Learning Gaussian Tree Models: Analysis of Error Exponents and Extremal Structures

Vincent Y. F. Tan, *Student Member, IEEE*, Animashree Anandkumar, *Member, IEEE*, and
Alan S. Willsky, *Fellow, IEEE*

Abstract—The problem of learning tree-structured Gaussian graphical models from independent and identically distributed (i.i.d.) samples is considered. The influence of the tree structure and the parameters of the Gaussian distribution on the learning rate as the number of samples increases is discussed. Specifically, the error exponent corresponding to the event that the estimated tree structure differs from the actual unknown tree structure of the distribution is analyzed. Finding the error exponent reduces to a least-squares problem in the very noisy learning regime. In this regime, it is shown that the extremal tree structure that minimizes the error exponent is the star for any fixed set of correlation coefficients on the edges of the tree. If the magnitudes of all the correlation coefficients are less than 0.63, it is also shown that the tree structure that maximizes the error exponent is the Markov chain. In other words, the star and the chain graphs represent the hardest and the easiest structures to learn in the class of tree-structured Gaussian graphical models. This result can also be intuitively explained by correlation decay: pairs of nodes which are far apart, in terms of graph distance, are unlikely to be mistaken as edges by the maximum-likelihood estimator in the asymptotic regime.

Index Terms—Error exponents, Euclidean information theory, Gauss-Markov random fields, Gaussian graphical models, large deviations, structure learning, tree distributions.

I. INTRODUCTION

LEARNING of structure and interdependencies of a large collection of random variables from a set of data samples is an important task in signal and image analysis and many other scientific domains (see examples in [1]–[4] and references therein). This task is extremely challenging when the dimensionality of the data is large compared to the number of samples. Furthermore, structure learning of multivariate distributions is also complicated as it is imperative to find the right balance between data fidelity and overfitting the data to the model. This problem is circumvented when we limit the distributions to the set of Markov tree distributions, which have a fixed number of parameters and are tractable for learning [5] and statistical inference [1], [4].

The problem of maximum-likelihood (ML) learning of a Markov tree distribution from i.i.d. samples has an elegant solution, proposed by Chow and Liu in [5]. The ML tree structure is given by the maximum-weight spanning tree (MWST) with empirical mutual information quantities as the edge weights. Furthermore, the ML algorithm is *consistent* [6], which implies that the error probability in learning the tree structure decays to zero with the number of samples available for learning.

While consistency is an important qualitative property, there is substantial motivation for additional and more quantitative characterization of performance. One such measure, which we investigate in this theoretical paper is the rate of decay of the error probability, i.e., the probability that the ML estimate of the edge set differs from the true edge set. When the error probability decays exponentially, the learning rate is usually referred to as the *error exponent*, which provides a careful measure of performance of the learning algorithm since a larger rate implies a faster decay of the error probability.

We answer three fundamental questions in this paper: i) Can we characterize the error exponent for structure learning by the ML algorithm for tree-structured Gaussian graphical models (also called Gauss-Markov random fields)? ii) How do the *structure* and *parameters* of the model influence the error exponent? iii) What are extremal tree distributions for learning, i.e., the distributions that maximize and minimize the error exponents? We believe that our intuitively appealing answers to these important questions provide key insights for learning tree-structured Gaussian graphical models from data, and thus, for modeling high-dimensional data using parameterized tree-structured distributions.

A. Summary of Main Results

We derive the error exponent as the optimal value of the objective function of a nonconvex optimization problem, which can only be solved numerically (Theorem 2). To gain better insights into when errors occur, we approximate the error exponent with a closed-form expression that can be interpreted as the signal-to-noise ratio (SNR) for structure learning (Theorem 4), thus showing how the parameters of the true model affect learning. Furthermore, we show that due to *correlation decay*, pairs of nodes which are far apart, in terms of their graph distance, are unlikely to be mistaken as edges by the ML estimator. This is not only an intuitive result, but also results in a significant reduction in the computational complexity to find the exponent—from $\mathcal{O}(d^{d-2})$ for exhaustive search and $\mathcal{O}(d^3)$ for discrete tree models [7] to $\mathcal{O}(d)$ for Gaussians (Proposition 7), where d is the number of nodes.

We then analyze extremal tree structures for learning, given a fixed set of correlation coefficients on the edges of the tree. Our main result is the following: The *star* graph minimizes the error exponent and if the absolute value of all the correlation coefficients of the variables along the edges is less than 0.63, then the *Markov chain* also maximizes the error exponent (Theorem 8). Therefore, the extremal tree structures in terms of the diameter are *also* extremal trees for learning Gaussian tree distributions. This agrees with the intuition that the amount of correlation decay increases with the tree diameter, and that correlation decay helps the ML estimator to better distinguish the edges from the nonneighbor pairs. Lastly, we analyze how changing the size of the tree influences the magnitude of the error exponent (Propositions 11 and 12).

B. Related Work

There is a substantial body of work on approximate learning of graphical models (also known as Markov random fields) from data e.g., [8]–[11]. The authors of these papers use various score-based approaches [8], the maximum entropy principle [9] or ℓ_1 regularization [10], [11] as approximate structure learning techniques. Consistency guarantees in terms of the number of samples, the number of variables and the maximum neighborhood size are provided. Information-theoretic limits [12] for learning graphical models have also been derived. In [13], bounds on the error rate for learning the structure of Bayesian networks were provided but in contrast to our work, these bounds are not asymptotically tight (cf. Theorem 2). Furthermore, the analysis in [13] is tied to the Bayesian Information Criterion. The focus of our paper is the analysis of the Chow-Liu [5] algorithm as an *exact* learning technique for estimating the tree structure and comparing error rates amongst different graphical models. In a recent paper [14], the authors concluded that if the graphical model possesses long range correlations, then it is difficult to learn. In this paper, we in fact identify the extremal structures and distributions in terms of error exponents for structure learning. The area of study in statistics known as *covariance selection* [15], [16] also has connections with structure learning in Gaussian graphical models. Covariance selection involves estimating the nonzero elements in the inverse covariance matrix and providing consistency guarantees of the estimate in some norm, e.g., the Frobenius norm in [17].

We previously analyzed the error exponent for learning discrete tree distributions in [7]. We proved that for every discrete spanning tree model, the error exponent for learning is strictly positive, which implies that the error probability decays exponentially fast. In this paper, we extend these results to Gaussian tree models and derive new results which are both explicit and intuitive by exploiting the properties of Gaussians. The results we obtain in Sections III and IV are analogous to the results in [7] obtained for discrete distributions, although the proof techniques are different. Sections V and VI contain new results thanks to simplifications which hold for Gaussians but which do not hold for discrete distributions.

C. Paper Outline

This paper is organized as follows: In Section II, we state the problem precisely and provide necessary preliminaries on

learning Gaussian tree models. In Section III, we derive an expression for the so-called crossover rate of two pairs of nodes. We then relate the set of crossover rates to the error exponent for learning the tree structure. In Section IV, we leverage on ideas from Euclidean information theory [18] to state conditions that allow accurate approximations of the error exponent. We demonstrate in Section V how to reduce the computational complexity for calculating the exponent. In Section VI, we identify extremal structures that maximize and minimize the error exponent. Numerical results are presented in Section VII and we conclude the discussion in Section VIII.

II. PRELIMINARIES AND PROBLEM STATEMENT

A. Basics of Undirected Gaussian Graphical Models

Undirected graphical models or *Markov random fields*[1] (MRFs) are probability distributions that factorize according to given undirected graphs [3]. In this paper, we focus solely on *spanning trees* (i.e., undirected, acyclic, connected graphs). A d-dimensional random vector $\mathbf{x} = [x_1, \ldots, x_d]^T \in \mathbb{R}^d$ is said to be *Markov* on a spanning tree $T_p = (\mathcal{V}, \mathcal{E}_p)$ with vertex (or node) set $\mathcal{V} = \{1, \ldots, d\}$ and edge set $\mathcal{E}_p \subset \binom{\mathcal{V}}{2}$ if its distribution $p(\mathbf{x})$ satisfies the (local) Markov property: $p(x_i | x_{\mathcal{V} \setminus \{i\}}) = p(x_i | x_{\mathrm{nbd}(i)})$, where $\mathrm{nbd}(i) := \{j \in \mathcal{V} : (i, j) \in \mathcal{E}_p\}$ denotes the set of neighbors of node i. We also denote the set of spanning trees with d nodes as $\mathcal{T}^d$, thus $T_p \in \mathcal{T}^d$. Since p is Markov on the tree T_p, its probability density function (pdf) factorizes according to T_p into node marginals $\{p_i : i \in \mathcal{V}\}$ and pairwise marginals $\{p_{i,j} : (i, j) \in \mathcal{E}_p\}$ in the following specific way [3] given the edge set $\mathcal{E}_p$:

$$p(\mathbf{x}) = \prod_{i \in \mathcal{V}} p_i(x_i) \prod_{(i,j) \in \mathcal{E}_p} \frac{p_{i,j}(x_i, x_j)}{p_i(x_i) p_j(x_j)}. \tag{1}$$

We assume that p, in addition to being Markov on the spanning tree $T_p = (\mathcal{V}, \mathcal{E}_p)$, is a *Gaussian graphical model* or *Gauss-Markov random field* (GMRF) with known zero mean[2] and unknown positive definite covariance matrix $\mathbf{\Sigma} \succ 0$. Thus, $p(\mathbf{x})$ can be written as

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2} \mathbf{x}^T \mathbf{\Sigma}^{-1} \mathbf{x}\right). \tag{2}$$

We also use the notation $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{\Sigma})$ as a shorthand for (2). For Gaussian graphical models, it is known that the fill-pattern of the inverse covariance matrix $\mathbf{\Sigma}^{-1}$ encodes the structure of $p(\mathbf{x})$ [3], i.e., $\mathbf{\Sigma}^{-1}(i, j) = 0$ if and only if (iff) $(i, j) \notin \mathcal{E}_p$.

We denote the set of pdfs on $\mathbb{R}^d$ by $\mathcal{P}(\mathbb{R}^d)$, the set of Gaussian pdfs on $\mathbb{R}^d$ by $\mathcal{P}_\mathcal{N}(\mathbb{R}^d)$ and the set of Gaussian graphical models which factorize according to some tree in $\mathcal{T}^d$ as $\mathcal{P}_\mathcal{N}(\mathbb{R}^d, \mathcal{T}^d)$. For learning the structure of $p(\mathbf{x})$ (or equivalently the fill-pattern of $\mathbf{\Sigma}^{-1}$), we are provided with a set of d-dimensional samples $\mathbf{x}^n := \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ drawn from p, where $\mathbf{x}_k := [x_{k,1}, \ldots, x_{k,d}]^T \in \mathbb{R}^d$.

---

[1]In this paper, we use the terms "graphical models" and "Markov random fields" interchangeably.

[2]Our results also extend to the scenario where the mean of the Gaussian is unknown and has to be estimated from the samples.

