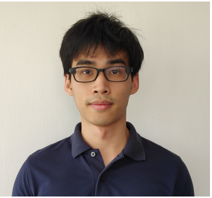
# Learning Tree Models in Noise: Exact Asymptotics and Robust Algorithms

**Georgia Tech ML Seminar**


**Vincent Y. F. Tan (ECE, Maths)**
**National University of Singapore**


Joint work with:

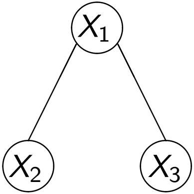
Anshoo Tandon (ECE)



Aldric Han (Maths)

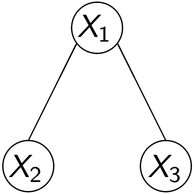
Shiyao Zhu (ECE)

# Graphical Models




- Marriage of probability theory and graph theory
- Nodes correspond to random variables
- Edges represent statistical dependence between variables

# Graphical Models




- Marriage of probability theory and graph theory
- Nodes correspond to random variables
- Edges represent statistical dependence between variables
- Graphical models encode conditional independence between variables
- $X_2$ and $X_3$ are conditionally independent, given $X_1$, i.e.,

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1)$$

# Graphical Models

- In general, the distribution $P$, of a random vector $\mathbf{X} := [X_1, \ldots, X_p]$, with corresponding graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, satisfies the property:

$$P(x_i | x_{\mathcal{V} \setminus i}) = P(x_i | x_{\mathrm{nbd}(i)}),$$

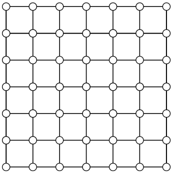
where $\mathrm{nbd}(i) := \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}\}$ is the neighborhood of node $i$.

# Graphical Models

- In general, the distribution $P$, of a random vector $\mathbf{X} := [X_1, \ldots, X_p]$, with corresponding graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, satisfies the 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 $i$.
- Graphical models have found extensive application in
  - Image denoising
  - Natural language processing
  - Combinatorial optimization
- Example: Lattice graphical model for image pixels



# Tree-structured Graphical Models

- We study tree-structured graphical models over $p$ nodes
- In an undirected tree, we may assume that node 1 is the *root* node
- So, if $\mathbf{x} := (x_1, \ldots, x_p)$, then graphical model $P$ factors as

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

where $\mathrm{pa}(i)$ denotes the unique parent node of node $i$.

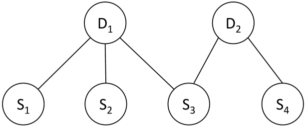
# Tree-structured Graphical Models

- We study tree-structured graphical models over $p$ nodes
- In an undirected tree, we may assume that node 1 is the *root* node
- So, if $\mathbf{x} := (x_1, \ldots, x_p)$, then graphical model $P$ factors as

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

where $\mathrm{pa}(i)$ denotes the unique parent node of node $i$.

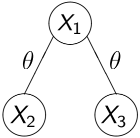
- A simple example (biomedical):
  $D_i$ nodes: Diseases
  $S_j$ nodes: Symptoms



- Part I
  - Homogeneous tree model
  - Identically distributed noise
  - Exact asymptotics using *strong* large deviation theory
- Part II
  - Non-identically distributed noise
  - Exact tree structure recovery may be impossible in some cases
  - Robust Learning: Partial tree structure recovery (up to equivalence class) under non-identical noise

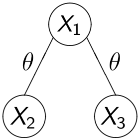
- # Part I
  - Homogeneous tree model
  - Identically distributed noise
  - Exact asymptotics using *strong* large deviation theory

# Tree-structured Graphical Models: System Model




- We consider binary random variables with alphabet $\mathcal{X} = \{0, 1\}$

# Tree-structured Graphical Models: System Model




- We consider binary random variables with alphabet $\mathcal{X} = \{0, 1\}$
- Graphical model $P$, for $p > 2$ nodes, has the following properties:

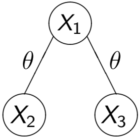
P1: Zero external field: The marginals are uniform, i.e.

$$P_i(0) = P_i(1) = 0.5, \qquad 1 \leq i \leq p.$$

P2: $\theta$-Homogeneity: For every edge $\{i, j\} \in \mathcal{E}_P$, we have

$$P_{i,j}(0, 1) = P_{i,j}(1, 0) = \frac{\theta}{2}, \qquad \theta \in (0, 0.5).$$

# Tree-structured Graphical Models: System Model




- We consider binary random variables with alphabet $\mathcal{X} = \{0, 1\}$
- Graphical model $P$, for $p > 2$ nodes, has the following properties:

P1: Zero external field: The marginals are uniform, i.e.

$$P_i(0) = P_i(1) = 0.5, \qquad 1 \leq i \leq p.$$

P2: $\theta$-Homogeneity: For every edge $\{i, j\} \in \mathcal{E}_P$, we have

$$P_{i,j}(0, 1) = P_{i,j}(1, 0) = \frac{\theta}{2}, \qquad \theta \in (0, 0.5).$$

- Corresponds to homogeneous Ising model with zero external field
- $\theta$ can be viewed as the crossover probability
- $0 < \theta < 0.5$ implies a positive correlation along the edges

# Problem: Tree Learning with Side Information

- Given $n$ i.i.d. $p$-dimensional samples $\mathbf{x}^n := \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ from an unknown $P \in \mathcal{D}(\mathcal{T}^p, \theta)$, where

$$\mathcal{D}(\mathcal{T}^p, \theta) = \left\{ \begin{array}{l} \text{tree distributions on } \{0,1\}^p \text{ satisfying} \\ \text{Zero external field } \& \ \theta\text{-Homogeneity} \end{array} \right\}.$$

- Problem: Given $\mathbf{x}^n$, learn the underlying tree structure of $P$ with side information that $P$ satisfies Zero external field & $\theta$-Homogeneity.

# Problem: Tree Learning with Side Information

- Given $n$ i.i.d. $p$-dimensional samples $\mathbf{x}^n := \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ from an unknown $P \in \mathcal{D}(\mathcal{T}^p, \theta)$, where

$$\mathcal{D}(\mathcal{T}^p, \theta) = \left\{ \begin{array}{l} \text{tree distributions on } \{0,1\}^p \text{ satisfying} \\ \text{Zero external field } \& \ \theta\text{-Homogeneity} \end{array} \right\}.$$

- Problem: Given $\mathbf{x}^n$, learn the underlying tree structure of $P$ with side information that $P$ satisfies Zero external field & $\theta$-Homogeneity.
- Error event:

$$\mathcal{A}_P(n) := \{\mathcal{E}_{\mathrm{ML}}(\mathbf{x}^n) \neq \mathcal{E}_P\}$$

- Given $\mathbf{x}^n$, the ML estimator of the unknown distribution $P$ is

$$P_{\mathrm{ML}}(\mathbf{x}^n) := \mathop{\arg\max}_{Q \in \mathcal{D}(\mathcal{T}^p, \theta)} \sum_{k=1}^{n} \log Q(\mathbf{x}_k)$$

# Maximum Likelihood Estimation

- Given $\mathbf{x}^n$, its empirical distribution (or joint type) is

$$\widehat{P}(\mathbf{x}) := \frac{1}{n}\sum_{k=1}^{n} 1\{\mathbf{x}_k = \mathbf{x}\}, \quad \mathbf{x} \in \mathcal{X}^p,$$

# Maximum Likelihood Estimation

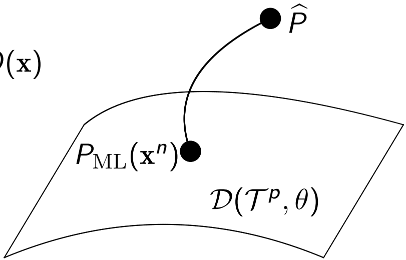
- Given $\mathbf{x}^n$, its empirical distribution (or joint type) is

$$\widehat{P}(\mathbf{x}) := \frac{1}{n}\sum_{k=1}^{n} 1\{\mathbf{x}_k = \mathbf{x}\}, \quad \mathbf{x} \in \mathcal{X}^p,$$

- We have

$$\begin{aligned} P_{\mathrm{ML}}(\mathbf{x}^n) &= \operatorname*{arg\,max}_{Q \in \mathcal{D}(\mathcal{T}^p,\theta)} \sum_{\mathbf{x} \in \mathcal{X}^p} \widehat{P}(\mathbf{x}) \log Q(\mathbf{x}) \\ &= \operatorname*{arg\,min}_{Q \in \mathcal{D}(\mathcal{T}^p,\theta)} D(\widehat{P} \| Q) \end{aligned}$$

Reverse I-Projection



# Maximum Likelihood Estimation

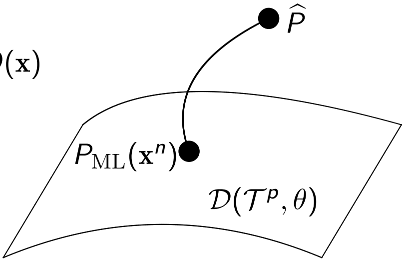
- Given $\mathbf{x}^n$, its empirical distribution (or joint type) is

$$\widehat{P}(\mathbf{x}) := \frac{1}{n}\sum_{k=1}^{n} 1\{\mathbf{x}_k = \mathbf{x}\}, \quad \mathbf{x} \in \mathcal{X}^p,$$

- We have

$$P_{\mathrm{ML}}(\mathbf{x}^n) = \underset{Q\in\mathcal{D}(\mathcal{T}^p,\theta)}{\arg\max} \sum_{\mathbf{x}\in\mathcal{X}^p} \widehat{P}(\mathbf{x}) \log Q(\mathbf{x})$$
$$= \underset{Q\in\mathcal{D}(\mathcal{T}^p,\theta)}{\arg\min} \; D(\widehat{P}\|Q)$$

Reverse I-Projection




- When $\theta$ is known, $\{\mathcal{E}_{\mathrm{ML}}(\mathbf{x}^n) \neq \mathcal{E}_P\} = \{P_{\mathrm{ML}}(\mathbf{x}^n) \neq P\}$

# Maximum Likelihood Estimation: Simplified formulation

- Let $\widehat{P}_{i,j}(x_i, x_j)$ denote the marginal of $\widehat{P}(\mathbf{x})$ on the pair of nodes $(i, j)$, with $i \neq j$, and define $\widehat{A}_{i,j}$ as

$$\widehat{A}_{i,j} := \widehat{P}_{i,j}(0,0) + \widehat{P}_{i,j}(1,1)$$

  $\widehat{A}_{i,j}$ can be interpreted at the agreement of nodes $i$ and $j$.

