text{if } i = j; \\ 0 & \text{otherwise} \end{cases} \tag{52}$$

for some parameter $w \in \mathbb{R}$. This matrix is then inverted to get the covariance matrix $\mathbf{\Sigma}^*$ of the distribution P. The correlation matrix $\mathbf{K}^*$ is calculated from $\mathbf{\Sigma}^*$ using the formula $\mathbf{K}^* = (\mathrm{diag}(\mathbf{\Sigma}^*))^{-\frac{1}{2}} \mathbf{\Sigma}^* (\mathrm{diag}(\mathbf{\Sigma}^*))^{-\frac{1}{2}}$. This is used to compute the minimum and maximum correlation coefficients $\rho_{\min}$ and $\rho_{\max}$. The parameter $w \in \mathbb{R}$ is chosen so as to ensure that $\rho_{\max} \approx 0.8$ from the resulting $\mathbf{K}^*$. Additionally, for the noiseless case, the diagonal matrix $\mathbf{D}^*$ is taken to be the zero matrix, and for the noisy case $[\mathbf{D}^*]_{i,i} = 2$ for $i \in \{1,3,5,7,9\}$; thus Gaussian noise of variance 2 is directly added to the node observations for nodes with odd indices. Finally, samples were generated from the joint Gaussian distribution $\tilde{P}(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mathbf{0}, \mathbf{\Sigma}^* + \mathbf{D}^*)$.

J.2. Modifications to the Algorithm

Even though Katiyar et al. (2019) proposed an algorithm for the partial learning of Gaussian graphical models (given noisy observations), it is not directly implementable to the case in which we have a *finite* number of samples n. An algorithm for learning trees with noisy samples up to their equivalence classes was proposed by Katiyar et al. (2020) but the algorithm

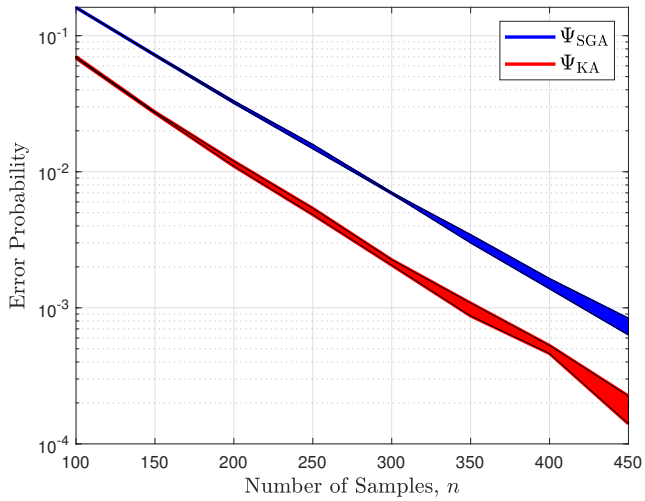
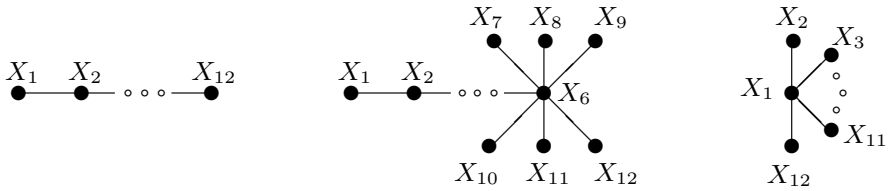


Figure 10. Comparison of error probabilities for 4-node star structured trees when all the edge correlation are equal to $\rho = 0.6$ and $q_{\max} = 0$.



(i) Chain (ii) Hybrid (iii) Star

Figure 11. Three different 12-node tree structures

provided therein was originally used to recover *Ising models*. Hence, some modifications to the algorithm had to be made so that it is amenable to Gaussian graphical models. In particular, there are two thresholds given to form the proximal set of any node i in the Ising case. The first, $t_1 \triangleq (1-2q_{\max})^2\rho_{\min}^4$, gives a lower bound for the correlation between i any other node with distance at most 4 between them. The second, $t_2 \triangleq \min\left\{t_1, \frac{t_1(1-2q_{\max})}{\rho_{\max}}\right\}$, gives another lower bound for the correlation of i and the first node in the path i to some other node j where $\tilde{\rho}_{i,j} \geq t_1$.