B. ML Estimation of Gaussian Tree Models

In this subsection, we review the Chow-Liu ML learning algorithm [5] for estimating the structure of p given samples $\mathbf{x}^n$. Denoting $D(p_1 \| p_2) := \mathbb{E}_{p_1} \log(p_1/p_2)$ as the Kullback-Leibler (KL) divergence [19] between p_1 and p_2, the ML estimate of the structure $\mathcal{E}_{\mathrm{CL}}(\mathbf{x}^n)$ is given by the optimization problem[3]

$$\mathcal{E}_{\mathrm{CL}}(\mathbf{x}^n) := \underset{\mathcal{E}_q : q \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d)}{\arg \min} D(\widehat{p} \| q) \qquad (3)$$

where $\widehat{p}(\mathbf{x}) := \mathcal{N}(\mathbf{x}; \mathbf{0}, \widehat{\boldsymbol{\Sigma}})$ and $\widehat{\boldsymbol{\Sigma}} := 1/n \sum_{k=1}^{n} \mathbf{x}_k \mathbf{x}_k^T$ is the *empirical covariance matrix*. Given $\widehat{p}$, and exploiting the fact that q in (3) factorizes according to a tree as in (1), Chow and Liu [5] showed that the optimization for the optimal edge set in (3) can be reduced to a MWST problem:

$$\mathcal{E}_{\mathrm{CL}}(\mathbf{x}^n) = \underset{\mathcal{E}_q : q \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d)}{\arg \max} \sum_{e \in \mathcal{E}_q} I(\widehat{p}_e) \qquad (4)$$

where the edge weights are the *empirical mutual information quantities* [19] given by[4]

$$I(\widehat{p}_e) = -\frac{1}{2} \log \left(1 - \widehat{\rho}_e^2\right) \qquad (5)$$

and where the *empirical correlation coefficients* are given by $\widehat{\rho}_e = \widehat{\rho}_{i,j} := \widehat{\boldsymbol{\Sigma}}(i,j)/(\widehat{\boldsymbol{\Sigma}}(i,i)\widehat{\boldsymbol{\Sigma}}(j,j))^{1/2}$. Note that in (4), the estimated edge set $\mathcal{E}_{\mathrm{CL}}(\mathbf{x}^n)$ depends on n and, specifically, on the samples in $\mathbf{x}^n$ and we make this dependence explicit. We assume that T_p is a spanning tree because with probability 1, the resulting optimization problem in (4) produces a spanning tree as all the mutual information quantities in (5) will be nonzero. If T_p *were* allowed to be a *proper forest* (a tree that is not connected), the estimation of $\mathcal{E}_p$ will be inconsistent because the learned edge set will be different from the true edge set.

C. Problem Statement

We now state our problem formally. Given a set of i.i.d. samples $\mathbf{x}^n$ drawn from an unknown Gaussian tree model p with edge set $\mathcal{E}_p$, we define the error event that the set of edges is estimated incorrectly as

$$\mathcal{A}_n := \{\mathbf{x}^n : \mathcal{E}_{\mathrm{CL}}(\mathbf{x}^n) \neq \mathcal{E}_p\} \qquad (6)$$

where $\mathcal{E}_{\mathrm{CL}}(\mathbf{x}^n)$ is the edge set of the Chow-Liu ML estimator in (3). In this paper, we are interested to *compute* and subsequently *study* the *error exponent* K_p, or the rate that the error probability of the event $\mathcal{A}_n$ with respect to the *true* model p decays with the number of samples n. K_p is defined as

$$K_p := \lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}(\mathcal{A}_n) \qquad (7)$$

assuming the limit exists and where $\mathbb{P}$ is the product probability measure with respect to the true model p. We prove that the limit

---

[3]Note that it is unnecessary to impose the Gaussianity constraint on q in (3). We can optimize over $\mathcal{P}(\mathbb{R}^d, \mathcal{T}^d)$ instead of $\mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d)$. It can be shown that the optimal distribution is still Gaussian. We omit the proof for brevity.

[4]Our notation for the mutual information between two random variables differs from the conventional one in [19].

in (7) exists in Section III (Corollary 3). The value of K_p for different tree models p provides an indication of the relative ease of estimating such models. Note that both the *parameters* and *structure* of the model influence the magnitude of K_p.

III. DERIVING THE ERROR EXPONENT

A. Crossover Rates for Mutual Information Quantities

To compute K_p, consider first two pairs of nodes $e, e' \in \binom{\mathcal{V}}{2}$ such that $I(p_e) > I(p_{e'})$. We now derive a large-deviation principle (LDP) for the *crossover event of empirical mutual information quantities*

$$\mathcal{C}_{e,e'} := \{\mathbf{x}^n : I(\widehat{p}_e) \leq I(\widehat{p}_{e'})\}. \qquad (8)$$

This is an important event for the computation of K_p because if two pairs of nodes (or node pairs) e and e' happen to *crossover*, this *may* lead to the event $\mathcal{A}_n$ occurring (see the next subsection). We define $J_{e,e'} = J_{e,e'}(p_{e,e'})$, the *crossover rate of empirical mutual information quantities*, as

$$J_{e,e'} := \lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}(\mathcal{C}_{e,e'}). \qquad (9)$$

Here we remark that the following analysis does not depend on whether e and e' share a node. If e and e' do share a node, we say they are an *adjacent* pair of nodes. Otherwise, we say e and e' are *disjoint*. We also reserve the symbol m to denote the total number of distinct nodes in e and e'. Hence, $m = 3$ if e and e' are adjacent and $m = 4$ if e and e' are disjoint.

Theorem 1 (LDP for Crossover of Empirical MI): For two node pairs $e, e' \in \binom{\mathcal{V}}{2}$ with pdf $p_{e,e'} \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^m)$ (for $m = 3$ or $m = 4$), the crossover rate for empirical mutual information quantities is

$$J_{e,e'} = \inf_{q \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^m)} \left\{ D(q \| p_{e,e'}) : I(q_e) = I(q_{e'}) \right\}. \qquad (10)$$

The crossover rate $J_{e,e'} > 0$ iff the correlation coefficients of $p_{e,e'}$ satisfy $|\rho_e| \neq |\rho_{e'}|$.

Proof (Sketch): This is an application of Sanov's Theorem [20, Ch. 3], and the contraction principle [21, Ch. 3] in large deviations theory, together with the maximum entropy principle [19, Ch. 12]. We remark that the proof is different from the corresponding result in [7]. See Appendix A. ∎

Theorem 1 says that in order to compute the crossover rate $J_{e,e'}$, we can restrict our attention to a problem that involves only an optimization over Gaussians, which is a finite-dimensional optimization problem.

B. Error Exponent for Structure Learning

We now relate the set of crossover rates $\{J_{e,e'}\}$ over all the node pairs e, e' to the error exponent K_p, defined in (7). The primary idea behind this computation is the following: We consider a fixed non-edge $e' \notin \mathcal{E}_p$ in the true tree T_p which may be erroneously selected during learning process. Because of the global tree constraint, this non-edge e' must *replace* some edge along its unique path in the original model. We only need to consider a single such crossover event because K_p will be larger if there are multiple crossovers (see formal proof in [7]). Finally,
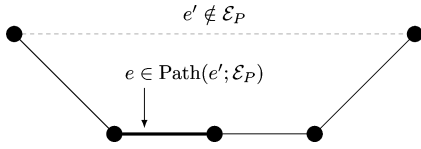
Fig. 1. If the error event occurs during the learning process, an edge $e \in \mathrm{Path}(e'; \mathcal{E}_P)$ is replaced by a non-edge $e' \notin \mathcal{E}_p$ in the original model. We identify the crossover event that has the minimum rate $J_{e,e'}$ and its rate is K_p.

we identify the crossover event that has the minimum rate. See Fig. 1 for an illustration of this intuition.

Theorem 2 (Exponent as a Crossover Event): The error exponent for structure learning of tree-structured Gaussian graphical models, defined in (7), is given as

$$K_p = \min_{e' \notin \mathcal{E}_p} \min_{e \in \mathrm{Path}(e'; \mathcal{E}_p)} J_{e,e'} \qquad (11)$$

where $\mathrm{Path}(e'; \mathcal{E}_p) \subset \mathcal{E}_p$ is the unique path joining the nodes in e' in the original tree $T_p = (\mathcal{V}, \mathcal{E}_p)$.

This theorem implies that the *dominant error tree* [7], which is the asymptotically most-likely estimated error tree under the error event $\mathcal{A}_n$, differs from the true tree T_p in exactly one edge. Note that in order to compute the error exponent K_p in (11), we need to compute at most $\mathrm{diam}(T_p)(d-1)(d-2)/2$ crossover rates, where $\mathrm{diam}(T_p)$ is the diameter of T_p. Thus, this is a significant reduction in the complexity of computing K_p as compared to performing an exhaustive search over all possible error events which requires a total of $\mathcal{O}(d^{d-2})$ computations [22] (equal to the number of spanning trees with d nodes).

In addition, from the result in Theorem 2, we can derive conditions to ensure that $K_p > 0$ and hence for the error probability to decay exponentially.

Corollary 3 (Condition for Positive Error Exponent): The error probability $\mathbb{P}(\mathcal{A}_n)$ decays exponentially, i.e., $K_p > 0$ iff Σ has full rank and T_p is not a forest (as was assumed in Section II).

Proof: See Appendix B for the proof. ∎

The above result provides necessary and sufficient conditions for the error exponent K_p to be positive, which implies exponential decay of the error probability in n, the number of samples. Our goal now is to analyze the influence of structure and parameters of the Gaussian distribution p on the *magnitude* of the error exponent K_p. Such an exercise requires a closed-form expression for K_p, which in turn, requires a closed-form expression for the crossover rate $J_{e,e'}$. However, the crossover rate, despite having an exact expression in (10), can only be found numerically, since the optimization is nonconvex (due to the highly nonlinear equality constraint $I(q_e) = I(q_{e'})$). Hence, we provide an approximation to the crossover rate in the next section which is tight in the so-called very noisy learning regime.

IV. EUCLIDEAN APPROXIMATIONS

In this section, we use an approximation that only considers parameters of Gaussian tree models that are "hard" for learning. There are three reasons for doing this. First, we expect parameters which result in easy problems to have large error exponents and so the structures can be learned accurately from a moderate number of samples. Hard problems thus lend much more insight into when and how errors occur. Second, it allows us to

approximate the intractable problem in (10) with an intuitive, closed-form expression. Finally, such an approximation allows us to compare the relative ease of learning various tree structures in the subsequent sections.

Our analysis is based on Euclidean information theory [18], which we exploit to approximate the crossover rate $J_{e,e'}$ and the error exponent K_p, defined in (9) and (7), respectively. The key idea is to impose suitable "noisy" conditions on $p_{e,e'}$ (the joint pdf on node pairs e and e') so as to enable us to relax the nonconvex optimization problem in (10) to a convex program.

Definition 1 (ε-Very Noisy Condition): The joint pdf $p_{e,e'}$ on node pairs e and e' is said to satisfy the *ε-very noisy condition* if the correlation coefficients on e and e' satisfy $\||\rho_e| - |\rho_{e'}|\| < \epsilon$.

By continuity of the mutual information in the correlation coefficient, given any fixed ϵ and ρ_e, there exists a $\delta = \delta(\epsilon, \rho_e) > 0$ such that $|I(p_e) - I(p_{e'})| < \delta$, which means that if ϵ is small, it is difficult to distinguish which node pair e or e' has the larger mutual information given the samples $\mathbf{x}^n$. Therefore the ordering of the empirical mutual information quantities $I(\widehat{p}_e)$ and $I(\widehat{p}_{e'})$ may be incorrect. Thus, if ϵ is small, we are in the very noisy learning regime, where learning is difficult.