# Maximum Likelihood Estimation: Simplified formulation

- Let $\widehat{P}_{i,j}(x_i, x_j)$ denote the marginal of $\widehat{P}(\mathbf{x})$ on the pair of nodes $(i, j)$, with $i \neq j$, and define $\widehat{A}_{i,j}$ as

$$\widehat{A}_{i,j} := \widehat{P}_{i,j}(0,0) + \widehat{P}_{i,j}(1,1)$$

  $\widehat{A}_{i,j}$ can be interpreted at the agreement of nodes $i$ and $j$.
- $\mathcal{E}_{\mathrm{ML}}(\mathbf{x}^n)$ can be obtained as the edge set of a maximum weight spanning tree (MWST)

# Maximum Likelihood Estimation: Simplified formulation

- Let $\widehat{P}_{i,j}(x_i, x_j)$ denote the marginal of $\widehat{P}(\mathbf{x})$ on the pair of nodes $(i, j)$, with $i \neq j$, and define $\widehat{A}_{i,j}$ as

$$\widehat{A}_{i,j} := \widehat{P}_{i,j}(0,0) + \widehat{P}_{i,j}(1,1)$$

  $\widehat{A}_{i,j}$ can be interpreted at the agreement of nodes $i$ and $j$.
- $\mathcal{E}_{\mathrm{ML}}(\mathbf{x}^n)$ can be obtained as the edge set of a maximum weight spanning tree (MWST)
- Weights of the MWST are $\{\widehat{A}_{i,j}\}$. Equivalently,

$$P_{\mathrm{ML}}(\mathbf{x}^n) = \operatorname*{arg\,max}_{Q \in \mathcal{D}(\mathcal{T}^p, \theta)} \sum_{\{i,j\} \in \mathcal{E}_Q} \widehat{A}_{i,j},$$

  where $\mathcal{E}_Q$ denotes the edge set of the tree distribution $Q$

# Classical Chow-Liu alogorithm

- In the absence of the side information—Zero external field and $\theta$-Homogeneity—tree can be learned via the Chow-Liu algorithm [IT'68] where

$$\mathcal{E}_{\mathrm{CL}}(\mathbf{x}^n) = \operatorname*{arg\,max}_{\mathcal{E} \text{ is a tree}} \sum_{\{i,j\}\in\mathcal{E}} \widehat{I}_{i,j},$$

where $\widehat{I}_{i,j}$ is the empirical mutual information

$$\widehat{I}_{i,j} = I(\widehat{P}_{i,j}) := \sum_{(x_i,x_j)\in\mathcal{X}^2} \widehat{P}_{i,j}(x_i,x_j) \log \frac{\widehat{P}_{i,j}(x_i,x_j)}{\widehat{P}_i(x_i)\widehat{P}_j(x_j)}$$

# Classical Chow-Liu alogorithm

- In the absence of the side information—Zero external field and $\theta$-Homogeneity—tree can be learned via the Chow-Liu algorithm [IT'68] where

$$\mathcal{E}_{\mathrm{CL}}(\mathbf{x}^n) = \operatorname*{arg\,max}_{\mathcal{E} \text{ is a tree}} \sum_{\{i,j\} \in \mathcal{E}} \widehat{I}_{i,j},$$

where $\widehat{I}_{i,j}$ is the empirical mutual information

$$\widehat{I}_{i,j} = I(\widehat{P}_{i,j}) := \sum_{(x_i,x_j) \in \mathcal{X}^2} \widehat{P}_{i,j}(x_i, x_j) \log \frac{\widehat{P}_{i,j}(x_i, x_j)}{\widehat{P}_i(x_i)\widehat{P}_j(x_j)}$$

- Agreement $\widehat{A}_{i,j}$ simplifies $\widehat{I}_{i,j}$ with side information.

$$\widehat{A}_{i,j} := \widehat{P}_{i,j}(0,0) + \widehat{P}_{i,j}(1,1)$$

# Error Exponent

The error exponent (using the ML algorithm) is defined as

$$K_P := \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(\mathcal{E}_{\mathrm{ML}}(\mathbf{x}^n) \neq \mathcal{E}_P\right)$$

$K_P$ characterizes the exponential decay rate of error probability with $n$, i.e.,

$$\mathbb{P}\left(\mathcal{E}_{\mathrm{ML}}(\mathbf{x}^n) \neq \mathcal{E}_P\right) \approx \exp\big(-nK_P\big).$$

## Error Exponent

The error exponent (using the ML algorithm) is defined as

$$K_P := \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(\mathcal{E}_{\mathrm{ML}}(\mathbf{x}^n) \neq \mathcal{E}_P\right)$$

$K_P$ characterizes the exponential decay rate of error probability with $n$, i.e.,

$$\mathbb{P}\left(\mathcal{E}_{\mathrm{ML}}(\mathbf{x}^n) \neq \mathcal{E}_P\right) \approx \exp\big(-nK_P\big).$$

**Theorem 1 (Tandon, T. and Zhu (2020))**

*For $P \in \mathcal{D}(\mathcal{T}^p, \theta)$, we have*

$$K_P = \log \frac{1}{1-\theta\big(1-\sqrt{4\theta(1-\theta)}\big)} \quad \textit{and} \quad K_P = K_P^{\mathrm{CL}}.$$

# Error Exponent

The error exponent (using the ML algorithm) is defined as

$$K_P := \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(\mathcal{E}_{\mathrm{ML}}(\mathbf{x}^n) \neq \mathcal{E}_P\right)$$

$K_P$ characterizes the exponential decay rate of error probability with $n$, i.e.,

$$\mathbb{P}\left(\mathcal{E}_{\mathrm{ML}}(\mathbf{x}^n) \neq \mathcal{E}_P\right) \approx \exp\left( - nK_P\right).$$

**Theorem 1 (Tandon, T. and Zhu (2020))**

*For $P \in \mathcal{D}(\mathcal{T}^p, \theta)$, we have*

$$K_P = \log \frac{1}{1-\theta\left(1-\sqrt{4\theta(1-\theta)}\right)} \quad \textit{and} \quad K_P = K_P^{\mathrm{CL}}.$$

- $K_P^{\mathrm{CL}} = K_P \implies$ No advantage (from the error exponent perspective) in having the side information of zero external field and Homogeneity.
- When the sample size is extremely small, side information yields smaller error probabilities over the vanilla Chow-Liu procedure

# Related Work: Bresler and Karzand [Ann. Statist. 2020]

- Bresler and Karzand considered general Ising tree models, that allowed for different correlations along the edges
- Provided a non-asymptotic upper bound on the error probability

$$\mathbb{P}\left(\mathcal{E}_{\mathrm{ML}}(\mathbf{x}^n) \neq \mathcal{E}_P\right) \leq 2p^2 \exp\left(-nK_P^{\mathrm{BK}}\right),$$

where the exponent $K_P^{\mathrm{BK}}$, upon specializing to our model, is

$$K_P^{\mathrm{BK}} = \frac{\theta\,(1-2\theta)^2}{8}$$

# Related Work: Bresler and Karzand [Ann. Statist. 2020]

- Bresler and Karzand considered general Ising tree models, that allowed for different correlations along the edges
- Provided a non-asymptotic upper bound on the error probability

$$\mathbb{P}\left(\mathcal{E}_{\mathrm{ML}}(\mathbf{x}^n) \neq \mathcal{E}_P\right) \leq 2p^2 \exp\left(-nK_P^{\mathrm{BK}}\right),$$

where the exponent $K_P^{\mathrm{BK}}$, upon specializing to our model, is

$$K_P^{\mathrm{BK}} = \frac{\theta\left(1-2\theta\right)^2}{8}$$

- For any $P \in \mathcal{D}(\mathcal{T}^p, \theta)$, we have

$$K_P^{\mathrm{BK}} < \frac{K_P}{3}$$

# Related Work: Bresler and Karzand [Ann. Statist. 2020]

- Bresler and Karzand considered general Ising tree models, that allowed for different correlations along the edges
- Provided a non-asymptotic upper bound on the error probability

$$\mathbb{P}\left(\mathcal{E}_{\mathrm{ML}}(\mathbf{x}^n) \neq \mathcal{E}_P\right) \leq 2p^2 \exp\left(-nK_P^{\mathrm{BK}}\right),$$

where the exponent $K_P^{\mathrm{BK}}$, upon specializing to our model, is

$$K_P^{\mathrm{BK}} = \frac{\theta\,(1-2\theta)^2}{8}$$

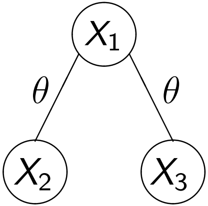
- For any $P \in \mathcal{D}(\mathcal{T}^p, \theta)$, we have

$$K_P^{\mathrm{BK}} < \frac{K_P}{3}$$

- Implies that BK's upper bound on the error probability is rather loose asymptotically

# Exact Asymptotics: 3 nodes

Let $P \in \mathcal{D}(\mathcal{T}^3, \theta)$, and define




$$\tilde{f}(n) := \frac{\exp(-nK_P)}{\sqrt{2\pi\sigma^2 n}} \left[1 + \frac{1-3\sigma^2}{8\sigma^2 n}\right],$$

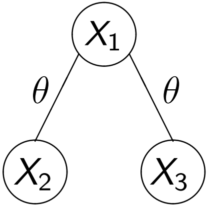
$$f(n) := \frac{\tilde{f}(n)}{1-z} \left[1 - \frac{z(1+z)}{2(1-z)^2\sigma^2 n}\right],$$

$$z := \sqrt{\frac{\theta}{1-\theta}},$$

$$\sigma^2 := \theta\sqrt{4\theta(1-\theta)}\,\exp(K_P)$$

# Exact Asymptotics: 3 nodes

Let $P \in \mathcal{D}(\mathcal{T}^3, \theta)$, and define




$$\tilde{f}(n) := \frac{\exp(-nK_P)}{\sqrt{2\pi\sigma^2 n}} \left[1 + \frac{1-3\sigma^2}{8\sigma^2 n}\right],$$

$$f(n) := \frac{\tilde{f}(n)}{1-z} \left[1 - \frac{z(1+z)}{2(1-z)^2\sigma^2 n}\right],$$

$$z := \sqrt{\frac{\theta}{1-\theta}},$$

$$\sigma^2 := \theta\sqrt{4\theta(1-\theta)}\,\exp(K_P)$$

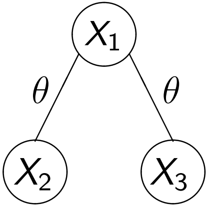
**Theorem 2 (Tandon, T. and Zhu (2020))**

*When ties are randomly broken in an MWST algorithm, then*

$$\mathbb{P}\left(\mathcal{E}_{\mathrm{ML}}(\mathbf{x}^n) \neq \mathcal{E}_P\right) = \left(2f(n) - \tilde{f}(n)\right)\left(1 + o(n^{-1})\right)$$

# Exact Asymptotics: 3 nodes

Let $P \in \mathcal{D}(\mathcal{T}^3, \theta)$, and define




$$\tilde{f}(n) := \frac{\exp(-nK_P)}{\sqrt{2\pi\sigma^2 n}} \left[1 + \frac{1-3\sigma^2}{8\sigma^2 n}\right],$$

$$f(n) := \frac{\tilde{f}(n)}{1-z} \left[1 - \frac{z(1+z)}{2(1-z)^2\sigma^2 n}\right],$$

$$z := \sqrt{\frac{\theta}{1-\theta}},$$

$$\sigma^2 := \theta\sqrt{4\theta(1-\theta)}\,\exp(K_P)$$

**Theorem 2 (Tandon, T. and Zhu (2020))**

*When ties are randomly broken in an MWST algorithm, then*

$$\mathbb{P}\left(\mathcal{E}_{\mathrm{ML}}(\mathbf{x}^n) \neq \mathcal{E}_P\right) = \left(2f(n) - \tilde{f}(n)\right)\left(1 + o(n^{-1})\right)$$

Strong large deviations by Blackwell and Hodges [Ann. Math. Statist.'59]

# Main Result: Exact Asymptotics ($p$ nodes)

- For $P \in \mathcal{D}(\mathcal{T}^p, \theta)$, let $d_i$ denote the degree of node $i$ in the tree corresponding to $P$, and define