A similar idea can be used to construct proximal sets for the Gaussian case using the correlation decay property. First, note that the correlation coefficient between the variable X_i and its noisy counterpart $Y_i = X_i + N_i$ is

$$\rho_{X_iY_i} = \frac{\mathbb{E}[X_iY_i]}{\sqrt{\mathbb{E}[X_i^2]\mathbb{E}[Y_i^2]}} = \frac{\mathbb{E}[X_i(X_i+N_i)]}{\sqrt{\mathbb{E}[X_i^2]\mathbb{E}[(X_i+N_i)^2]}} = \frac{\mathbb{E}[X_i^2]}{\sqrt{\mathbb{E}[X_i^2](\mathbb{E}[X_i^2]+\mathbb{E}[N_i^2])}} = \frac{1}{\sqrt{1+\frac{\sigma_i^2}{\mathbb{E}[X_i^2]}}}. \tag{53}$$

Let $S_i \triangleq \sigma_i^2/\mathbb{E}[X_i^2]$ and $S_{\max} \triangleq \max_{1\leq i\leq d} S_i$. With the correlation decay property for Gaussian tree models (Tan et al., 2010, Eqn. (18)), any node j within radius 4 of i will have its noisy correlation coefficient bounded as follows

$$\rho_{Y_iY_j} = \rho_{Y_iX_i}\cdot\rho_{X_iX_j}\cdot\rho_{X_jY_j} \geq \frac{1}{\sqrt{1+\frac{\sigma_i^2}{\mathbb{E}[X_i^2]}}}\cdot\rho_{\min}^4\cdot\frac{1}{\sqrt{1+\frac{\sigma_j^2}{\mathbb{E}[X_j^2]}}} \geq \rho_{\min}^4\cdot\frac{1}{1+S_{\max}}. \tag{54}$$

In this way, we obtain the first threshold $h_1 \triangleq \frac{\rho_{\min}^4}{1+S_{\max}}$. Similarly, if we let node k be the first node in the path from j to i (i.e., k is at distance 1 from j), then it holds that

$$\frac{\rho_{Y_iY_j}}{\rho_{X_kX_j}} = \underbrace{\rho_{X_iY_i}\cdot\rho_{X_iX_k}\cdot\rho_{X_kY_k}}_{\rho_{Y_iY_k}}\cdot\frac{\rho_{X_jY_j}}{\rho_{X_kY_k}}, \tag{55}$$

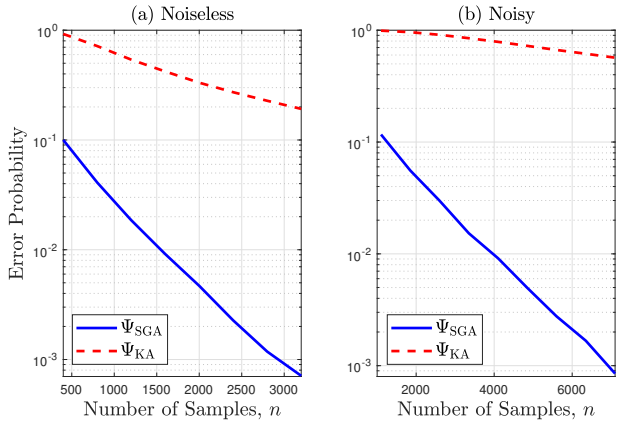


Figure 12. Comparison of error probabilities for a 10-node chain, with $w = 0.5$ (see (52)).

whence, by (53) and the definition of S_i, the correlation coefficient between Y_i and Y_k can be bounded as follows

$$\rho_{Y_iY_k} = \frac{\rho_{Y_iY_j}}{\rho_{X_kX_j}} \cdot \sqrt{\frac{1+S_j}{1+S_k}} \geq \frac{h_1}{\rho_{\max}\sqrt{1+S_{\max}}}. \tag{56}$$

Thus, we get our second threshold $h_2 \triangleq \min\left\{h_1, \frac{h_1}{\rho_{\max}\sqrt{1+S_{\max}}}\right\}$. So the proximal set of node i is defined as the set of all nodes j that satisfy $|\hat{\rho}_{i,j}| \geq 0.5h_2$.