To perform our analysis, we recall from Verdu [23, Sec. IV-E] that we can bound the KL-divergence between two zero-mean Gaussians with covariance matrices $\Sigma_{e,e'} + \Delta_{e,e'}$ and $\Sigma_{e,e'}$ as

$$D(\mathcal{N}(\mathbf{0}, \Sigma_{e,e'} + \Delta_{e,e'}) \| \mathcal{N}(\mathbf{0}, \Sigma_{e,e'})) \leq \frac{\|\Sigma_{e,e'}^{-1} \Delta_{e,e'}\|_F^2}{4} \quad (12)$$

where $\|\mathbf{M}\|_F$ is the Frobenius norm of the matrix $\mathbf{M}$. Furthermore, the inequality in (12) is tight when the perturbation matrix $\Delta_{e,e'}$ is small. More precisely, as the ratio of the singular values $\sigma_{\max}(\Delta_{e,e'})/\sigma_{\min}(\Sigma_{e,e'})$ tends to zero, the inequality in (12) becomes tight. To convexify the problem, we also perform a linearization of the nonlinear constraint set in (10) around the unperturbed covariance matrix $\Sigma_{e,e'}$. This involves taking the derivative of the mutual information with respect to the covariance matrix in the Taylor expansion. We denote this derivative as $\nabla_{\Sigma_e} I(\Sigma_e)$ where $I(\Sigma_e) = I(\mathcal{N}(\mathbf{0}, \Sigma_e))$ is the mutual information between the two random variables of the Gaussian joint pdf $p_e = \mathcal{N}(\mathbf{0}, \Sigma_e)$. We now define the *linearized constraint set* of (10) as the affine subspace

$$L_\Delta(p_{e,e'}) := \{\Delta_{e,e'} \in \mathbb{R}^{m \times m} : I(\Sigma_e) + \langle \nabla_{\Sigma_e} I(\Sigma_e), \Delta_e \rangle$$
$$= I(\Sigma_{e'}) + \langle \nabla_{\Sigma_{e'}} I(\Sigma_{e'}), \Delta_{e'} \rangle\}, \quad (13)$$

where $\Delta_e \in \mathbb{R}^{2 \times 2}$ is the sub-matrix of $\Delta_{e,e'} \in \mathbb{R}^{m \times m}$ ($m = 3$ or 4) that corresponds to the covariance matrix of the node pair e. We also define the *approximate crossover rate* of e and e' as the minimization of the quadratic in (12) over the affine subspace $L_\Delta(p_{e,e'})$ defined in (13)

$$\widetilde{J}_{e,e'} := \min_{\Delta_{e,e'} \in L_\Delta(p_{e,e'})} \frac{1}{4} \|\Sigma_{e,e'}^{-1} \Delta_{e,e'}\|_F^2. \quad (14)$$

Equation (14) is a *convexified* version of the original optimization in (10). This problem is not only much easier to solve, but also provides key insights as to when and how errors occur when learning the structure. We now define an additional information-theoretic quantity before stating the Euclidean approximation.
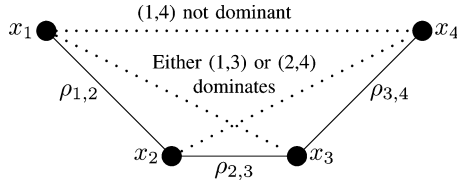
Fig. 2. Illustration of correlation decay in a Markov chain. By Lemma 5(b), only the node pairs $(1,3)$ and $(2,4)$ need to be considered for computing the error exponent $\widetilde{K}_p$. By correlation decay, the node pair $(1,4)$ will not be mistaken as a true edge by the estimator because its distance, which is equal to 3, is longer than either $(1,3)$ or $(2,4)$, whose distances are equal to 2.

Definition 2 (Information Density): Given a pairwise joint pdf $p_{i,j}$ with marginals p_i and p_j, the *information density* denoted by $s_{i,j} : \mathbb{R}^2 \rightarrow \mathbb{R}$, is defined as

$$s_{i,j}(x_i, x_j) := \log \frac{p_{i,j}(x_i, x_j)}{p_i(x_i)p_j(x_j)}. \tag{15}$$

Hence, for each pair of variables x_i and x_j, its associated information density $s_{i,j}$ is a random variable whose expectation is the mutual information of x_i and x_j, i.e., $\mathbb{E}[s_{i,j}] = I(p_{i,j})$.

Theorem 4 (Euclidean Approx. of Crossover Rate): The approximate crossover rate for the empirical mutual information quantities, defined in (14), is given by

$$\widetilde{J}_{e,e'} = \frac{(\mathbb{E}[s_{e'} - s_e])^2}{2 \operatorname{Var}(s_{e'} - s_e)} = \frac{(I(p_{e'}) - I(p_e))^2}{2 \operatorname{Var}(s_{e'} - s_e)}. \tag{16}$$

In addition, the approximate error exponent corresponding to $\widetilde{J}_{e,e'}$ in (14) is given by

$$\widetilde{K}_p = \min_{e' \in \mathcal{E}_p} \min_{e \in \operatorname{Path}(e'; \mathcal{E}_p)} \widetilde{J}_{e,e'}. \tag{17}$$

Proof: The proof involves solving the least squares problem in (14). See Appendix C. ∎

We have obtained a closed-form expression for the approximate crossover rate $\widetilde{J}_{e,e'}$ in (16). It is proportional to the square of the difference between the mutual information quantities. This corresponds to our intuition—that if $I(p_e)$ and $I(p_{e'})$ are relatively well separated ($I(p_e) \gg I(p_{e'})$) then the rate $\widetilde{J}_{e,e'}$ is large. In addition, the SNR is also weighted by the inverse variance of the difference of the information densities $s_e - s_{e'}$. If the variance is large, then we are uncertain about the estimate $I(\widehat{p}_e) - I(\widehat{p}_{e'})$, thereby reducing the rate. Theorem 4 illustrates how *parameters* of Gaussian tree models affect the crossover rate. In the sequel, we limit our analysis to the very noisy regime where the above expressions apply.

V. SIMPLIFICATION OF THE ERROR EXPONENT

In this section, we exploit the properties of the approximate crossover rate in (16) to significantly reduce the complexity in finding the error exponent $\widetilde{K}_p$ to $\mathcal{O}(d)$. As a motivating example, consider the Markov chain in Fig. 2. From our analysis to this point, it appears that, when computing the approximate error exponent $\widetilde{K}_p$ in (17), we have to consider all possible replacements between the non-edges $(1,4)$, $(1,3)$, and $(2,4)$ and the true edges along the paths connecting these non-edges. For example, $(1,3)$ might be mistaken as a true edge, replacing either $(1,2)$ or $(2,3)$.

We will prove that, in fact, to compute $\widetilde{K}_p$ we can ignore the possibility that longest non-edge $(1,4)$ is mistaken as a true edge, thus reducing the number of computations for the approximate crossover rate $\widetilde{J}_{e,e'}$. The key to this result is the exploitation of *correlation decay*, i.e., the decrease in the absolute value of the correlation coefficient between two nodes as the *distance* (the number of edges along the path between two nodes) between them increases. This follows from the Markov property:

$$\rho_{e'} = \prod_{e \in \operatorname{Path}(e'; \mathcal{E}_p)} \rho_e, \quad \forall e' \notin \mathcal{E}_p. \tag{18}$$

For example, in Fig. 2, $|\rho_{1,4}| \leq \min\{|\rho_{1,3}|, |\rho_{2,4}|\}$ and because of this, the following lemma implies that $(1,4)$ is less likely to be mistaken as a true edge than $(1,3)$ or $(2,4)$.

It is easy to verify that the crossover rate $\widetilde{J}_{e,e'}$ in (16) depends *only* on the correlation coefficients ρ_e and $\rho_{e'}$ and not the variances σ_i^2. Thus, without loss of generality, we assume that all random variables have unit variance (which is still unknown to the learner) and to make the dependence clear, we now write $\widetilde{J}_{e,e'} = \widetilde{J}(\rho_e, \rho_{e'})$. Finally define $\rho_{\text{crit}} := 0.63055$.

Lemma 5 (Monotonicity of $\widetilde{J}(\rho_e, \rho_{e'})$): $\widetilde{J}(\rho_e, \rho_{e'})$, derived in (16), has the following properties:

a) $\widetilde{J}(\rho_e, \rho_{e'})$ is an even function of both ρ_e and $\rho_{e'}$;

b) $\widetilde{J}(\rho_e, \rho_{e'})$ is monotonically *decreasing* in $|\rho_{e'}|$ for fixed $\rho_e \in (-1, 1)$;

c) Assuming that $|\rho_{e_1}| < \rho_{\text{crit}}$, then $\widetilde{J}(\rho_{e_1}, \rho_{e_1}\rho_{e_2})$ is monotonically *increasing* in $|\rho_{e_1}|$ for fixed ρ_{e_2};

d) Assuming that $|\rho_e| < \rho_{\text{crit}}$, then $\widetilde{J}(\rho_e, \rho_{e'})$ is monotonically *increasing* in $|\rho_e|$ for fixed $\rho_{e'}$.

See Fig. 3 for an illustration of the properties of $\widetilde{J}(\rho_e, \rho_{e'})$.

Proof (Sketch): Statement (a) follows from (16). We prove (b) by showing that $\partial \widetilde{J}(\rho_e, \rho_{e'})/\partial|\rho_{e'}| \leq 0$ for all $|\rho_{e'}| \leq |\rho_e|$. Statements (c) and (d) follow similarly. See Appendix D for the details. ∎

Our intuition about correlation decay is substantiated by Lemma 5(b), which implies that for the example in Fig. 2, $\widetilde{J}(\rho_{2,3}, \rho_{1,3}) \leq \widetilde{J}(\rho_{2,3}, \rho_{1,4})$, since $|\rho_{1,4}| \leq |\rho_{1,3}|$ due to Markov property on the chain (18). Therefore, $\widetilde{J}(\rho_{2,3}, \rho_{1,4})$ can be ignored in the minimization to find $\widetilde{K}_p$ in (17). Interestingly while Lemma 5(b) is a statement about correlation decay, Lemma 5(c) states that the absolute strengths of the correlation coefficients also influence the magnitude of the crossover rate.

From Lemma 5(b) (and the above motivating example in Fig. 2), finding the approximate error exponent $\widetilde{K}_p$ now reduces to finding the minimum crossover rate only over *triangles* ($(1,2,3)$ and $(2,3,4)$) in the tree as shown in Fig. 2, i.e., we only need to consider $\widetilde{J}(\rho_e, \rho_{e'})$ for *adjacent* edges.