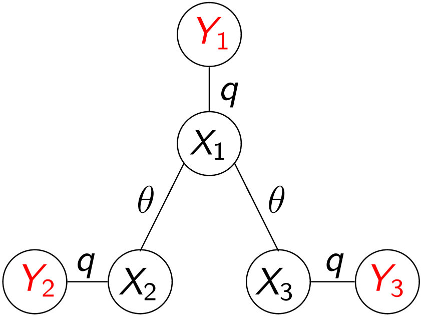
$$\zeta_P := \sum_{i=1}^{p} \frac{d_i(d_i - 1)}{2}$$

When ties are randomly broken in an MWST algorithm, then we have

$$\mathbb{P}\left(\mathcal{E}_{\mathrm{ML}}(\mathbf{x}^n) \neq \mathcal{E}_P\right) = \zeta_P\left(2f(n) - \tilde{f}(n)\right)\left(1 + o(n^{-1})\right)$$

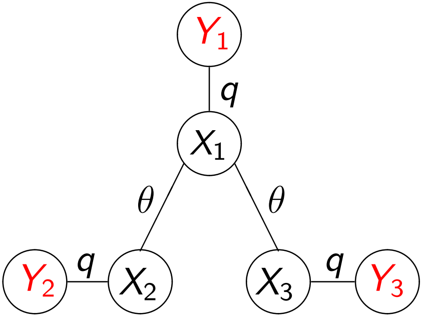
- $\zeta_P$ accounts for the number of 3-node sub-trees of $\mathcal{T}_P$ that contribute to dominant errors
- $f(n)$ and $\tilde{f}(n)$ do not depend on the particular choice of $P$, but the multiplicative factor $\zeta_P$ depends on the underlying tree structure

- Observed samples are noise-corrupted versions of the samples generated from the underlying tree-structured graphical model



# **Extension**: Noisy Samples Setting

- Observed samples are noise-corrupted versions of the samples generated from the underlying tree-structured graphical model




- Observe samples from $(Y_1, Y_2, Y_3)$ instead of $(X_1, X_2, X_3)$.
- Noise crossover probability $q$ constant across nodes.

# **Extension**: Noisy Samples Setting

- Observe noisy sample $\mathbf{y} = [y_1, \ldots, y_p] \sim P^{(q)}$, where $\mathbf{y}$ is the output when each component of $\mathbf{x}$ is passed through a BSC with crossover probability $0 \leq q < 0.5$

# **Extension**: Noisy Samples Setting

- Observe noisy sample $\mathbf{y} = [y_1, \ldots, y_p] \sim P^{(q)}$, where $\mathbf{y}$ is the output when each component of $\mathbf{x}$ is passed through a BSC with crossover probability $0 \leq q < 0.5$
- The distribution of the noisy samples $P^{(q)}$ is

$$P^{(q)}(\mathbf{y}) = \sum_{\mathbf{x} \in \mathcal{X}^p} q^{\delta_{\mathbf{x},\mathbf{y}}} (1-q)^{p-\delta_{\mathbf{x},\mathbf{y}}} P(\mathbf{x}), \qquad \mathbf{y} \in \mathcal{Y}^p = \{0,1\}^p,$$

where $\delta_{\mathbf{x},\mathbf{y}}$ denotes the Hamming distance between $\mathbf{x}$ and $\mathbf{y}$

# **Extension**: Noisy Samples Setting

- Observe noisy sample $\mathbf{y} = [y_1, \ldots, y_p] \sim P^{(q)}$, where $\mathbf{y}$ is the output when each component of $\mathbf{x}$ is passed through a BSC with crossover probability $0 \leq q < 0.5$

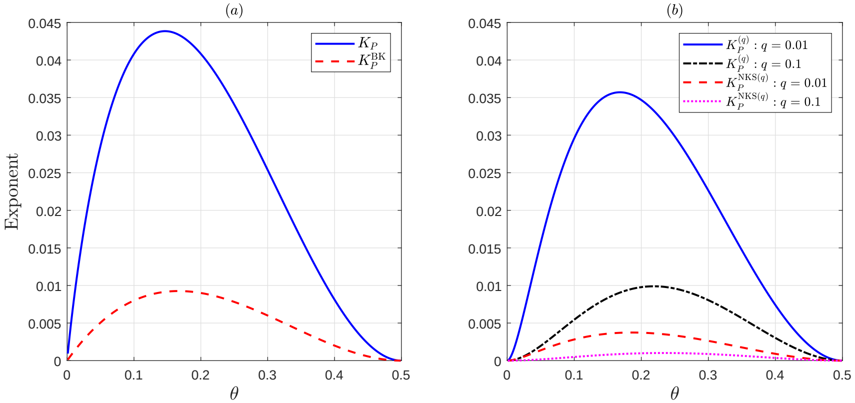
- The distribution of the noisy samples $P^{(q)}$ is

$$P^{(q)}(\mathbf{y}) = \sum_{\mathbf{x} \in \mathcal{X}^p} q^{\delta_{\mathbf{x},\mathbf{y}}} (1-q)^{p-\delta_{\mathbf{x},\mathbf{y}}} P(\mathbf{x}), \qquad \mathbf{y} \in \mathcal{Y}^p = \{0,1\}^p,$$

  where $\delta_{\mathbf{x},\mathbf{y}}$ denotes the Hamming distance between $\mathbf{x}$ and $\mathbf{y}$

- Extend our results for the tree learning problem with noisy samples, providing an explicit characterization of the error exponent and exact error asymptotics

# Numerical Results: Comparing Error Exponents

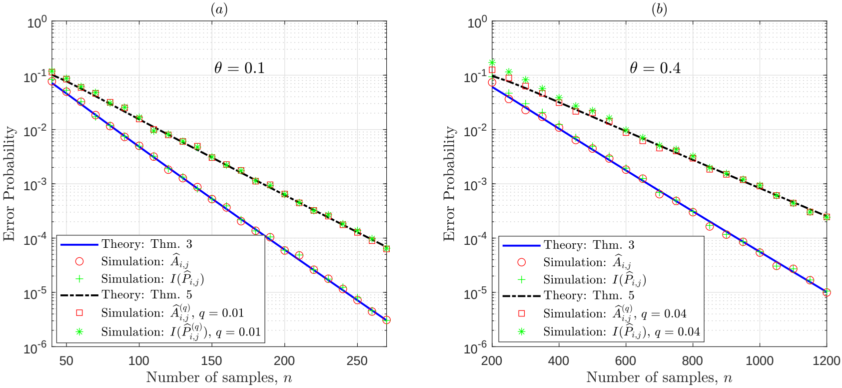



Comparison of error exponents using $(a)$ noiseless samples, and $(b)$ noisy samples

— $K_P^{\mathrm{BK}}$ is the exponent by Bresler and Karzand [Ann. Statist.'20]

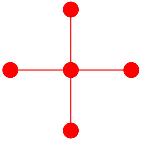
— $K_P^{\mathrm{NKS}(q)}$ is the exponent by Nikolakakis, Kalogerias, and Sarwate [AISTATS'19]

# Numerical Results: Error Asymptotics ($p = 3$)




Comparison of the theoretical error asymptotics for the noiseless and noisy sample setting, for a 3-node tree, with corresponding simulation results

# Extremal Tree Structures

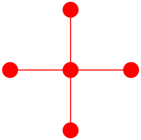


Star tree

Markov chain

- For a given $p$, $\zeta_P = \sum_{i=1}^{p} \frac{d_i(d_i-1)}{2}$ is maximized (resp. minimized) when the underlying tree structure is a star (resp. Markov chain)

# Extremal Tree Structures

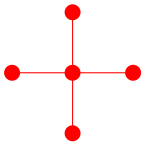


Star tree

Markov chain

- For a given $p$, $\zeta_P = \sum_{i=1}^{p} \frac{d_i(d_i-1)}{2}$ is maximized (resp. minimized) when the underlying tree structure is a star (resp. Markov chain)
- Specifically, $\zeta_P^{\mathrm{star}} = (p-1)(p-2)/2$ and $\zeta_P^{\mathrm{MC}} = p-2$.

# Extremal Tree Structures

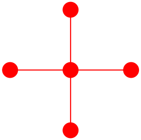


Star tree

Markov chain

- For a given $p$, $\zeta_P = \sum_{i=1}^{p} \frac{d_i(d_i-1)}{2}$ is maximized (resp. minimized) when the underlying tree structure is a star (resp. Markov chain)
- Specifically, $\zeta_P^{\mathrm{star}} = (p-1)(p-2)/2$ and $\zeta_P^{\mathrm{MC}} = p-2$.
- Error prob. is also max (resp. min) when tree is a star (resp. MC)

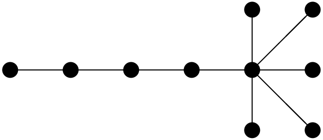
# Extremal Tree Structures



Star tree

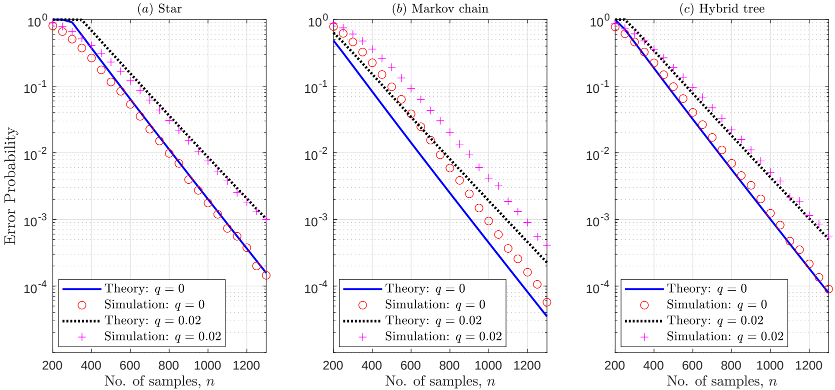
Markov chain

- For a given $p$, $\zeta_P = \sum_{i=1}^{p} \frac{d_i(d_i-1)}{2}$ is maximized (resp. minimized) when the underlying tree structure is a star (resp. Markov chain)
- Specifically, $\zeta_P^{\mathrm{star}} = (p-1)(p-2)/2$ and $\zeta_P^{\mathrm{MC}} = p-2$.
- Error prob. is also max (resp. min) when tree is a star (resp. MC)
- An intermediate tree-structure:



Hybrid tree

# Numerical Results: Error Asymptotics ($p = 10$)




Comparison of theoretical and simulation results for the noiseless ($q = 0$) and noisy sample setting ($q = 0.02$) for 10-node trees with $\theta = 0.4$

# Conclusions

- Strong large deviations approach to compute the exact asymptotics for learning trees given noiseless and noisy samples
- Refined estimates of the error probability in learning graphical models
  - For the noiseless and noisy cases respectively, we significantly improved on the error exponents derived by Bresler-Karzand [Ann. Statist.'20] and Nikolakakis-Kalogerias-Sarwate [AISTATS'19]
  - Our theoretical results were in keen agreement with numerical simulations at relatively small sample sizes
- Future work: High-dimensional setting where $p$ grows with $n$

# Part II: Partial tree recovery under non-identical noise

- Part I
  - Homogeneous tree model
  - Identically distributed noise
  - Exact asymptotics using *strong* large deviation theory
- Part II
  - Non-identically distributed noise
  - Exact tree structure recovery may be impossible in some cases
  - Robust Learning: Partial tree structure recovery (up to equivalence class) under non-identical noise

- Part II
  - Non-identically distributed noise
  - Exact tree structure recovery may be impossible in some cases
  - Robust Learning: Partial tree structure recovery (up to equivalence class) under non-identical noise