J.3. Numerical Results

We now present numerical results for Gaussian tree models, comparing the performance of Ψ_{SGA} and Ψ_{KA} for three different tree structures with $d = 10$ nodes: (i) Chain, (ii) Hybrid, and (iii) Star. 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 (or runs) in the Monte Carlo simulation framework, where an error is declared if the estimated tree does not belong to the equivalence class $[\mathrm{T}]$. For the noisy case, we set $[\mathbf{D}^*]_{i,i} = 2$ for $i \in \{1, 3, 5, 7, 9\}$.

J.3.1. 10-NODE CHAIN

Fig. 12 plots the results for a 10-node chain for the (a) noiseless and (b) noisy cases, with $w = 0.5$ (see (52)). It is seen that Ψ_{SGA} significantly outperforms Ψ_{KA} for the Gaussian chain; a similar trend was observed for the Ising chain in Fig. 3.

J.3.2. 10-NODE HYBRID TREE STRUCTURE

Fig. 13 plots the results for a 10-node hybrid tree structure for the (a) noiseless and (b) noisy cases, with $w = 0.38$ (see (52)). The hybrid tree structure is a combination of chain and star structures where nodes 1 to 5 are linked in the form of a chain, while nodes 6 to 10 are directly connected to node 5. Similar to the performance comparison for the Ising hybrid tree in Fig. 4, it is seen that Ψ_{SGA} significantly outperforms Ψ_{KA} for the Gaussian hybrid tree.

J.3.3. 10-NODE STAR

Fig. 14 plots the results for a 10-node star tree structure for the (a) noiseless and (b) noisy cases, with $w = 0.325$ (see (52)), where nodes 2 to 10 are directly connected to node 1. It is seen that the performance of Ψ_{SGA} is only slightly better than that of Ψ_{KA}; a similar trend was observed for the Ising star tree in Fig. 5. This is corroborated by the error exponent results in Fig. 2, which is for Ising models but we expect the same behavior for Gaussian models.

These experiments for Gaussian tree models learned using noisy samples demonstrate that the behavior of SGA and the algorithm by Katiyar et al. (2020) are qualitatively very similar to the Ising case as detailed in Sec. 7. In particular, the performance of SGA is far superior to that of Katiyar et al. (2020) for chains and hybrid trees (i.e., trees with moderate to large diameter). SGA's performance is comparable to the algorithm by Katiyar et al. (2020) for stars (i.e., trees with small diameter), though in this case, SGA's performance is still marginally better.

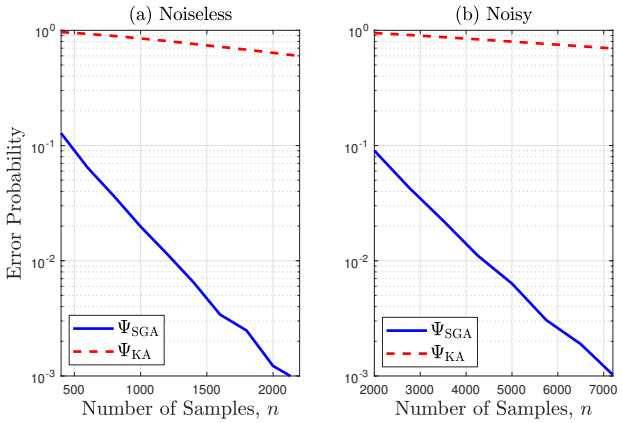


Figure 13. Comparison of error probabilities for a 10-node hybrid tree structure, with $w = 0.38$.

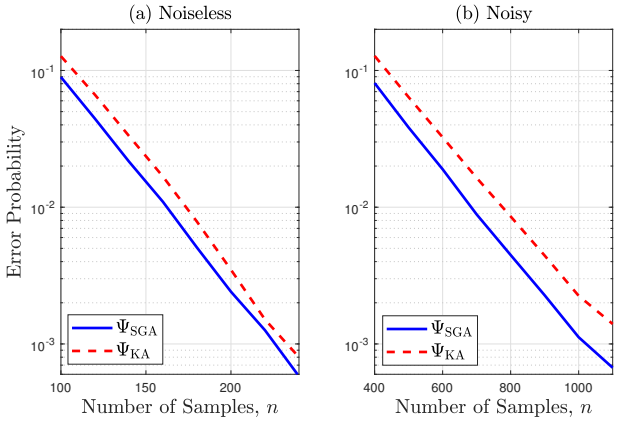


Figure 14. Comparison of error probabilities for a 10-node star tree, with $w = 0.325$.