Corollary 6 (Computation of $\widetilde{K}_p$): Under the very noisy learning regime, the error exponent $\widetilde{K}_p$ is

$$\widetilde{K}_p = \min_{e_i, e_j \in \mathcal{E}_p, e_i \sim e_j} W(\rho_{e_i}, \rho_{e_j}) \tag{19}$$

where $e_i \sim e_j$ means that the edges e_i and e_j are adjacent and the weights are defined as

$$W(\rho_{e_1}, \rho_{e_2}) := \min\left\{\widetilde{J}(\rho_{e_1}, \rho_{e_1}\rho_{e_2}), \widetilde{J}(\rho_{e_2}, \rho_{e_1}\rho_{e_2})\right\}. \tag{20}$$

If we carry out the computations in (19) independently, the complexity is $\mathcal{O}(d\deg_{\max})$, where $\deg_{\max}$ is the maximum de-
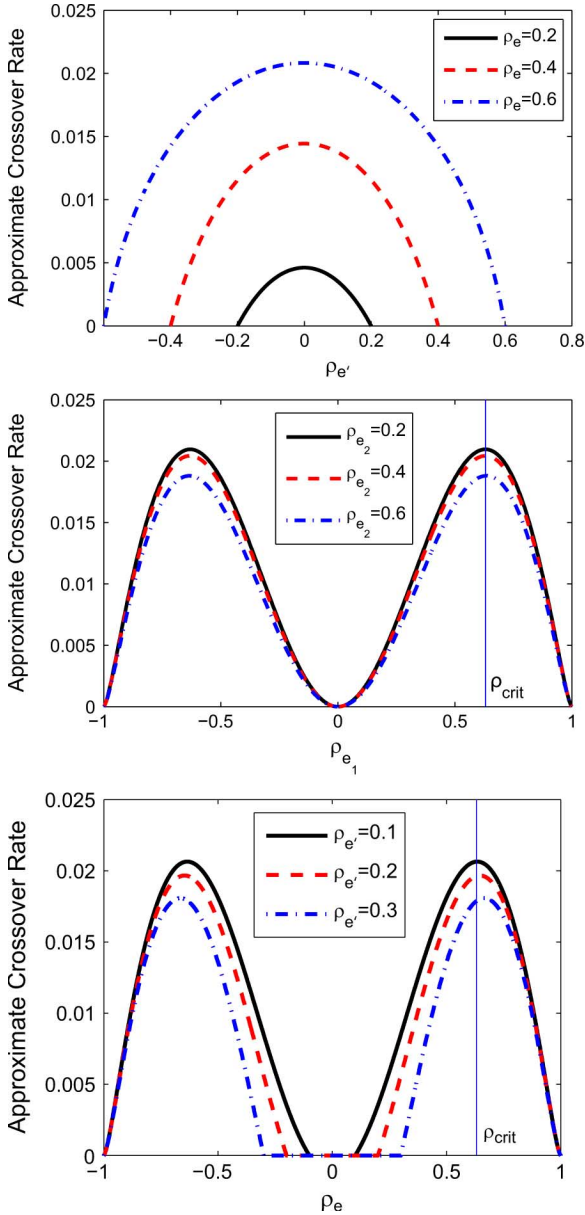
Fig. 3. Illustration of the properties of $\widetilde{J}(\rho_e, \rho_{e'})$ in Lemma 5. $\widetilde{J}(\rho_e, \rho_{e'})$ is decreasing in $|\rho_{e'}|$ for fixed ρ_e (top) and $\widetilde{J}(\rho_{e_1}, \rho_{e_1}\rho_{e_2})$ is increasing in $|\rho_{e_1}|$ for fixed ρ_{e_2} if $|\rho_{e_1}| < \rho_{\mathrm{crit}}$ (middle). Similarly, $\widetilde{J}(\rho_e, \rho_{e'})$ is increasing in $|\rho_e|$ for fixed $\rho_{e'}$ if $|\rho_e| < \rho_{\mathrm{crit}}$ (bottom).

gree of the nodes in the tree graph. Hence, in the worst case, the complexity is $\mathcal{O}(d^2)$, instead of $\mathcal{O}(d^3)$ if (17) is used. We can, in fact, reduce the number of computations to $\mathcal{O}(d)$.

Proposition 7 (Complexity in Computing $\widetilde{K}_p$): The approximate error exponent $\widetilde{K}_p$, derived in (17), can be computed in linear time ($d-1$ operations) as

$$\widetilde{K}_p = \min_{e \in \mathcal{E}_p} \widetilde{J}(\rho_e, \rho_e \rho_e^*) \tag{21}$$

where the maximum correlation coefficient on the edges adjacent to $e \in \mathcal{E}_p$ is defined as

$$\rho_e^* := \max\{|\rho_{\tilde{e}}| : \tilde{e} \in \mathcal{E}_p, \tilde{e} \sim e\}. \tag{22}$$

Proof: By Lemma 5(b) and the definition of ρ_e^*, we obtain the smallest crossover rate associated to edge e. We obtain the approximate error exponent $\widetilde{K}_p$ by minimizing over all edges $e \in \mathcal{E}_p$ in (21). ∎

Recall that $\mathrm{diam}(T_p)$ is the diameter of T_p. The computation of K_p is reduced significantly from $\mathcal{O}(\mathrm{diam}(T_p)d^2)$ in (11) to $\mathcal{O}(d)$. Thus, there is a further reduction in the complexity to estimate the error exponent K_p as compared to exhaustive search which requires $\mathcal{O}(d^{d-2})$ computations. This simplification only holds for Gaussians under the very noisy regime.

VI. EXTREMAL STRUCTURES FOR LEARNING

In this section, we study the influence of graph structure on the approximate error exponent $\widetilde{K}_p$ using the concept of correlation decay and the properties of the crossover rate $\widetilde{J}_{e,e'}$ in Lemma 5. We have already discussed the connection between the error exponent and correlation decay. We also proved that non-neighbor node pairs which have shorter distances are more likely to be mistaken as edges by the ML estimator. Hence, we expect that a tree T_p which contains non-edges with shorter distances to be "harder" to learn (i.e., has a smaller error exponent $\widetilde{K}_p$) as compared to a tree which contains non-edges with longer distances. In subsequent subsections, we formalize this intuition in terms of the diameter of the tree $\mathrm{diam}(T_p)$, and show that the extremal trees, in terms of their diameter, are also extremal trees for learning. We also analyze the effect of changing the size of the tree on the error exponent.

From the Markov property in (18), we see that for a Gaussian tree distribution, the set of correlation coefficients fixed on the edges of the tree, along with the structure T_p, are sufficient statistics and they completely characterize p. Note that this parameterization neatly decouples the structure from the correlations. We use this fact to study the influence of changing the structure T_p while keeping the set of correlations on the edges fixed.[5] Before doing so, we provide a review of some basic graph theory.

A. Basic Notions in Graph Theory

Definition 3 (Extremal Trees in Terms of Diameter): Assume that $d > 3$. Define the *extremal trees* with d nodes in terms of the tree diameter $\mathrm{diam} : \mathcal{T}^d \to \{2, \ldots, d-1\}$ as

$$T_{\max}(d) := \arg\max_{T \in \mathcal{T}^d} \mathrm{diam}(T)$$
$$T_{\min}(d) := \arg\min_{T \in \mathcal{T}^d} \mathrm{diam}(T) \tag{23}$$

Then it is clear that the two extremal structures, the *chain* (where there is a simple path passing through all nodes and edges exactly once) and the *star* (where there is one central node) have the largest and smallest diameters, respectively, i.e., $T_{\max}(d) = T_{\mathrm{chain}}(d)$, and $T_{\min}(d) = T_{\mathrm{star}}(d)$.

Definition 4 (Line Graph): The *line graph* [22] H of a graph G, denoted by $H = \mathcal{L}(G)$, is one in which, roughly speaking, the vertices and edges of G are interchanged. More precisely, H is the undirected graph whose vertices are the edges of G and there is an edge between any two vertices in the line graph

---

[5]Although the set of correlation coefficients on the edges is fixed, the elements in this set can be arranged in different ways on the edges of the tree. We formalize this concept in (24).
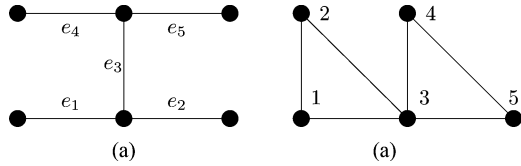
Fig. 4. (a): A graph G. (b): The line graph $H = \mathcal{L}(G)$ that corresponds to G is the graph whose vertices are the edges of G (denoted as e_i) and there is an edge between any two vertices i and j in H if the corresponding edges in G share a node.

if the corresponding edges in G have a common node, i.e., are adjacent. See Fig. 4 for a graph G and its associated line graph H.

B. Formulation: Extremal Structures for Learning

We now formulate the problem of finding the best and worst tree structures for learning and also the distributions associated with them. At a high level, our strategy involves two distinct steps. First, and primarily, we find the *structure* of the optimal distributions in Section VI-D. It turns out that the optimal structures that maximize and minimize the exponent are the Markov chain (under some conditions on the correlations) and the star, respectively, and these are the extremal structures in terms of the diameter. Second, we optimize over the *positions* (or placement) of the correlation coefficients on the edges of the optimal structures.

Let $\boldsymbol{\rho} := [\rho_1, \rho_2, \ldots, \rho_{d-1}]$ be a *fixed* vector of feasible[6] correlation coefficients, i.e., $\rho_i \in (-1, 1) \setminus \{0\}$ for all i. For a tree, it follows from (18) that if ρ_i's are the correlation coefficients on the edges, then $|\rho_i| < 1$ is a necessary and sufficient condition to ensure that $\boldsymbol{\Sigma} \succ 0$. Define $\boldsymbol{\Pi}_{d-1}$ to be the group of permutations of order $d-1$, hence elements in $\boldsymbol{\Pi}_{d-1}$ are permutations of a given ordered set with cardinality $d-1$. Also denote the set of tree-structured, d-variate Gaussians which have unit variances at all nodes and $\boldsymbol{\rho}$ as the correlation coefficients on the edges in some order as $\mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d; \boldsymbol{\rho})$. Formally,

$$\mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d; \boldsymbol{\rho}) := \{p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \boldsymbol{\Sigma}) \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d) :$$
$$\boldsymbol{\Sigma}(i, i) = 1, \forall i \in \mathcal{V}, \exists \boldsymbol{\pi}_p \in \boldsymbol{\Pi}_{d-1} : \boldsymbol{\sigma}_{\mathcal{E}_p} = \boldsymbol{\pi}_p(\boldsymbol{\rho})\}, \quad (24)$$

where $\boldsymbol{\sigma}_{\mathcal{E}_p} := [\boldsymbol{\Sigma}(i, j) : (i, j) \in \mathcal{E}_p]$ is the length-$(d-1)$ vector consisting of the covariance elements[7] on the edges (arranged in lexicographic order) and $\boldsymbol{\pi}_p(\boldsymbol{\rho})$ is the permutation of $\boldsymbol{\rho}$ according to $\boldsymbol{\pi}_p$. The tuple $(T_p, \boldsymbol{\pi}_p, \boldsymbol{\rho})$ uniquely parameterizes a Gaussian tree distribution with unit variances. Note that we can regard the permutation $\boldsymbol{\pi}_p$ as a nuisance parameter for solving the optimization for the best structure given $\boldsymbol{\rho}$. Indeed, it can happen that there are different $\boldsymbol{\pi}_p$'s such that the error exponent $\widetilde{K}_p$ is the same. For instance, in a star graph, all permutations $\boldsymbol{\pi}_p$ result in the same exponent. Despite this, we show that extremal tree *structures* are invariant to the specific choice of $\boldsymbol{\pi}_p$ and $\boldsymbol{\rho}$.

For distributions in the set $\mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d; \boldsymbol{\rho})$, our goal is to find the best (easiest to learn) and the worst (most difficult to learn)

[6]We do not allow any of the correlation coefficient to be zero because otherwise, this would result in T_p being a forest.

[7]None of the elements in $\boldsymbol{\Sigma}$ are allowed to be zero because $\rho_i \neq 0$ for every $i \in \mathcal{V}$ and the Markov property in (18).

distributions for learning. Formally, the optimization problems for the best and worst distributions for learning are given by

$$p_{\max, \boldsymbol{\rho}} := \arg\max_{p \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d; \boldsymbol{\rho})} \widetilde{K}_p, \quad (25)$$

$$p_{\min, \boldsymbol{\rho}} := \arg\min_{p \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d; \boldsymbol{\rho})} \widetilde{K}_p. \quad (26)$$

Thus, $p_{\max, \boldsymbol{\rho}}$ (respectively, $p_{\min, \boldsymbol{\rho}}$) corresponds to the Gaussian tree model which has the largest (respectively, smallest) approximate error exponent.