# Ising Models with Non-Identical Noisy Observations

- Random variables are zero mean with alphabet $\mathcal{X} = \{+1, -1\}$
- The joint distribution of $\mathbf{X} = (X_1, \ldots, X_p)$ is given by

$$P_{\mathbf{X}}(\mathbf{x}) = \frac{1}{Z} \exp\left( \sum_{\{i,j\} \in \mathcal{E}} \beta_{i,j}\, x_i x_j \right),$$

  where $Z$ is a normalization factor called the partition function
- For a tree, if $\{i, j\} \in \mathcal{E}$ then

$$\rho_{i,j} = \mathbb{E}[X_i X_j] = \tanh(\beta_{i,j})$$

# Ising Models with Non-Identical Noisy Observations

- Random variables are zero mean with alphabet $\mathcal{X} = \{+1, -1\}$
- The joint distribution of $\mathbf{X} = (X_1, \ldots, X_p)$ is given by

$$P_{\mathbf{X}}(\mathbf{x}) = \frac{1}{Z} \exp\Bigg( \sum_{\{i,j\} \in \mathcal{E}} \beta_{i,j}\, x_i x_j \Bigg),$$

  where $Z$ is a normalization factor called the partition function
- For a tree, if $\{i, j\} \in \mathcal{E}$ then

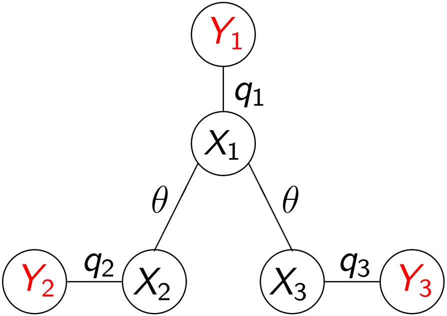
$$\rho_{i,j} = \mathbb{E}[X_i X_j] = \tanh(\beta_{i,j})$$

- Noise Model: Observe $Y_i = X_i N_i$, where

$$\Pr(N_i = -1) = q_i \quad \text{and} \quad \Pr(N_i = +1) = 1 - q_i$$

  with $0 \le q_i < 0.5$
- Observations are corrupted by independent, non-identical noise

# Ising Models with Non-Identical Noisy Observations




The $q_i$'s need not be the same!

# Classical Chow-Liu algorithm fails in general!

- Chow and Liu [IT'68] gave an elegant algorithm for learning a tree

$$\mathcal{E}_{\mathrm{CL}}(\mathbf{x}^n) = \operatorname*{arg\,max}_{\mathcal{E} \text{ is a tree}} \sum_{\{i,j\}\in\mathcal{E}} \widehat{I}_{i,j},$$

- Also works when $q_i = q$ for all $1 \le i \le p$ – Error exponent optimal!
- However, with non-identical noise, the Chow-Liu algorithm may not be able to recover the structure of the tree
- Note noisy correlation is $\tilde{\rho}_{i,j} = (1-2q_i)(1-2q_j)\rho_{i,j}$

Example:

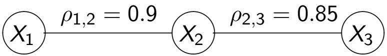
# Classical Chow-Liu algorithm fails in general!

- Chow and Liu [IT'68] gave an elegant algorithm for learning a tree

$$\mathcal{E}_{\mathrm{CL}}(\mathbf{x}^n) = \underset{\mathcal{E} \text{ is a tree}}{\arg\max} \sum_{\{i,j\} \in \mathcal{E}} \widehat{I}_{i,j},$$

- Also works when $q_i = q$ for all $1 \le i \le p$ – Error exponent optimal!
- However, with non-identical noise, the Chow-Liu algorithm may not be able to recover the structure of the tree
- Note noisy correlation is $\tilde{\rho}_{i,j} = (1 - 2q_i)(1 - 2q_j)\rho_{i,j}$

Example:



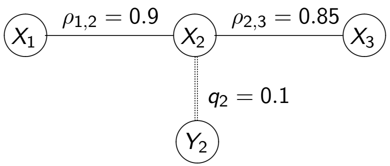
# Classical Chow-Liu algorithm fails in general!

- Chow and Liu [IT'68] gave an elegant algorithm for learning a tree

$$\mathcal{E}_{\mathrm{CL}}(\mathbf{x}^n) = \operatorname*{arg\,max}_{\mathcal{E} \text{ is a tree}} \sum_{\{i,j\} \in \mathcal{E}} \widehat{I}_{i,j},$$

- Also works when $q_i = q$ for all $1 \le i \le p$ – Error exponent optimal!
- However, with non-identical noise, the Chow-Liu algorithm may not be able to recover the structure of the tree
- Note noisy correlation is $\tilde{\rho}_{i,j} = (1 - 2q_i)(1 - 2q_j)\rho_{i,j}$

Example:



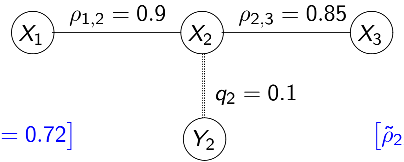
# Classical Chow-Liu algorithm fails in general!

- Chow and Liu [IT'68] gave an elegant algorithm for learning a tree

$$\mathcal{E}_{\mathrm{CL}}(\mathbf{x}^n) = \operatorname*{arg\,max}_{\mathcal{E} \text{ is a tree}} \sum_{\{i,j\}\in\mathcal{E}} \widehat{I}_{i,j},$$

- Also works when $q_i = q$ for all $1 \le i \le p$ – Error exponent optimal!
- However, with non-identical noise, the Chow-Liu algorithm may not be able to recover the structure of the tree
- Note noisy correlation is $\tilde{\rho}_{i,j} = (1-2q_i)(1-2q_j)\rho_{i,j}$

Example:




$[\tilde{\rho}_{1,3} = \rho_{1,3} = 0.765]$

$[\tilde{\rho}_{1,2} = 0.72]$

$[\tilde{\rho}_{2,3} = 0.68]$

# Partial Tree Recovery under Non-Identical Noise

- Katiyar, Shah, and Caramanis [arXiv, Jun 2020] proposed an algorithm for partial tree structure recovery under non-identical noise for different nodes

## Robust Estimation of Tree Structured Ising Models

Ashish Katiyar
a.katiyar@utexas.edu

Vatsal Shah
vatsalshah1106@utexas.edu

Constantine Caramanis
constantine@utexas.edu

The University of Texas at Austin

**Abstract**

We consider the task of learning Ising models when the signs of different random variables are flipped independently with possibly unequal, unknown probabilities. In this paper, we focus on the problem of robust estimation of tree-structured Ising models. Without any additional assumption of side information, this is an open problem. We first prove that this problem is unidentifiable, however, this unidentifiability is limited to a small equivalence class of trees formed by leaf nodes exchanging positions with their neighbors. Next, we propose an algorithm to solve the above problem with logarithmic sample complexity in the number of nodes and polynomial run-time complexity. Lastly, we empirically demonstrate that, as expected, existing algorithms are not inherently robust in the proposed setting whereas our algorithm correctly recovers the underlying equivalence class.

- Extension of previous work for Gaussian tree models [ICML'19].

# Equivalent Tree-Structures

- Partial tree structure $\iff$ Trees in an equivalence class
- The equivalence relation is defined as follows:

$$\mathcal{T}_p = \text{set of trees on } p \text{ nodes}$$

$$\mathcal{L}_{\mathrm{T}} = \text{set of leaf nodes of } \mathrm{T}$$

$$\mathscr{S}_{\mathrm{T}} = \{\mathcal{S} \subset \mathcal{L}_{\mathrm{T}} : \text{no two nodes in } \mathcal{S} \text{ have the same neighbor}\}$$

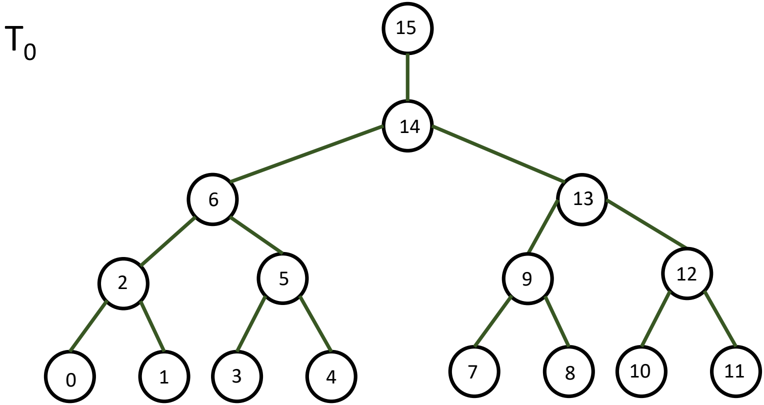
# Equivalent Tree-Structures

- Partial tree structure $\iff$ Trees in an equivalence class
- The equivalence relation is defined as follows:

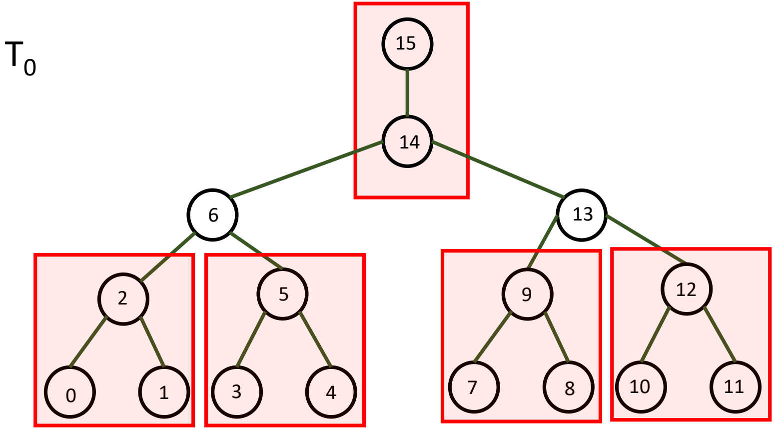
$$\mathcal{T}_p = \text{set of trees on } p \text{ nodes}$$
$$\mathcal{L}_{\mathrm{T}} = \text{set of leaf nodes of } \mathrm{T}$$
$$\mathscr{S}_{\mathrm{T}} = \{\mathcal{S} \subset \mathcal{L}_{\mathrm{T}} : \text{no two nodes in } \mathcal{S} \text{ have the same neighbor}\}$$

- For all $\mathcal{S} \in \mathscr{S}_{\mathrm{T}}$, let $\mathrm{T}_{\mathcal{S}}$ be the tree obtained by in interchanging the nodes in $\mathcal{S}$ with their corresponding neighbor node in $\mathrm{T}$
- $[\mathrm{T}] = \{\mathrm{T}_{\mathcal{S}} : \mathcal{S} \in \mathscr{S}_{\mathrm{T}}\}$ is our desired equivalence class.

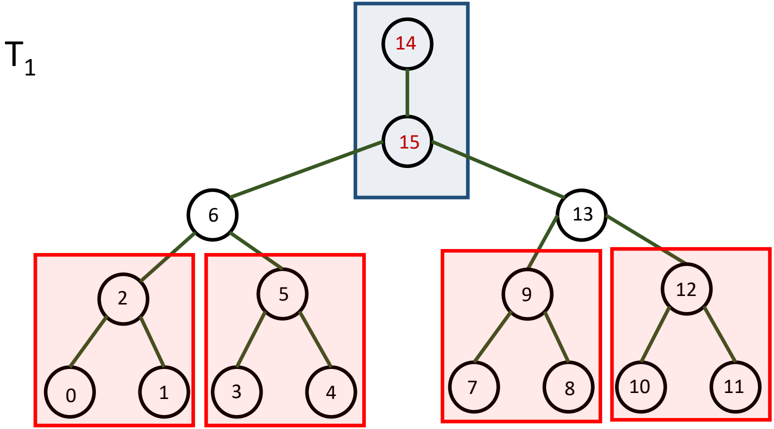
# Equivalent Tree-Structures: An Example



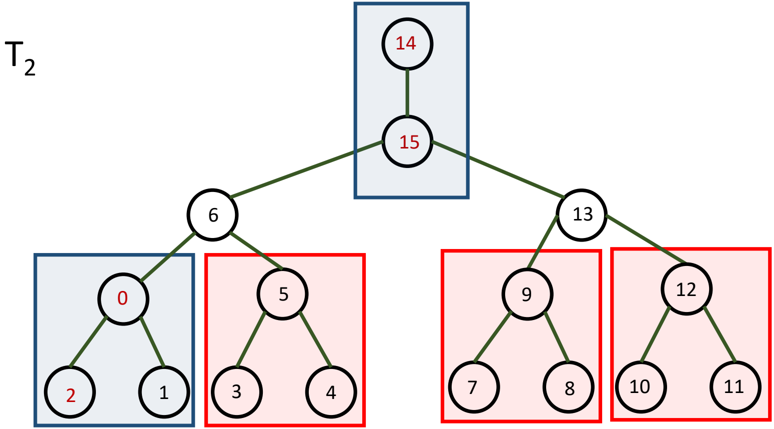
# Equivalent Tree-Structures: An Example



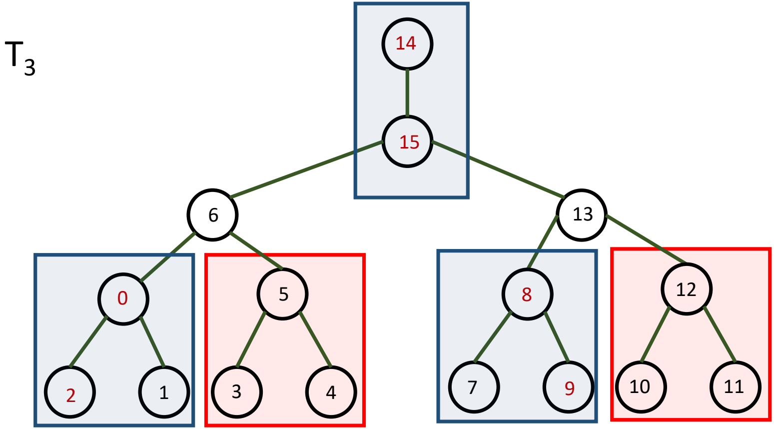
# Equivalent Tree-Structures: An Example



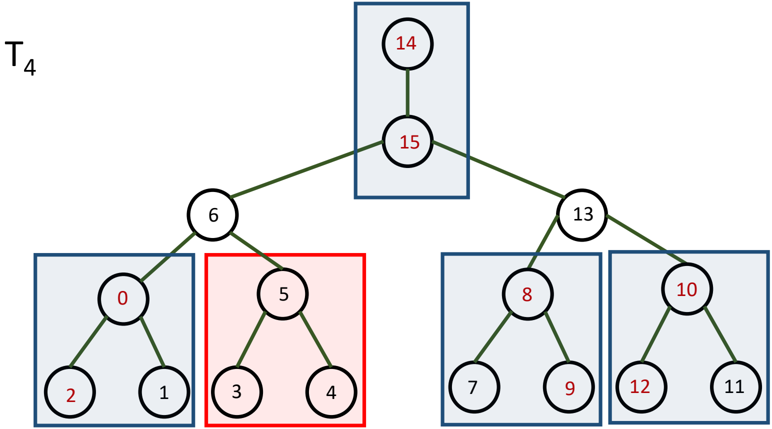
# Equivalent Tree-Structures: An Example



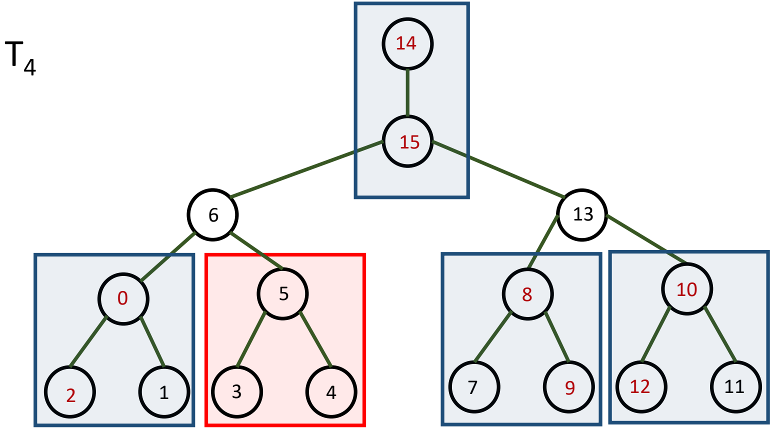
# Equivalent Tree-Structures: An Example



# Equivalent Tree-Structures: An Example



# Equivalent Tree-Structures: An Example




**Theorem 3 (Informal version of Katiyar, Shah, Caramanis (2020))**

*For arb. noise $\{q_i\}_{i=1}^{p}$, the "best one can do" is to learn trees up to $[\mathrm{T}]$.*

# Partial Tree Recovery under Non-Identical Noise

- Let the conditional independence among noiseless variables $X_1, \ldots, X_p$ be encoded by an unknown tree $\mathrm{T}$
- Let $\mathbf{y}_1^n = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$ denote $n$ independent noisy observations
- $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,p})$ with $y_{i,j}$ denoting the $i$th observation corresponding to the $j$th node, where $y_{i,j} \in \mathcal{Y} \triangleq \{+1, -1\}$

# Partial Tree Recovery under Non-Identical Noise

- Let the conditional independence among noiseless variables $X_1, \ldots, X_p$ be encoded by an unknown tree $\mathrm{T}$
- Let $\mathbf{y}_1^n = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$ denote $n$ independent noisy observations
- $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,p})$ with $y_{i,j}$ denoting the $i$th observation corresponding to the $j$th node, where $y_{i,j} \in \mathcal{Y} \triangleq \{+1, -1\}$
- Given $\mathbf{y}_1^n$, a learning algorithm (or estimator)

$$\Psi : \mathcal{Y}^{p \times n} \to \mathcal{T}_p$$

  provides an estimate of the underlying tree structure $\mathrm{T}$
- Noise statistics are completely unknown to the learning algorithm
- Interested in partial tree recovery (up to equivalence class $[\mathrm{T}]$), and an error is declared in the event

$$\text{Error} = \left\{\Psi(\mathbf{y}_1^n) \notin [\mathrm{T}]\right\}$$

# Algorithm for Partial Tree Structure Recovery

- Katiyar, Shah, and Caramanis presented an algorithm for partial tree structure recovery using $\mathbf{y}_1^n$ assuming

$$\text{(i) } 0 < \rho_{\min} \le |\rho_{i,j}| \le \rho_{\max} < 1 \qquad \text{(ii) } 0 \le q_i \le q_{\max} < 0.5$$

- Classification of any set of 4 distinct nodes as non-star or star

# Algorithm for Partial Tree Structure Recovery

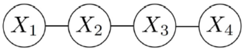
- Katiyar, Shah, and Caramanis presented an algorithm for partial tree structure recovery using $\mathbf{y}_1^n$ assuming

$$\text{(i) } 0 < \rho_{\min} \leq |\rho_{i,j}| \leq \rho_{\max} < 1 \qquad \text{(ii) } 0 \leq q_i \leq q_{\max} < 0.5$$

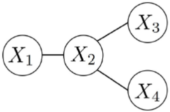
- Classification of any set of 4 distinct nodes as non-star or star

**Definition 4 (Non-star and star)**

- Any set of 4 distinct nodes forms a *non-star* if there exists 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*.




Non-star




Star

# Procedure for declaring Non-star or Star

Given noisy samples $\mathbf{y}_1^n$, algorithm calculates the empirical correlations

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

```
procedure IS-NON-STAR                ▷ Let the set of 4 nodes be {X_1, X_2, X_3, X_4}
    α = (1+ρ_max^2)/2
    if (ρ̂_{1,3} ρ̂_{2,4})/(ρ̂_{1,2} ρ̂_{3,4}) < α and (ρ̂_{1,3} ρ̂_{2,4})/(ρ̂_{1,4} ρ̂_{2,3}) > α then
        Declare Non-star where {X_1, X_2} forms a pair
    else if (ρ̂_{1,2} ρ̂_{3,4})/(ρ̂_{1,3} ρ̂_{2,4}) < α and (ρ̂_{1,2} ρ̂_{3,4})/(ρ̂_{1,4} ρ̂_{2,3}) > α then
        Declare Non-star where {X_1, X_3} forms a pair
    else if (ρ̂_{1,2} ρ̂_{3,4})/(ρ̂_{1,4} ρ̂_{2,3}) < α and (ρ̂_{1,2} ρ̂_{3,4})/(ρ̂_{1,3} ρ̂_{2,4}) > α then
        Declare Non-star where {X_1, X_4} forms a pair
    else
        Declare Star
    end if
end procedure
```

# Intuition behind the Non-star/Star procedure

- Consider 4 nodes that form a Markov-chain

$$X_1 — X_2 — X_3 — X_4$$

- If the noisy correlations are denoted $\tilde{\rho}_{i,j} \triangleq \mathbb{E}[Y_i Y_j]$, then we have

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

# Intuition behind the Non-star/Star procedure

- Consider 4 nodes that form a Markov-chain

$$X_1 \text{---} X_2 \text{---} X_3 \text{---} X_4$$

- If the noisy correlations are denoted $\tilde{\rho}_{i,j} \triangleq \mathbb{E}[Y_i Y_j]$, then we have

$$\frac{\tilde{\rho}_{1,3}\,\tilde{\rho}_{2,4}}{\tilde{\rho}_{1,2}\,\tilde{\rho}_{3,4}} \le \rho_{\max}^2, \quad \text{and} \quad \frac{\tilde{\rho}_{1,3}\,\tilde{\rho}_{2,4}}{\tilde{\rho}_{1,4}\,\tilde{\rho}_{2,3}} = 1$$

- Hence we would expect empirical correlations to satisfy

$$\frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} < \alpha \quad \text{and} \quad \frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}} > \alpha \quad \text{where} \quad \alpha = \frac{1+\rho_{\max}^2}{2}.$$

If all sets of 4 nodes are correctly declared as star or non-star (with appropriate pairing of nodes), then the equivalence class $[\mathrm{T}]$ is successfully detected, i.e., no error $\Psi(\mathbf{y}_1^n) \in [\mathrm{T}]$.

# Achievability Result by Katiyar, Shah, and Caramanis

**Theorem 5 (Katiyar, Shah, and Caramanis (Jun 2020))**

*The equivalence class* $[\mathrm{T}]$ *can be correctly recovered with probability at least* $1-\tau$ *if the number of samples satisfies*

$$n \geq \Omega\left(\frac{\log(p/\tau)}{(1-\rho_{\mathsf{max}})^2(1-2q_{\mathsf{max}})^{18}\rho_{\mathsf{min}}^{24}}\right).$$

# Achievability Result by Katiyar, Shah, and Caramanis

**Theorem 5 (Katiyar, Shah, and Caramanis (Jun 2020))**

*The equivalence class* $[\mathrm{T}]$ *can be correctly recovered with probability at least* $1-\tau$ *if the number of samples satisfies*

$$n \geq \Omega\left(\frac{\log(p/\tau)}{(1-\rho_{\max})^2(1-2q_{\max})^{18}\rho_{\min}^{24}}\right).$$

- As $q_{\max} \to (1/2)^-$, learning becomes more difficult because nodes suffer from too much noise.
- As $\rho_{\min} \to 0^+$, learning also becomes more difficult as minimum correlation is too small.