C. Reformulation as Optimization Over Line Graphs

Since the number of permutations $\boldsymbol{\pi}$ and number of spanning trees are prohibitively large, finding the optimal distributions cannot be done through a brute-force search unless d is small. Our main idea in this section is to use the notion of line graphs to simplify the problems in (25) and (26). In subsequent sections, we identify the extremal tree structures before identifying the precise best and worst distributions.

Recall that the approximate error exponent $\widetilde{K}_p$ can be expressed in terms of the weights $W(\rho_{e_i}, \rho_{e_j})$ between two adjacent edges e_i, e_j as in (19). Therefore, we can write the extremal distribution in (25) as

$$p_{\max, \boldsymbol{\rho}} = \arg\max_{p \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d; \boldsymbol{\rho})} \min_{e_i, e_j \in \mathcal{E}_p, e_i \sim e_j} W(\rho_{e_i}, \rho_{e_j}). \quad (27)$$

Note that in (27), $\mathcal{E}_p$ is the edge set of a weighted graph whose edge weights are given by $\boldsymbol{\rho}$. Since the weight is between two edges, it is more convenient to consider line graphs defined in Section VI-A.

We now transform the intractable optimization problem in (27) over the set of trees to an optimization problem over all the set of line graphs:

$$p_{\max, \boldsymbol{\rho}} = \arg\max_{p \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d; \boldsymbol{\rho})} \min_{(i, j) \in H, H = \mathcal{L}(T_p)} W(\rho_i, \rho_j) \quad (28)$$

and $W(\rho_i, \rho_j)$ can be considered as an edge weight between nodes i and j in a weighted line graph H. Equivalently, (26) can also be written as in (28) but with then argmax replaced by an argmin.

D. Main Results: Best and Worst Tree Structures

In order to solve (28), we need to characterize the set of line graphs of spanning trees $\mathcal{L}(\mathcal{T}^d) = \{\mathcal{L}(T) : T \in \mathcal{T}^d\}$. This has been studied before [24, Theorem 8.5], but the set $\mathcal{L}(\mathcal{T}^d)$ is nonetheless still very complicated. Hence, solving (28) directly is intractable. Instead, our strategy now is to identify the *structures* corresponding to the optimal distributions, $p_{\max, \rho}$ and $p_{\min, \rho}$ by exploiting the monotonicity of $\widetilde{J}(\rho_e, \rho_{e'})$ given in Lemma 5.

Theorem 8 (Extremal Tree Structures): The tree structure that minimizes the approximate error exponent $\widetilde{K}_p$ in (26) is given by
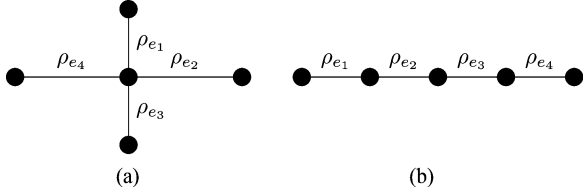
$$T_{p_{\min, \rho}} = T_{\text{star}}(d) \quad (29)$$

Fig. 5. Illustration for Theorem 8: The star (a) and the chain (b) minimize and maximize the approximate error exponent, respectively.

for all feasible correlation coefficient vectors $\boldsymbol{\rho}$ with $\rho_i \in (-1,1) \setminus \{0\}$. In addition, if $\rho_i \in (-\rho_{\mathrm{crit}}, \rho_{\mathrm{crit}}) \setminus \{0\}$ (where $\rho_{\mathrm{crit}} = 0.63055$), then the tree structure that maximizes the approximate error exponent $\widetilde{K}_p$ in (25) is given by

$$T_{p_{\max,\boldsymbol{\rho}}} = T_{\mathrm{chain}}(d) \tag{30}$$

Proof (Idea): The assertion that $T_{p_{\min,\boldsymbol{\rho}}} = T_{\mathrm{star}}(d)$ follows from the fact that all the crossover rates for the star graph are the minimum possible, hence $\widetilde{K}_{\mathrm{star}} \leq \widetilde{K}_p$. See Appendix E for the details. ∎

See Fig. 5. This theorem agrees with our intuition: for the star graph, the nodes are strongly correlated (since its diameter is the smallest) while in the chain, there are many weakly correlated pairs of nodes for the same set of correlation coefficients on the edges thanks to correlation decay. Hence, it is hardest to learn the star while it is easiest to learn the chain. It is interesting to observe Theorem 8 implies that the extremal tree structures $T_{p_{\max,\boldsymbol{\rho}}}$ and $T_{p_{\min,\boldsymbol{\rho}}}$ are *independent of* the correlation coefficients $\boldsymbol{\rho}$ (if $|\rho_i| < \rho_{\mathrm{crit}}$ in the case of the star). Indeed, the experiments in Section VII-B also suggest that Theorem 8 may likely be true for larger ranges of problems (without the constraint that $|\rho_i| < \rho_{\mathrm{crit}}$) but this remains open.

The results in (29) and (30) do not yet provide the complete solution to $p_{\max,\boldsymbol{\rho}}$ and $p_{\min,\boldsymbol{\rho}}$ in (25) and (26) since there are many possible pdfs in $\mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d; \boldsymbol{\rho})$ corresponding to a fixed tree because we can rearrange the correlation coefficients along the edges of the tree in multiple ways. The only exception is if T_p is known to be a star then there is only one pdf in $\mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d; \boldsymbol{\rho})$, and we formally state the result below.

Corollary 9 (Most Difficult Distribution to Learn): The Gaussian $p_{\min,\boldsymbol{\rho}}(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \boldsymbol{\Sigma}_{\min,\boldsymbol{\rho}})$ defined in (26), corresponding to the most difficult distribution to learn for fixed $\boldsymbol{\rho}$, has the covariance matrix whose upper triangular elements are given as $\boldsymbol{\Sigma}_{\min,\boldsymbol{\rho}}(i,j) = \rho_i$ if $i = 1, j \neq 1$ and $\boldsymbol{\Sigma}_{\min,\rho}(i,j) = \rho_i \rho_j$ otherwise. Moreover, if $|\rho_1| \geq \cdots \geq |\rho_{d-1}|$ and $|\rho_1| < \rho_{\mathrm{crit}} = 0.63055$, then $\widetilde{K}_p$ corresponding to the star graph can be written explicitly as a minimization over only two crossover rates: $\widetilde{K}_{p_{\min,\rho}} = \min\{\widetilde{J}(\rho_1, \rho_1\rho_2), \widetilde{J}(\rho_{d-1}, \rho_{d-1}\rho_1)\}$.

Proof: The first assertion follows from the Markov property (18) and Theorem 8. The next result follows from Lemma 5(c) which implies that $\widetilde{J}(\rho_{d-1}, \rho_{d-1}\rho_1) \leq \widetilde{J}(\rho_k, \rho_k \rho_1)$ for all $2 \leq k \leq d - 1$. ∎

In other words, $p_{\min,\rho}$ is a *star* Gaussian graphical model with correlation coefficients ρ_i on its edges. This result can also be explained by correlation decay. In a star graph, since the distances between non-edges are small, the estimator in (3) is more likely to mistake a non-edge with a true edge. It is often useful

in applications to compute the minimum error exponent for a fixed vector of correlations $\boldsymbol{\rho}$ as it provides a lower bound of the decay rate of $\mathbb{P}(\mathcal{A}_n)$ for any tree distribution with parameter vector $\boldsymbol{\rho}$. Interestingly, we also have a result for the easiest tree distribution to learn.

Corollary 10 (Easiest Distribution to Learn): Assume that $\rho_{\mathrm{crit}} > |\rho_1| \geq |\rho_2| \geq \cdots \geq |\rho_{d-1}|$. Then, the Gaussian $p_{\max,\boldsymbol{\rho}}(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \boldsymbol{\Sigma}_{\max,\boldsymbol{\rho}})$ defined in (25), corresponding to the easiest distribution to learn for fixed $\boldsymbol{\rho}$, has the covariance matrix whose upper triangular elements are $\boldsymbol{\Sigma}_{\max,\boldsymbol{\rho}}(i, i+1) = \rho_i$ for all $i = 1, \ldots, d - 1$ and $\boldsymbol{\Sigma}_{\max,\boldsymbol{\rho}}(i,j) = \prod_{k=i}^{j-1} \rho_k$ for all $j > i$.

Proof: The first assertion follows from the proof of Theorem 8 in Appendix E and the second assertion from the Markov property in (18). ∎

In other words, in the regime where $|\rho_i| < \rho_{\mathrm{crit}}$, $p_{\max,\boldsymbol{\rho}}$ is a *Markov chain* Gaussian graphical model with correlation coefficients arranged in increasing (or decreasing) order on its edges. We now provide some intuition for why this is so. If a particular correlation coefficient ρ_i (such that $|\rho_i| < \rho_{\mathrm{crit}}$) is fixed, then the edge weight $W(\rho_i, \rho_j)$, defined in (20), is maximized when $|\rho_j| = |\rho_i|$. Otherwise, if $|\rho_i| < |\rho_j|$ the event that the non-edge with correlation $\rho_i \rho_j$ replaces the edge with correlation ρ_i (and hence results in an error) has a higher likelihood than if equality holds. Thus, correlations ρ_i and ρ_j that are close in terms of their absolute values should be placed closer to one another (in terms of graph distance) for the approximate error exponent to be maximized. See Fig. 6.

E. Influence of Data Dimension on Error Exponent

We now analyze the influence of *changing* the *size* of the tree on the error exponent, i.e., adding and deleting nodes and edges while satisfying the tree constraint and observing samples from the modified graphical model. This is of importance in many applications. For example, in *sequential* problems, the learner receives data at different times and would like to update the estimate of the tree structure learned. In *dimensionality reduction*, the learner is required to estimate the structure of a smaller model given high-dimensional data. Intuitively, learning only a tree with a smaller number of nodes is easier than learning the entire tree since there are fewer ways for errors to occur during the learning process. We prove this in the affirmative in Proposition 11.

Formally, we start with a d-variate Gaussian $p \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d; \boldsymbol{\rho})$ and consider a d'-variate pdf $p' \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^{d'}, \mathcal{T}^{d'}; \boldsymbol{\rho}')$, obtained by marginalizing p over a subset of variables and $T_{p'}$ is the tree[8] associated to the distribution p'. Hence $d' < d$ and $\boldsymbol{\rho}'$ is a subvector of $\boldsymbol{\rho}$. See Fig. 7. In our formulation, the only available observations are those sampled from the smaller Gaussian graphical model p'.

Proposition 11 (Error Exponent of Smaller Trees): The approximate error exponent for learning p' is at least that of p, i.e., $\widetilde{K}_{p'} \geq \widetilde{K}_p$.