# Achievability Result by Katiyar, Shah, and Caramanis

**Theorem 5 (Katiyar, Shah, and Caramanis (Jun 2020))**

*The equivalence class* $[\mathrm{T}]$ *can be correctly recovered with probability at least* $1-\tau$ *if the number of samples satisfies*

$$n \geq \Omega\left(\frac{\log(p/\tau)}{(1-\rho_{\max})^2(1-2q_{\max})^{18}\rho_{\min}^{24}}\right).$$

- As $q_{\max} \to (1/2)^-$, learning becomes more difficult because nodes suffer from too much noise.
- As $\rho_{\min} \to 0^+$, learning also becomes more difficult as minimum correlation is too small.
- Polynomial orders (18, 24) very large! Can we improve?

# Our Contributions

- Significantly improved analysis of algorithm by Katiyar, Shah, and Caramanis (KSC)
- Significantly improved algorithm – Symmetrized Geometric Averaging (SGA)
  - Provable improvement of sample complexity vis-à-vis KSC's algorithm via error exponents
  - Much superior experimental results
  - Applicable to Gaussian graphical models
- Novel impossibility result in terms on the minimax error probability

# Improved Achievability Result

**Theorem 6 (Tandon, Han and T., Jan 2021)**

*Using KSC's algorithm, the equivalence class* $[\mathrm{T}]$ *can be correctly recovered with probability at least* $1-\tau$ *if the number of samples satisfy*

$$n \geq \Omega\left(\frac{\log(p/\tau)}{(1-\rho_{\max})^2(1-2q_{\max})^{6}\rho_{\min}^{8}}\right).$$

# Improved Achievability Result

**Theorem 6 (Tandon, Han and T., Jan 2021)**

*Using KSC's algorithm, the equivalence class* $[\mathrm{T}]$ *can be correctly recovered with probability at least* $1-\tau$ *if the number of samples satisfy*

$$n \geq \Omega\left(\frac{\log(p/\tau)}{(1-\rho_{\max})^2(1-2q_{\max})^{6}\rho_{\min}^{8}}\right).$$

- Refined probability bounds for events such as

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

- Polynomial orders $(6, 8)$ still very large. Can we do better?

# Improved Achievability Result

**Theorem 6 (Tandon, Han and T., Jan 2021)**

*Using KSC's algorithm, the equivalence class* $[\mathrm{T}]$ *can be correctly recovered with probability at least* $1-\tau$ *if the number of samples satisfy*

$$n \geq \Omega\left(\frac{\log(p/\tau)}{(1-\rho_{\max})^2(1-2q_{\max})^6\rho_{\min}^8}\right).$$

- Refined probability bounds for events such as

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

- Polynomial orders $(6, 8)$ still very large. Can we do better?
- Yes: We can get a better algorithm.
- No: Cannot reduce the polynomial orders analytically but can provide a distribution-dependent bound via error exponents.

# SGA Procedure to declare Non-star or Star

**procedure** $\textsc{SGA-Is-Non-star}$ $\triangleright$ The 4 nodes are $\{X_1, X_2, X_3, X_4\}$

$\alpha = (1 + \rho^2_{\mathsf{max}})/2$

$v_2 = \frac{\sqrt{|\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}\,\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}|}}{|\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}|}, \quad v_3 = \frac{\sqrt{|\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}\,\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}|}}{|\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}|}, \quad v_4 = \frac{\sqrt{|\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}\,\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}|}}{|\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}|}$

Let $v = \min_{2 \le i \le 4} v_i$ and $i^* = \arg\min_{2 \le i \le 4} v_i$

**if** $v < \alpha$ **then**

Declare Non-star where $\{X_1, X_{i^*}\}$ forms a pair

**else**

Declare Star

**end if**

**end procedure**

# SGA Procedure to declare Non-star or Star

**procedure** SGA-IS-NON-STAR ▷ The 4 nodes are $\{X_1, X_2, X_3, X_4\}$

$\alpha = (1 + \rho_{\max}^2)/2$

$v_2 = \frac{\sqrt{|\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}\,\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}|}}{|\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}|}, \quad v_3 = \frac{\sqrt{|\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}\,\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}|}}{|\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}|}, \quad v_4 = \frac{\sqrt{|\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}\,\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}|}}{|\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}|}$

Let $v = \min_{2 \le i \le 4} v_i$ and $i^* = \arg\min_{2 \le i \le 4} v_i$

**if** $v < \alpha$ **then**

Declare Non-star where $\{X_1, X_{i^*}\}$ forms a pair

**else**

Declare Star

**end if**

**end procedure**

Advantages of SGA over KSC's procedure

- Symmetry: Invariant to permutation of the node indices
- Averaging: Takes the Geometric Mean of the empirical statistics

$$\sqrt{\left|\frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}}\right| \cdot \left|\frac{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}}\right|} = \frac{\sqrt{|\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}\,\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}|}}{|\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}|}$$

# Intuition behind the SGA procedure

- Intuition behind SGA can be highlighted by an example where $\{X_1, X_2, X_3, X_4\}$ forms a non-star with pair $\{X_1, X_2\}$
  - If $\tilde{\rho}_{i,j} \triangleq \mathbb{E}[Y_i Y_j]$, then

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

# Intuition behind the SGA procedure

- Intuition behind SGA can be highlighted by an example where $\{X_1, X_2, X_3, X_4\}$ forms a non-star with pair $\{X_1, X_2\}$
  - If $\tilde{\rho}_{i,j} \triangleq \mathbb{E}[Y_i Y_j]$, then

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

  - Hence, we would expect the following metrics, based on empirical correlations, to satisfy

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

  - KSC checks (i) but ignores (ii)
  - SGA compares the geometric average of the metrics in (i) and (ii) against the threshold $\alpha$

# Intuition behind the SGA procedure

- Intuition behind SGA can be highlighted by an example where $\{X_1, X_2, X_3, X_4\}$ forms a non-star with pair $\{X_1, X_2\}$
  - If $\tilde{\rho}_{i,j} \triangleq \mathbb{E}[Y_i Y_j]$, then

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

  - Hence, we would expect the following metrics, based on empirical correlations, to satisfy

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

  - KSC checks (i) but ignores (ii)
  - SGA compares the geometric average of the metrics in (i) and (ii) against the threshold $\alpha$
- Folklore theorem: "Averaging cannot hurt and generally helps"

# Katiyar's Algorithm: Error Exponent for Chains

- Consider a 4-node Markov chain structure

$$X_1 \text{---} X_2 \text{---} X_3 \text{---} X_4,$$

and let $\tilde{P}$ denote the joint distribution of the noisy samples

# Katiyar's Algorithm: Error Exponent for Chains

- Consider a 4-node Markov chain structure

$$X_1 — X_2 — X_3 — X_4,$$

and let $\tilde{P}$ denote the joint distribution of the noisy samples

- Two events that lead to error using Katiyar's algorithm are

$$\mathscr{E}_1 = \left\{ \frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} \geq \alpha \right\} \quad \text{and} \quad \mathscr{E}_2 = \left\{ \frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}} \leq \alpha \right\}$$

  - Using Sanov's theorem, we have
    $e_1 = \min_{Q\in\mathcal{P}(\mathcal{Y}^4)} \left\{ D(Q\|\tilde{P}) \,:\, \frac{\rho_{1,3}^{(Q)}\,\rho_{2,4}^{(Q)}}{\rho_{1,2}^{(Q)}\,\rho_{3,4}^{(Q)}} \geq \alpha \right\}$, where $\rho_{i,j}^{(Q)} \triangleq \mathbb{E}_Q[Y_i Y_j]$
  - $e_2 = \min_{Q\in\mathcal{P}(\mathcal{Y}^4)} \left\{ D(Q\|\tilde{P}) \,:\, \frac{\rho_{1,3}^{(Q)}\,\rho_{2,4}^{(Q)}}{\rho_{1,4}^{(Q)}\,\rho_{2,3}^{(Q)}} \leq \alpha \right\}$

- Consider a 4-node Markov chain structure

$$X_1 - X_2 - X_3 - X_4,$$

and let $\tilde{P}$ denote the joint distribution of the noisy samples

- Two events that lead to error using Katiyar's algorithm are

$$\mathscr{E}_1 = \left\{ \frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,2}\,\widehat{\rho}_{3,4}} \ge \alpha \right\} \quad \text{and} \quad \mathscr{E}_2 = \left\{ \frac{\widehat{\rho}_{1,3}\,\widehat{\rho}_{2,4}}{\widehat{\rho}_{1,4}\,\widehat{\rho}_{2,3}} \le \alpha \right\}$$

  - Using Sanov's theorem, we have
    $e_1 = \min_{Q \in \mathcal{P}(\mathcal{Y}^4)} \left\{ D(Q \| \tilde{P}) \,:\, \frac{\rho_{1,3}^{(Q)}\,\rho_{2,4}^{(Q)}}{\rho_{1,2}^{(Q)}\,\rho_{3,4}^{(Q)}} \ge \alpha \right\}$, where $\rho_{i,j}^{(Q)} \triangleq \mathbb{E}_Q[Y_i Y_j]$
  - $e_2 = \min_{Q \in \mathcal{P}(\mathcal{Y}^4)} \left\{ D(Q \| \tilde{P}) \,:\, \frac{\rho_{1,3}^{(Q)}\,\rho_{2,4}^{(Q)}}{\rho_{1,4}^{(Q)}\,\rho_{2,3}^{(Q)}} \le \alpha \right\}$

- The overall error exponent for Katiyar's algorithm is given by

$$E(\Psi_{\mathrm{KA}}, \tilde{P}) \triangleq \lim_{n \to \infty} -\frac{1}{n} \log \Pr\left(\mathscr{E}_1 \cup \mathscr{E}_2\right) = \min\{e_1, e_2\}$$

# SGA Algorithm: Error Exponent for Chains

- Error events using SGA:

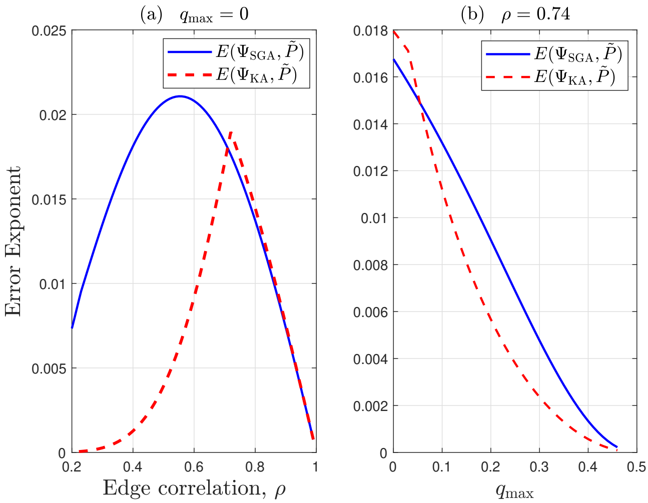
$$\mathscr{E}_3 = \left\{ \frac{\sqrt{|\widehat{\rho}_{1,3}\, \widehat{\rho}_{2,4}\, \widehat{\rho}_{1,4}\, \widehat{\rho}_{2,3}|}}{|\widehat{\rho}_{1,2}\, \widehat{\rho}_{3,4}|} \geq \alpha \right\}$$

$$\mathscr{E}_4 = \{|\widehat{\rho}_{1,3}\, \widehat{\rho}_{2,4}| \geq |\widehat{\rho}_{1,2}\, \widehat{\rho}_{3,4}|\}, \quad \text{and} \quad \mathscr{E}_5 = \{|\widehat{\rho}_{1,4}\, \widehat{\rho}_{2,3}| \geq |\widehat{\rho}_{1,2}\, \widehat{\rho}_{3,4}|\}.$$

- The corresponding error exponents are given by:
  - $e_3 = \min_{Q \in \mathcal{P}(\mathcal{Y}^4)} \left\{ D(Q\|\tilde{P}) \,:\, \frac{\sqrt{|\rho_{1,3}^{(Q)}\, \rho_{2,4}^{(Q)}\, \rho_{1,4}^{(Q)}\, \rho_{2,3}^{(Q)}|}}{|\rho_{1,2}^{(Q)}\, \rho_{3,4}^{(Q)}|} \geq \alpha \right\}$
  - $e_4 = \min_{Q \in \mathcal{P}(\mathcal{Y}^4)} \left\{ D(Q\|\tilde{P}) \,:\, |\rho_{1,3}^{(Q)}\, \rho_{2,4}^{(Q)}| \geq |\rho_{1,2}^{(Q)}\, \rho_{3,4}^{(Q)}| \right\}$
  - $e_5 = \min_{Q \in \mathcal{P}(\mathcal{Y}^4)} \left\{ D(Q\|\tilde{P}) \,:\, |\rho_{1,4}^{(Q)}\, \rho_{2,3}^{(Q)}| \geq |\rho_{1,2}^{(Q)}\, \rho_{3,4}^{(Q)}| \right\}$
- The overall error exponent using SGA algorithm is given by

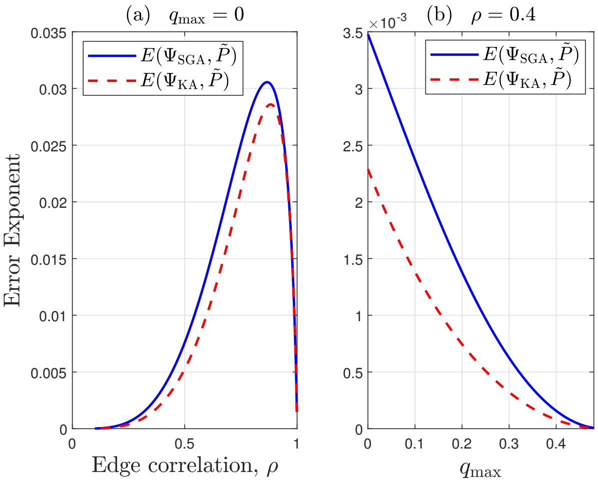
$$E(\Psi_{\mathrm{SGA}}, \tilde{P}) \triangleq \lim_{n\to\infty} -\frac{1}{n} \log \Pr\left(\mathscr{E}_3 \cup \mathscr{E}_4 \cup \mathscr{E}_5\right) = \min\{e_3, e_4, e_5\}$$

# Numerical Results: Error Exponent for a 4 node chain




Error exponents for a 4-node homogeneous chain with edge correlation $\rho$

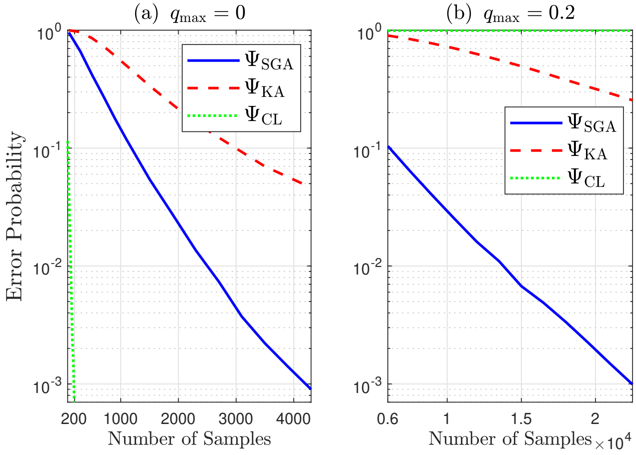
# Numerical Results: Error Exponent for a 4 node star




Error exponents for a 4-node (homogeneous) star with edge correlation $\rho$

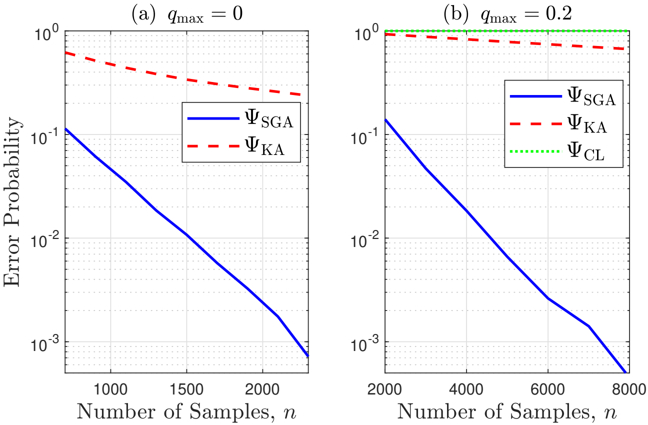
# Simulation Results: 12-node chain tree structure

Edge correlation $\rho = 0.6$



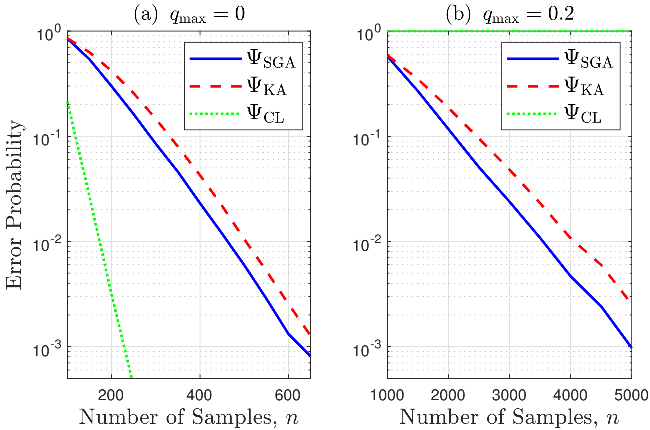
# Simulation Results: 12-node hybrid tree structure

Edge correlation $\rho = 0.6$



# Simulation Results: 12-node star tree structure

Edge correlation $\rho = 0.6$



# Extension to Gaussian tree models

- Observe $Y_i = X_i + N_i$, where noise $N_i \sim \mathcal{N}(0, \sigma_i^2)$ for some $\sigma_i > 0$.
- Experiment Setup:
  - Choose a tree structure $\mathrm{T}_P = (\mathcal{V}, \mathcal{E}_P)$ with $p = 10$ nodes
  - Generate the inverse covariance matrix $(\mathbf{\Sigma}^*)^{-1}$ as follows

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

for some parameter $w \in \mathbb{R}$. Invert $(\mathbf{\Sigma}^*)^{-1}$ to obtain $\mathbf{\Sigma}^*$

# Extension to Gaussian tree models

- Observe $Y_i = X_i + N_i$, where noise $N_i \sim \mathcal{N}(0, \sigma_i^2)$ for some $\sigma_i > 0$.
- Experiment Setup:
  - Choose a tree structure $\mathrm{T}_P = (\mathcal{V}, \mathcal{E}_P)$ with $p = 10$ nodes
  - Generate the inverse covariance matrix $(\boldsymbol{\Sigma}^*)^{-1}$ as follows

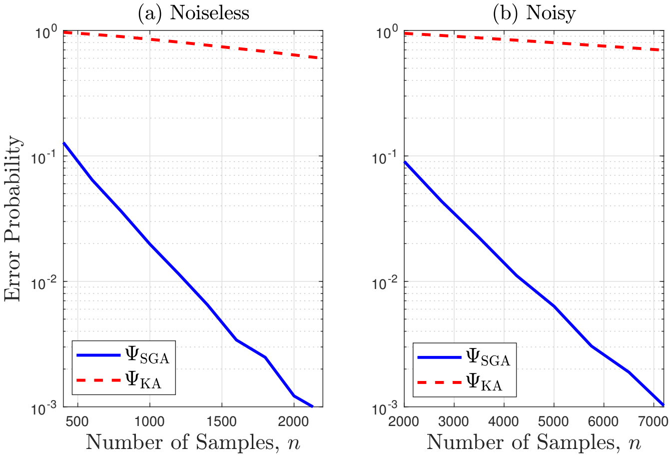
$$[(\boldsymbol{\Sigma}^*)^{-1}]_{i,j} = \begin{cases} w, & \text{if } \{i,j\} \in \mathcal{E}_P; \\ 1, & \text{if } i = j; \\ 0 & \text{otherwise} \end{cases}$$

    for some parameter $w \in \mathbb{R}$. Invert $(\boldsymbol{\Sigma}^*)^{-1}$ to obtain $\boldsymbol{\Sigma}^*$
  - The correlation matrix $\mathbf{K}^*$ is calculated from $\boldsymbol{\Sigma}^*$ using the formula

$$\mathbf{K}^* = (\mathrm{diag}(\boldsymbol{\Sigma}^*))^{-\frac{1}{2}} \boldsymbol{\Sigma}^* (\mathrm{diag}(\boldsymbol{\Sigma}^*))^{-\frac{1}{2}}$$

  - For the noisy case, $[\mathbf{D}^*]_{i,i} = 2$ for $i \in \{1, 3, 5, 7, 9\}$
  - Generate samples from distribution $\mathcal{N}(0, \boldsymbol{\Sigma}^* + \mathbf{D}^*)$

## Gaussian Results: 10-node hybrid tree with $w = 0.38$



# Converse: Number of necessary samples

- For a given tree $\mathrm{T} = (\mathcal{V}, \mathcal{E})$, let $\mathcal{P}_{\mathrm{T}}(\rho_{\mathsf{min}}, \rho_{\mathsf{max}})$ denote the set of all tree structured Ising models that satisfy

$$0 < \rho_{\mathsf{min}} \le |\rho_{i,j}| \le \rho_{\mathsf{max}} < 1 \qquad \forall \{i,j\} \in \mathcal{E}$$

- For a given tree $\mathrm{T} = (\mathcal{V}, \mathcal{E})$, let $\mathcal{P}_{\mathrm{T}}(\rho_{\min}, \rho_{\max})$ denote the set of all tree structured Ising models that satisfy

$$0 < \rho_{\min} \leq |\rho_{i,j}| \leq \rho_{\max} < 1 \qquad \forall \{i,j\} \in \mathcal{E}$$

- The minimax error probability for partial tree structure recovery up to equivalence class $[\mathrm{T}]$ is

$$\mathcal{M}_n(q_{\max}, \rho_{\min}, \rho_{\max}) \triangleq \inf_{\Psi} \sup_{\substack{\mathrm{T} \in \mathcal{T}_p, \\ P \in \mathcal{P}_{\mathrm{T}}(\rho_{\min}, \rho_{\max}), \\ 0 \leq q_i \leq q_{\max} < 0.5}} \mathbb{P}_P\big(\Psi(\mathbf{Y}_1^n) \notin [\mathrm{T}]\big)$$

where $\mathbb{P}_P(\cdot)$ denotes the probability when tree distribution is $P$, and noise crossover probabilities are given by $\{q_i\}_{i=1}^p$.