---

[8]Note that $T_{p'}$ still needs to satisfy the tree constraint so that the variables that are marginalized out are not arbitrary (but must be variables that form the first part of a node elimination order [3]). For example, we are not allowed to marginalize out the central node of a star graph since the resulting graph would not be a tree. However, we can marginalize out any of the other nodes. In effect, we can only marginalize out nodes with degree either 1 or 2.
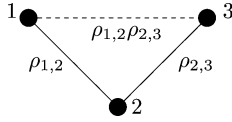
Fig. 6. If $|\rho_{1,2}| < |\rho_{2,3}|$, then the likelihood of the non-edge $(1,3)$ replacing edge $(1,2)$ would be higher than if $|\rho_{1,2}| = |\rho_{2,3}|$. Hence, the weight $W(\rho_{1,2}, \rho_{2,3})$ is maximized when equality holds.
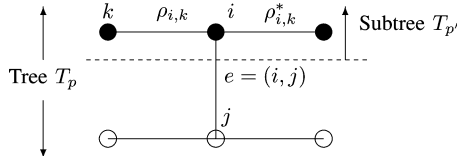

Fig. 7. Illustration of Proposition 11. $T_p = (\mathcal{V}, \mathcal{E}_p)$ is the original tree and $e \in \mathcal{E}_p$. $T_{p'} = (\mathcal{V}', \mathcal{E}_{p'})$ is a subtree. The observations for learning the structure p' correspond to the shaded nodes, the unshaded nodes correspond to unobserved variables.
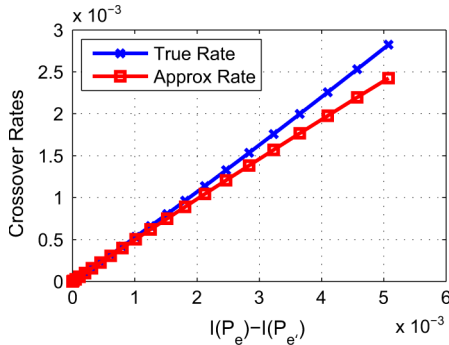

Fig. 8. Comparison of true and approximate crossover rates in (10) and (16), respectively.

Proof: Reducing the number of adjacent edges to a fixed edge $(i,k) \in \mathcal{E}_p$ as in Fig. 7 (where $k \in \text{nbd}(i) \setminus \{j\}$) ensures that the maximum correlation coefficient $\rho_{i,k}^*$, defined in (22), does not increase. By Lemma 5(b) and (17), the approximate error exponent $\widetilde{K}_p$ does not decrease. ∎

Thus, lower-dimensional models are easier to learn if the set of correlation coefficients is fixed and the tree constraint remains satisfied. This is a consequence of the fact that there are fewer crossover error events that contribute to the error exponent $\widetilde{K}_p$.

We now consider the "dual" problem of adding a new edge to an existing tree model, which results in a larger tree. We are now provided with $(d+1)$-dimensional observations to learn the larger tree. More precisely, given a d-variate tree Gaussian pdf p, we consider a $(d+1)$-variate pdf p'' such that T_p is a subtree of $T_{p''}$. Equivalently, let $\boldsymbol{\rho} := [\rho_{e_1}, \rho_{e_2}, \ldots, \rho_{e_{d-1}}]$ be the vector of correlation coefficients on the edges of the graph of p and let $\boldsymbol{\rho}'' := [\boldsymbol{\rho}, \rho_{\text{new}}]$ be that of p''.

By comparing the error exponents $\widetilde{K}_p$ and $\widetilde{K}_{p''}$, we can address the following question: Given a new edge correlation coefficient ρ_{new}, how should one adjoin this new edge to the existing tree such that the resulting error exponent is maximized or minimized? Evidently, from Proposition 11, it is not possible to increase the error exponent by growing the tree but can we devise a strategy to place this new edge judiciously (respectively, adversarially) so that the error exponent deteriorates as little (respectively, as much) as possible?

To do so, we say edge e *contains* node v if $e = (v,i)$ and we define the nodes in the smaller tree T_p

$$v_{\text{min}}^* := \arg\min_{v \in \mathcal{V}} \max_{e \in \mathcal{E}_p} \{|\rho_e| : e \text{ contains node } v\}. \quad (31)$$

$$v_{\text{max}}^* := \arg\max_{v \in \mathcal{V}} \max_{e \in \mathcal{E}_p} \{|\rho_e| : e \text{ contains node } v\}. \quad (32)$$

Proposition 12 (Error Exponent of Larger Trees): Assume that $|\rho_{\text{new}}| < |\rho_e| \ \forall e \in \mathcal{E}_p$. Then,
  (a) The difference between the error exponents $\widetilde{K}_p - \widetilde{K}_{p''}$ is *minimized* when $T_{p''}$ is obtained by adding to T_p a new edge with correlation coefficient ρ_{new} at vertex v_{min}^* given by (31) as a leaf.
  (b) The difference $\widetilde{K}_p - \widetilde{K}_{p''}$ is *maximized* when the new edge is added to v_{max}^* given by (32) as a leaf.
  Proof: The vertex given by (31) is the best vertex to attach the new edge by Lemma 5(b). ∎

This result implies that if we receive data dimensions sequentially, we have a straightforward rule in (31) for identifying larger trees such that the exponent decreases as little as possible at each step.

VII. NUMERICAL EXPERIMENTS

We now perform experiments with the following two objectives. Firstly, we study the accuracy of the Euclidean approximations (Theorem 4) to identify regimes in which the approximate crossover rate $\widetilde{J}_{e,e'}$ is close to the true crossover rate $J_{e,e'}$. Secondly, by performing simulations we study how various tree structures (e.g., chains and stars) influence the error exponents (Theorem 8).

A. Comparison Between True and Approximate Rates

In Fig. 8, we plot the *true* and *approximate* crossover rates[9] [given in (10) and (14), respectively] for a 4-node symmetric star graph, whose structure is shown in Fig. 9. The zero-mean Gaussian graphical model has a covariance matrix $\boldsymbol{\Sigma}$ such that $\boldsymbol{\Sigma}^{-1}$ is parameterized by $\gamma \in (0, 1/\sqrt{3})$ in the following way: $\boldsymbol{\Sigma}^{-1}(i,i) = 1$ for all i, $\boldsymbol{\Sigma}^{-1}(1,j) = \boldsymbol{\Sigma}^{-1}(j,1) = \gamma$ for all $j = 2,3,4$ and $\boldsymbol{\Sigma}^{-1}(i,j) = 0$ otherwise. By increasing γ, we increase the difference of the mutual information quantities on the edges e and non-edges e'. We see from Fig. 8 that both rates increase as the difference $I(p_e) - I(p_{e'})$ increases. This is in line with our intuition because if $p_{e,e'}$ is such that $I(p_e) - I(p_{e'})$ is large, the crossover rate is also large. We also observe that if $I(p_e) - I(p_{e'})$ is small, the true and approximate rates are close. This is also in line with the assumptions of Theorem 4. When the difference between the mutual information quantities increases, the true and approximate rates separate from each other.

B. Comparison of Error Exponents Between Trees

In Fig. 10, we simulate error probabilities by drawing i.i.d. samples from three $d = 10$ node tree graphs—a chain, a star and a hybrid between a chain and a star as shown in Fig. 9. We then used the samples to learn the structure via the Chow-Liu procedure [5] by solving the MWST in (4). The $d - 1 = 9$

---

[9]This small example has sufficient illustrative power because as we have seen, errors occur locally and only involve triangles.
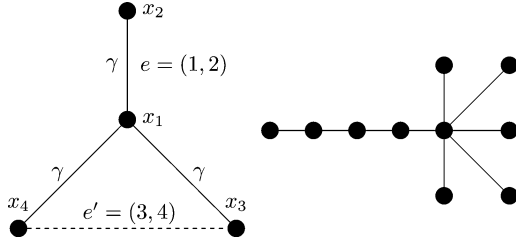
Fig. 9. Left: The symmetric star graphical model used for comparing the true and approximate crossover rates as described in Section VII-A. Right: The structure of a *hybrid* tree graph with $d = 10$ nodes as described in Section VII-B. This is a tree with a length-$d/2$ chain and a order $d/2$ star attached to one of the leaf nodes of the chain.
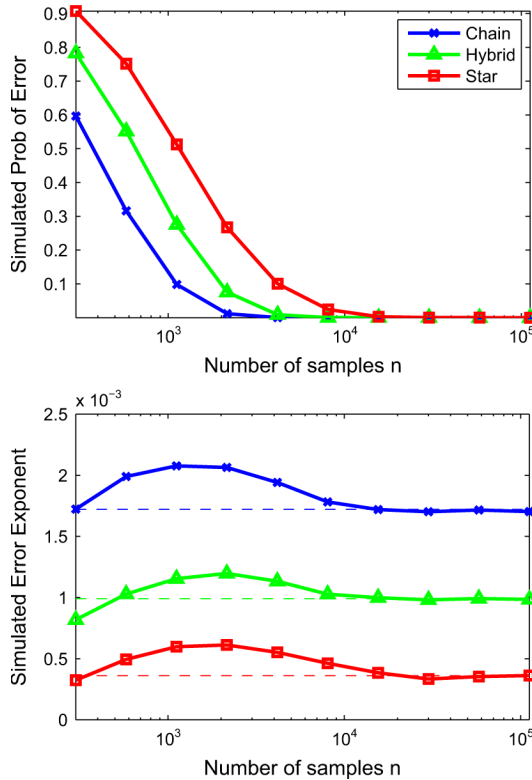

Fig. 10. Simulated error probabilities and error exponents for chain, hybrid and star graphs with fixed $\boldsymbol{\rho}$. The dashed lines show the true error exponent K_p computed numerically using (10) and (11). Observe that the simulated error exponent converges to the true error exponent as $n \to \infty$. The legend applies to both plots.

correlation coefficients were chosen to be equally spaced in the interval $[0.1, 0.9]$ and they were randomly placed on the edges of the three tree graphs. We observe from Fig. 10 that for fixed n, the star and chain have the highest and lowest error probabilities $\mathbb{P}(\mathcal{A}_n)$, respectively. The *simulated error exponents* given by $\{-n^{-1} \log \mathbb{P}(\mathcal{A}_n)\}_{n \in \mathbb{N}}$ also converge to their true values as $n \to \infty$. The exponent associated to the star is higher than that of the chain, which is corroborated by Theorem 8, even though the theorem only applies in the very-noisy case (and for $|\rho_i| < 0.63055$ in the case of the chain). From this experiment, the claim also seems to be true even though the setup is not very-noisy. We also observe that the error exponent of the hybrid is between that of the star and the chain.

VIII. CONCLUSION

Using the theory of large deviations, we have obtained the error exponent associated with learning the structure of a Gaussian tree model. Our analysis in this theoretical paper also answers the fundamental questions as to which set of parameters and which structures result in high and low error exponents. We conclude that Markov chains (respectively, stars) are the easiest (respectively, hardest) structures to learn as they maximize (respectively, minimize) the error exponent. Indeed, our numerical experiments on a variety of Gaussian graphical models validate the theory presented. We believe the intuitive results presented in this paper will lend useful insights for modeling high-dimensional data using tree distributions.

APPENDIX A
PROOF OF THEOREM 1

Proof: This proof borrows ideas from [25]. We assume $m = 4$ (i.e., disjoint edges) for simplicity. The result for $m = 3$ follows similarly. Let $\mathcal{V}' \subset \mathcal{V}$ be a set of $m = 4$ nodes corresponding to node pairs e and e'. Given a subset of node pairs $\mathcal{Y} \subset \mathcal{V}' \times \mathcal{V}'$ such that $(i, i) \in \mathcal{Y}, \forall i \in \mathcal{V}'$, the set of *feasible moments* [4] is defined as

$$\mathcal{M}_{\mathcal{Y}} := \left\{ \boldsymbol{\eta}_{e,e'} \in \mathbb{R}^{|\mathcal{Y}|} : \exists q(\cdot) \in \mathcal{P}(\mathbb{R}^m) \right.$$
$$\left. \text{s.t. } \mathbb{E}_q[x_i x_j] = \eta_{i,j}, \; \forall (i,j) \in \mathcal{Y} \right\}. \quad (33)$$

Let the set of densities with moments $\boldsymbol{\eta}_{e,e'} := \{\eta_{i,j} : (i,j) \in \mathcal{Y}\}$ be denoted as

$$\mathcal{B}_{\mathcal{Y}}(\boldsymbol{\eta}_{e,e'}) := \{q \in \mathcal{P}(\mathbb{R}^m) : \mathbb{E}_q[x_i x_j] = \eta_{i,j}, (i,j) \in \mathcal{Y}\}. \quad (34)$$

Lemma 13 (Sanov's Thm, Contraction Principle [20]): For the event that the empirical moments of the i.i.d. observations $\mathbf{x}^n$ are equal to $\eta_{e,e'} = \{\eta_{i,j} : (i,j) \in \mathcal{Y}\}$, we have the LDP

$$\lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P} \left[\bigcap_{(i,j) \in \mathcal{Y}} \left\{ \mathbf{x}^n : \frac{1}{n} \sum_{k=1}^{n} x_{k,i} x_{k,j} = \eta_{i,j} \right\} \right]$$
$$= \inf_{q_{e,e'} \in \mathcal{B}_{\mathcal{Y}}(\boldsymbol{\eta})} D(q_{e,e'} \| p_{e,e'}). \quad (35)$$

If $\boldsymbol{\eta}_{e,e'} \in \mathcal{M}_{\mathcal{Y}}$, the optimizing pdf $q_{e,e'}^*$ in (35) is given by $q_{e,e'}^*(\mathbf{x}) \propto p_{e,e'}(\mathbf{x}) \exp \left[\sum_{(i,j) \in \mathcal{Y}} \theta_{i,j} x_i x_j \right]$, where the set of constants $\{\theta_{i,j} : (i,j) \in \mathcal{Y}\}$ are chosen such that $q_{e,e'}^* \in \mathcal{B}_{\mathcal{Y}}(\boldsymbol{\eta}_{e,e'})$ given in (34).

From Lemma 13, we conclude that the optimal $q_{e,e'}^*$ in (35) is a Gaussian. Thus, we can restrict our search for the optimal distribution to a search over Gaussians, which are parameterized by means and covariances. The crossover event for mutual information defined in (8) is $\mathcal{C}_{e,e'} = \{\widehat{\rho}_{e'}^2 \geq \widehat{\rho}_e^2\}$, since in the Gaussian case, the mutual information is a monotonic function of the square of the correlation coefficient [cf. (5)]. Thus it suffices to consider $\{\widehat{\rho}_{e'}^2 \geq \widehat{\rho}_e^2\}$, instead of the event involving the mutual information quantities. Let $e = (i, j)$, $e' = (k, l)$ and $\boldsymbol{\eta}_{e,e'} := (\eta_e, \eta_{e'}, \eta_i, \eta_j, \eta_k, \eta_l) \in \mathcal{M}_{\mathcal{Y}} \subset \mathbb{R}^6$ be the moments of $p_{e,e'}$, where $\eta_e := \mathbb{E}[x_i x_j]$ is the covariance of x_i and x_j, and $\eta_i := \mathbb{E}[x_i^2]$ is the variance of x_i (and similarly for the other moments). Now apply the contraction principle [21, Ch. 3] to the
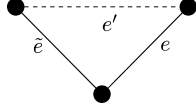
Fig. 11. Illustration for the proof of Corollary 3. The correlation coefficient on the non-edge is $\rho_{e'}$ and satisfies $|\rho_{e'}| = |\rho_e|$ if $|\rho_{\tilde{e}}| = 1$.

Fig. 12. Illustration for the proof of Corollary 3. The unique path between i_0 and i_M is $(i_0, i_1, \ldots, i_M) = \text{Path}(e'; \mathcal{E}_p)$.

continuous map $h : \mathcal{M}_{\mathcal{Y}} \to \mathbb{R}$, given by the difference between the square of correlation coefficients

$$h(\boldsymbol{\eta}_{e,e'}) := \frac{\eta_e^2}{\eta_i \eta_j} - \frac{\eta_{e'}^2}{\eta_k \eta_l}. \tag{36}$$

Following the same argument as in [7, Theorem 2], the equality case dominates $\mathcal{C}_{e,e'}$, i.e., the event $\{\widehat{\rho}_{e'}^2 = \widehat{\rho}_e^2\}$ dominates $\{\widehat{\rho}_{e'}^2 \geq \widehat{\rho}_e^2\}$.[10] Thus, by considering the set $\{\boldsymbol{\eta}_{e,e'} : h(\boldsymbol{\eta}_{e,e'}) = 0\}$, the rate corresponding to $\mathcal{C}_{e,e'}$ can be written as

$$J_{e,e'} = \inf_{\eta_{e,e'} \in \mathcal{M}_{\mathcal{Y}}} \left\{ g(\boldsymbol{\eta}_{e,e'}) : \frac{\eta_e^2}{\eta_i \eta_j} = \frac{\eta_{e'}^2}{\eta_k \eta_l} \right\} \tag{37}$$

where the function $g : \mathcal{M}_{\mathcal{Y}} \subset \mathbb{R}^6 \to [0, \infty)$ is defined as

$$g(\boldsymbol{\eta}_{e,e'}) := \inf_{q_{e,e'} \in \mathcal{B}_{\mathcal{Y}}(\boldsymbol{\eta}_{e,e'})} D(q_{e,e'} \| p_{e,e'}) \tag{38}$$

and the set $\mathcal{B}_{\mathcal{Y}}(\boldsymbol{\eta}_{e,e'})$ is defined in (34). Combining (37) and (38) and the fact that the optimal solution $q_{e,e'}^*$ is Gaussian yields $J_{e,e'}$ as given in the statement of the theorem [cf. (10)].

The second assertion in the theorem follows from the fact that since $p_{e,e}$ satisfies $I(p_e) \neq I(p_{e'})$, we have $|\rho_e| \neq |\rho_{e'}|$ since $I(p_e)$ is a monotonic function in $|\rho_e|$. Therefore, $q_{e,e'}^* \neq p_{e,e'}$ on a set whose (Lebesgue) measure ν is strictly positive. Since $D(q_{e,e'}^* \| p_{e,e'}) = 0$ if and only if $q_{e,e'}^* = p_{e,e'}$ almost everywhere- [ν], this implies that $D(q_{e,e'}^* \| p_{e,e'}) > 0$ [19, Theorem 8.6.1]. ∎

APPENDIX B
PROOF OF COROLLARY 3

Proof: ($\Rightarrow$) Assume that $K_p > 0$. Suppose, to the contrary, that either i) T_p is a forest or ii) $\text{rank}(\boldsymbol{\Sigma}) < d$ and T_p is not a forest. In i), structure estimation of p will be inconsistent (as described in Section II-B), which implies that $K_p = 0$, a contradiction. In ii), since p is a spanning tree, there exists an edge $\tilde{e} \in \mathcal{E}_p$ such that the correlation coefficient $\rho_{\tilde{e}} = \pm 1$ (otherwise $\boldsymbol{\Sigma}$ would be full rank). In this case, referring to Fig. 11 and assuming that $|\rho_e| \in (0, 1)$, the correlation on the non-edge e' satisfies $|\rho_{e'}| = |\rho_e||\rho_{\tilde{e}}| = |\rho_e|$, which implies that $I(p_e) = I(p_{e'})$. Thus, there is no unique maximizer in (4) with the empiricals $\widehat{p}_e$ replaced by p_e. As a result, ML for structure learning via (4) is inconsistent hence $K_p = 0$, a contradiction.

($\Leftarrow$) Suppose both $\boldsymbol{\Sigma} \succ 0$ and T_p *not* a proper forest, i.e., T_p is a spanning tree. Assume, to the contrary, that $K_p = 0$. Then from [7], $I(p_e) = I(p_{e'})$ for some $e' \notin \mathcal{E}_p$ and some $e \in \text{Path}(e'; \mathcal{E}_p)$. This implies that $|\rho_e| = |\rho_{e'}|$. Let $e' = (i_0, i_M)$ be a non-edge and let the unique path from node i_0 to node i_M be $(i_0, i_1, \ldots, i_M)$ for some $M \geq 2$. See Fig. 12. Then, $|\rho_{e'}| = |\rho_{i_0, i_M}| = |\rho_{i_0, i_1}||\rho_{i_1, i_2}| \ldots |\rho_{i_{M-1}, i_M}|$. Suppose, without loss
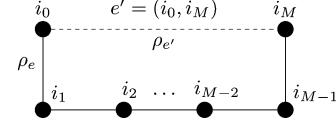
of generality, that edge $e = (i_0, i_1)$ is such that $|\rho_{e'}| = |\rho_e|$ holds, then we can cancel $|\rho_{e'}|$ and $|\rho_{i_0, i_1}|$ on both sides to give $|\rho_{i_1, i_2}||\rho_{i_2, i_3}| \ldots |\rho_{i_{M-1}, i_M}| = 1$. Cancelling $\rho_{e'}$ is legitimate because we assumed that $\rho_{e'} \neq 0$ for all $e' \in \mathcal{V} \times \mathcal{V}$, because p is a *spanning* tree. Since each correlation coefficient has magnitude not exceeding 1, this means that each correlation coefficient has magnitude 1, i.e., $|\rho_{i_1, i_2}| = \cdots = |\rho_{i_{M-1}, i_M}| = 1$. Since the correlation coefficients equal to ± 1, the submatrix of the covariance matrix $\boldsymbol{\Sigma}$ containing these correlation coefficients is not positive definite. Therefore by Sylvester's condition, the covariance matrix $\boldsymbol{\Sigma} \not\succ 0$, a contradiction. Hence, $K_p > 0$. ∎

APPENDIX C
PROOF OF THEOREM 4

Proof: We first assume that e and e' do not share a node. The approximation of the KL-divergence for Gaussians can be written as in (12). We now linearize the constraint set $L_{\boldsymbol{\Delta}}(p_{e,e'})$ as defined in (13). Given a positive definite covariance matrix $\boldsymbol{\Sigma}_e \in \mathbb{R}^{2 \times 2}$, to simplify the notation, let $I(\boldsymbol{\Sigma}_e) = I(\mathcal{N}(\mathbf{x}; \mathbf{0}, \boldsymbol{\Sigma}_e))$ be the mutual information of the two random variables with covariance matrix $\boldsymbol{\Sigma}_e$. We now perform a first-order Taylor expansion of the mutual information around $\boldsymbol{\Sigma}_e$. This can be expressed as

$$I(\boldsymbol{\Sigma}_e + \boldsymbol{\Delta}_e) = I(\boldsymbol{\Sigma}_e) + \text{Tr}\left(\nabla_{\boldsymbol{\Sigma}_e} I(\boldsymbol{\Sigma}_e)^T \boldsymbol{\Delta}_e\right) + o(\boldsymbol{\Delta}_e). \tag{39}$$