# Impossibility Result/Necessary Number of Samples

**Theorem 7 (Tandon, Han and T., Jan 2021)**

*Let* $\rho_q \triangleq (1 - 2q_{\max})\rho_{\min}$. *If* $p > 32$, *and the number of samples* $n$ *satisfy*

$$n < \frac{\log p}{4\,(1 - \rho_{\max})\,\rho_q\,\tanh^{-1}(\rho_q)}$$

*then we have* $\mathcal{M}_n(q_{\max}, \rho_{\min}, \rho_{\max}) \geq 1/2$.

# Impossibility Result/Necessary Number of Samples

## Theorem 7 (Tandon, Han and T., Jan 2021)

*Let* $\rho_q \triangleq (1-2q_{\max})\rho_{\min}$. *If* $p > 32$, *and the number of samples* $n$ *satisfy*

$$n < \frac{\log p}{4\,(1-\rho_{\max})\,\rho_q\,\tanh^{-1}(\rho_q)}$$

*then we have* $\mathcal{M}_n(q_{\max}, \rho_{\min}, \rho_{\max}) \geq 1/2$.

*In other words, the optimal sample complexity*

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

# Impossibility Result/Necessary Number of Samples

## Theorem 7 (Tandon, Han and T., Jan 2021)

*Let* $\rho_q \triangleq (1 - 2q_{\max})\rho_{\min}$. *If* $p > 32$, *and the number of samples* $n$ *satisfy*

$$n < \frac{\log p}{4\,(1-\rho_{\max})\,\rho_q\ \tanh^{-1}(\rho_q)}$$

*then we have* $\mathcal{M}_n(q_{\max}, \rho_{\min}, \rho_{\max}) \geq 1/2$.

*In other words, the optimal sample complexity*

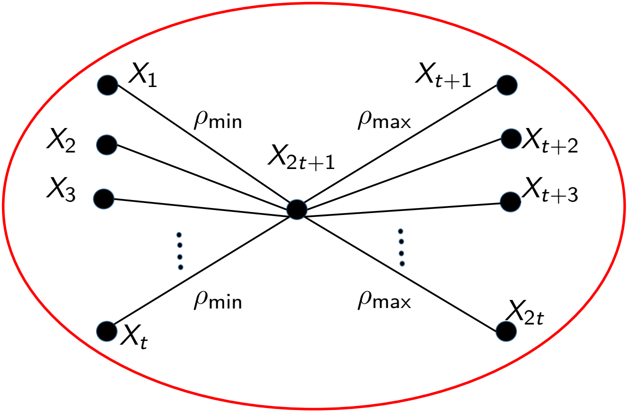
$$n^*(\rho_{\min}, \rho_{\max}, q_{\max}) = \Omega\left(\frac{\log p}{(1-\rho_{\max})^{1}(1-2q_{\max})^{2}\rho_{\min}^{2}}\right).$$

Compare to improved analysis of Katiyar *et al.*'s algorithm and SGA:

$$n^*(\rho_{\min}, \rho_{\max}, q_{\max}) = O\left(\frac{\log(p/\tau)}{(1-\rho_{\max})^{2}(1-2q_{\max})^{6}\rho_{\min}^{8}}\right).$$

# Ingredients in Impossibility Proof

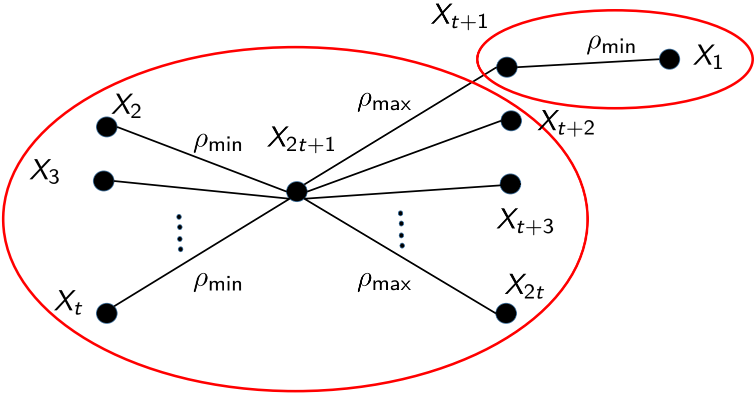
Choice of large $t^2 = O(p^2)$ number of trees, close to each other and respective equivalence classes are disjoint for an $t^2$-ary hypothesis test.




Tree structure $\mathrm{T}_0$

# Ingredients in Impossibility Proof

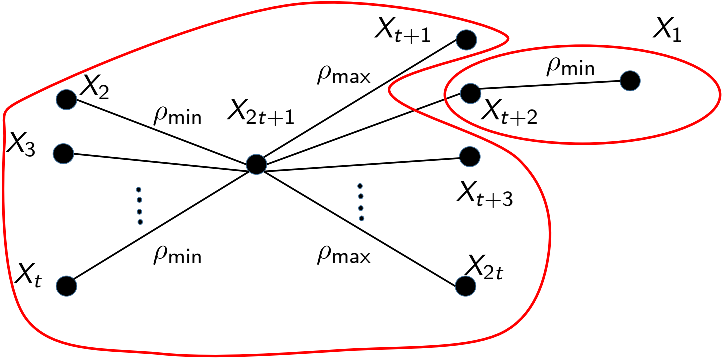
Choice of large $t^2 = O(p^2)$ number of trees, close to each other and respective equivalence classes are disjoint for an $t^2$-ary hypothesis test.




Tree structure $\mathrm{T}_1$

# Ingredients in Impossibility Proof

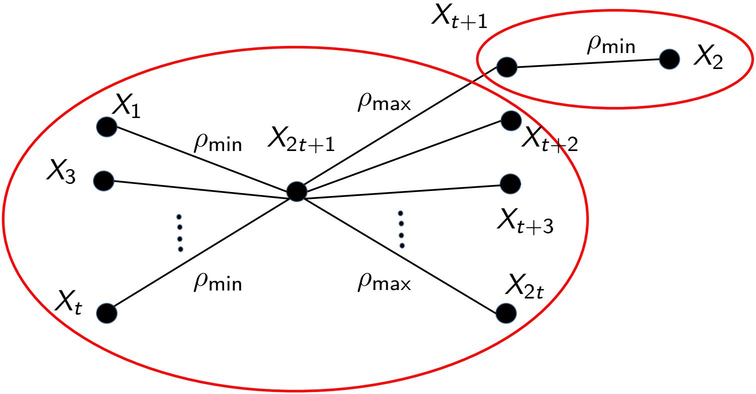
Choice of large $t^2 = O(p^2)$ number of trees, close to each other and respective equivalence classes are disjoint for an $t^2$-ary hypothesis test.




Tree structure $\mathrm{T}_2$

# Ingredients in Impossibility Proof

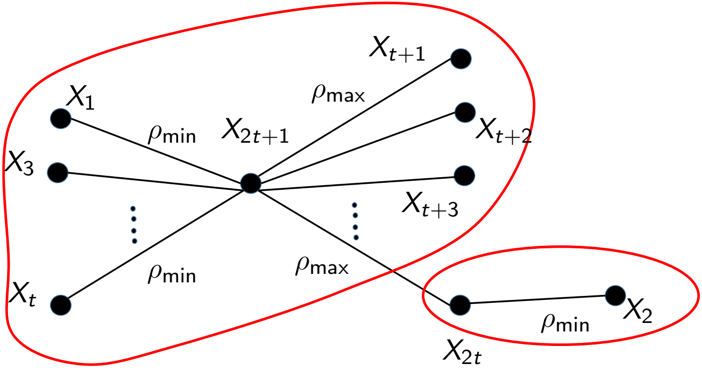
Choice of large $t^2 = O(p^2)$ number of trees, close to each other and respective equivalence classes are disjoint for an $t^2$-ary hypothesis test.




Tree structure $\mathrm{T}_{t+1}$ where $p = 2t + 1$

# Ingredients in Impossibility Proof

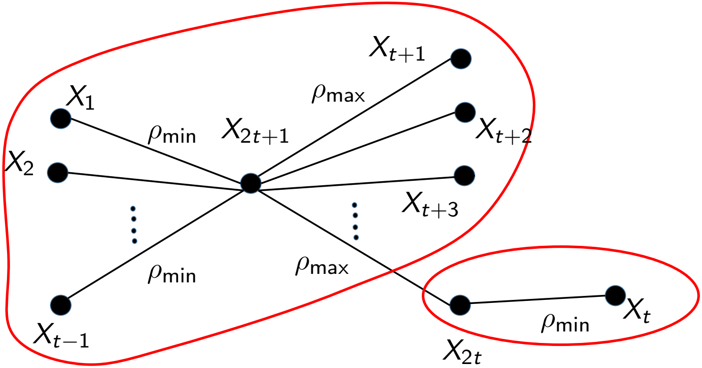
Choice of large $t^2 = O(p^2)$ number of trees, close to each other and respective equivalence classes are disjoint for an $t^2$-ary hypothesis test.




Tree structure $\mathrm{T}_{2t}$

# Ingredients in Impossibility Proof

Choice of large $t^2 = O(p^2)$ number of trees, close to each other and respective equivalence classes are disjoint for an $t^2$-ary hypothesis test.




Tree structure $\mathrm{T}_{t^2}$

# Converse Result: Discussion

- Nikolakakis *et al.* [AISTATS-2019] presented a bound on the necessary number of samples required for exact tree structure recovery for identical noise setting (i.e., $q_i = q, \forall i$)

# Converse Result: Discussion

- Nikolakakis *et al.* [AISTATS-2019] presented a bound on the necessary number of samples required for exact tree structure recovery for identical noise setting (i.e., $q_i = q, \forall i$)

- Impact of noise gets manifested as a multiplicative factor

$$[1 - (4q(1-q))^p]^{-1}.$$

Implies that if $q \neq 0, 1$ impact of noise becomes negligible because

$$\lim_{p \to \infty} [1 - (4q(1-q))^p]^{-1} = 1.$$

# Converse Result: Discussion

- Nikolakakis *et al.* [AISTATS-2019] presented a bound on the necessary number of samples required for exact tree structure recovery for identical noise setting (i.e., $q_i = q, \forall i$)

- Impact of noise gets manifested as a multiplicative factor

$$[1-(4q(1-q))^p]^{-1}.$$

  Implies that if $q \neq 0, 1$ impact of noise becomes negligible because

$$\lim_{p\to\infty}[1-(4q(1-q))^p]^{-1} = 1.$$

- In contrast, our result for non-identical noise

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

  shows that the necessary $n$ for $q_{\max} > 0$ is greater that for the noiseless setting by a factor of at least $(1-2q_{\max})^{-2}$, regardless of $p$.

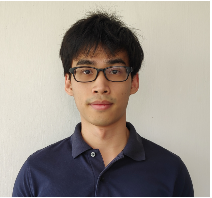
# Conclusions

- Improved sufficient sample complexity result compared to Katiyar, Shah and Caramanis (2020)
- Presented a modified procedure SGA for partial tree recovery
- Improved error exponents and numerical results for both discrete and Gaussian graphical models
- Novel converse result for partial tree structure recovery up to equivalence class under non-identical noise

# References and Acknowledgements


- A. Tandon, V. Y. F. Tan and S. Zhu, "Exact Asymptotics for Learning Tree-Structured Graphical Models: Noiseless and Noisy Samples", *IEEE Journal of Selected Areas in Inform. Th.*, Nov 2020
- A. Tandon, A. J. Y. Han and V. Y. F. Tan, "SGA: A Robust Algorithm for Partial Recovery of Tree-Structured Graphical Models with Noisy Samples", arXiv 2101.08917


Anshoo Tandon (ECE)



Aldric Han (Maths)

Shiyao Zhu (ECE)