Recall that the Taylor expansion of log-det [26] is $\log \det(\mathbf{A}) = \log \det(\mathbf{B}) + \langle \mathbf{A} - \mathbf{B}, \mathbf{B}^{-1} \rangle + o(\|\mathbf{A} - \mathbf{B}\|_F)$, with the notation $\langle \mathbf{A} - \mathbf{B}, \mathbf{B}^{-1} \rangle = \text{Tr}((\mathbf{A} - \mathbf{B})\mathbf{B}^{-1})$. Using this result we can conclude that the gradient of I with respect to $\boldsymbol{\Sigma}_e$ in the above expansion (39) can be simplified to give the matrix

$$\nabla_{\boldsymbol{\Sigma}_e} I(\boldsymbol{\Sigma}_e) = -\frac{1}{2} \begin{pmatrix} 0 & [\boldsymbol{\Sigma}_e^{-1}]_{\text{od}} \\ [\boldsymbol{\Sigma}_e^{-1}]_{\text{od}} & 0 \end{pmatrix} \tag{40}$$

where $[\mathbf{A}]_{\text{od}}$ is the (unique) off-diagonal element of the 2×2 symmetric matrix $\mathbf{A}$. By applying the same expansion to $I(\boldsymbol{\Sigma}_{e'} + \boldsymbol{\Delta}_{e'})$, we can express the linearized constraint as

$$\langle \mathbf{M}, \boldsymbol{\Delta} \rangle = \text{Tr}(\mathbf{M}^T \boldsymbol{\Delta}) = I(\boldsymbol{\Sigma}_e) - I(\boldsymbol{\Sigma}_{e'}), \tag{41}$$

where the symmetric matrix $\mathbf{M} = \mathbf{M}(\boldsymbol{\Sigma}_{e,e'})$ is defined in the following fashion: $\mathbf{M}(i,j) = 1/2[\boldsymbol{\Sigma}_e^{-1}]_{\text{od}}$ if $(i,j) = e$, $\mathbf{M}(i,j) = -1/2[\boldsymbol{\Sigma}_{e'}^{-1}]_{\text{od}}$ if $(i,j) = e'$ and $\mathbf{M}(i,j) = 0$ otherwise.

Thus, the problem reduces to minimizing (over $\boldsymbol{\Delta}$) the approximate objective in (12) subject to the linearized constraints in (41). This is a least-squares problem. By using the matrix derivative identities $\nabla_{\boldsymbol{\Delta}} \text{Tr}(\mathbf{M}\boldsymbol{\Delta}) = \mathbf{M}$ and $\nabla_{\boldsymbol{\Delta}} \text{Tr}((\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta})^2) = 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1}$, we can solve for the optimizer $\boldsymbol{\Delta}^*$ yielding:

$$\boldsymbol{\Delta}^* = \frac{I(\boldsymbol{\Sigma}_e) - I(\boldsymbol{\Sigma}_{e'})}{(\text{Tr}(\mathbf{M}\boldsymbol{\Sigma}))^2} \boldsymbol{\Sigma}\mathbf{M}\boldsymbol{\Sigma}. \tag{42}$$

---

[10]This is also intuitively true because the most likely way the error event $\mathcal{C}_{e,e'}$ occurs is when equality holds, i.e., $\{\widehat{\rho}_{e'}^2 = \widehat{\rho}_e^2\}$.

Substituting the expression for $\mathbf{\Delta}^*$ into (12) yields

$$\widetilde{J}_{e,e'} = \frac{(I(\mathbf{\Sigma}_e) - I(\mathbf{\Sigma}_{e'}))^2}{4\,\mathrm{Tr}((\mathbf{M}\mathbf{\Sigma})^2)} = \frac{(I(p_e) - I(p_{e'}))^2}{4\,\mathrm{Tr}((\mathbf{M}\mathbf{\Sigma})^2)}. \qquad (43)$$

Comparing (43) to our desired result (16), we observe that problem now reduces to showing that $\mathrm{Tr}((\mathbf{M}\mathbf{\Sigma})^2) = 1/2\mathrm{Var}(s_e - s_{e'})$. To this end, we note that for Gaussians, the information density is $s_e(x_i, x_j) = -1/2\log(1 - \rho_e^2) - [\mathbf{\Sigma}_e^{-1}]_{\mathrm{od}} x_i x_j$. Since the first term is a constant, it suffices to compute $\mathrm{Var}([\mathbf{\Sigma}_e^{-1}]_{\mathrm{od}} x_i x_j - [\mathbf{\Sigma}_{e'}^{-1}]_{\mathrm{od}} x_k x_l)$. Now, we define the matrices

$$\mathbf{C} := \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}, \quad \mathbf{C}_1 := \begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \mathbf{C}_2 := \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}$$

and use the following identity for the normal random vector $(x_i, x_j, x_k, x_l) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$

$$\mathrm{Cov}(a x_i x_j, b x_k x_l) = 2ab \cdot \mathrm{Tr}(\mathbf{C}_1 \mathbf{\Sigma} \mathbf{C}_2 \mathbf{\Sigma}), \quad \forall a, b \in \mathbb{R}$$

and the definition of $\mathbf{M}$ to conclude that $\mathrm{Var}(s_e - s_{e'}) = 2\mathrm{Tr}((\mathbf{M}\mathbf{\Sigma})^2)$. This completes the proof for the case when e and e' do not share a node. The proof for the case when e and e' share a node proceeds along exactly the same lines with a slight modification of the matrix $\mathbf{M}$. ∎

APPENDIX D
PROOF OF LEMMA 5

Proof: Denoting the correlation coefficient on edge e and non-edge e' as ρ_e and $\rho_{e'}$, respectively, the approximate crossover rate can be expressed as

$$\widetilde{J}(\rho_e, \rho_{e'}) = \frac{A(\rho_e^2, \rho_{e'}^2)}{B(\rho_e^2, \rho_{e'}^2)} \qquad (44)$$

where the numerator and the denominator are defined as

$$A(\rho_e^2, \rho_{e'}^2) := \left[\frac{1}{2} \log\left(\frac{1 - \rho_{e'}^2}{1 - \rho_e^2} \right) \right]^2,$$

$$B(\rho_e^2, \rho_{e'}^2) := \frac{2(\rho_{e'}^4 + \rho_{e'}^2)}{(1 - \rho_{e'}^2)^2} + \frac{2(\rho_e^4 + \rho_e^2)}{(1 - \rho_e^2)^2}$$
$$- \frac{4\rho_{e'}^2(\rho_e^2 + 1)}{(1 - \rho_{e'}^2)(1 - \rho_e^2)}.$$

The evenness result follows from A and B because $\widetilde{J}(\rho_e, \rho_{e'})$ is, in fact a function of $(\rho_e^2, \rho_{e'}^2)$. To simplify the notation, we make the following substitutions: $x := \rho_{e'}^2$ and $y := \rho_e^2$. Now we apply the quotient rule to (44). Defining $\mathcal{R} := \{(x, y) \in \mathbb{R}^2 : y \in (0, 1), x \in (0, y)\}$, it suffices to show that

$$C(x, y) := B(x, y) \frac{\partial A(x, y)}{\partial x} - A(x, y) \frac{\partial B(x, y)}{\partial x} \leq 0$$

for all $(x, y) \in \mathcal{R}$. Upon simplification, we have

$$C(x, y) = \frac{\log\left(\frac{1-x}{1-y} \right) \left[\log\left(\frac{1-x}{1-y} \right) C_1(x, y) + C_2(x, y) \right]}{2(1 - y)^2 (1 - x)^3}$$

where $C_1(x, y) := y^2 x - 6xy - 1 - 2y + 3y^2$ and $C_2(x, y) := 2x^2 y - 6x^2 + 2x - 2y^2 x + 8xy - 2y - 2y^2$. Since $x < y$, the logs in $C(x, y)$ are positive, i.e., $\log(1 - x/1 - y) > 0$, so it suffices to show that

$$\log\left(\frac{1 - x}{1 - y} \right) C_1(x, y) + C_2(x, y) \leq 0$$

for all $(x, y) \in \mathcal{R}$. By using the inequality $\log(1 + t) \leq t$ for all $t > -1$, it again suffices to show that

$$C_3(x, y) := (y - x) C_1(x, y) + (1 - y) C_2(x, y) \leq 0.$$

Now upon simplification, $C_3(x, y) = 3y^3 x - 19y^2 x - 3y - 2y^2 + 5y^3 - 3y^2 x^2 + 14x^2 y + 3x + 8xy - 6x^2$, and this polynomial is equal to zero in $\overline{\mathcal{R}}$ (the closure of $\mathcal{R}$) iff $x = y$. At all other points in $\mathcal{R}$, $C_3(x, y) < 0$. Thus, the derivative of $\widetilde{J}(\rho_e, \rho_{e'})$ with respect to $\rho_{e'}$ is indeed strictly negative on $\mathcal{R}$. Keeping ρ_e fixed, the function $\widetilde{J}(\rho_e, \rho_{e'})$ is monotonically decreasing in $\rho_{e'}^2$ and hence $|\rho_{e'}|$. Statements (c) and (d) follow along exactly the same lines and are omitted for brevity. ∎

APPENDIX E
PROOFS OF THEOREM 8 AND COROLLARY 10

Proof: Proof of $T_{p_{\min}(\rho)} = T_{\mathrm{star}}(d)$: Sort the correlation coefficients in decreasing order of magnitude and relabel the edges such that $|\rho_{e_1}| \geq \cdots \geq |\rho_{e_{d-1}}|$. Then, from Lemma 5(b), the set of crossover rates for the star graph is given by $\{\widetilde{J}(\rho_{e_1}, \rho_{e_1} \rho_{e_2})\} \cup \{\widetilde{J}(\rho_{e_i}, \rho_{e_i} \rho_{e_1}) : i = 2, \ldots, d-1\}$. For edge e_1, the correlation coefficient ρ_{e_2} is the largest correlation coefficient (hence, results in the smallest rate). For all other edges $\{e_i : i \geq 2\}$, the correlation coefficient ρ_{e_1} is the largest possible correlation coefficient (and hence results in the smallest rate). Since each member in the set of crossovers is the minimum possible, the minimum of these crossover rates is also the minimum possible among all tree graphs. ∎

Before we prove part (b), we present some properties of the edge weights $W(\rho_i, \rho_j)$, defined in (20).

Lemma 14 (Properties of Edge Weights): Assume that all the correlation coefficients are bounded above by ρ_{crit}, i.e., $|\rho_i| \leq \rho_{\mathrm{crit}}$. Then $W(\rho_i, \rho_j)$ satisfies the following properties:
  (a) The weights are symmetric, i.e., $W(\rho_i, \rho_j) = W(\rho_j, \rho_i)$.
  (b) $W(\rho_i, \rho_j) = \widetilde{J}(\min\{|\rho_i|, |\rho_j|\}, \rho_i \rho_j)$, where $\widetilde{J}$ is the approximate crossover rate given in (44).
  (c) If $|\rho_i| \geq |\rho_j| \geq |\rho_k|$, then

$$W(\rho_i, \rho_k) \leq \min\{W(\rho_i, \rho_j), W(\rho_j, \rho_k)\}. \qquad (45)$$

  (d) If $|\rho_1| \geq \cdots \geq |\rho_{d-1}|$, then

$$W(\rho_i, \rho_j) \leq W(\rho_i, \rho_{i+1}), \quad \forall j \geq i + 1 \qquad (46a)$$
$$W(\rho_i, \rho_j) \leq W(\rho_i, \rho_{i-1}), \quad \forall j \leq i - 1. \qquad (46b)$$

Proof: Claim (a) follows directly from the definition of $\widetilde{J}$ in (20). Claim (b) also follows from the definition of $\widetilde{J}$ and its monotonicity property in Lemma 5(d). Claim (c) follows by first using Claim (b) to establish that the right-hand side of (45) equals $\min\{\widetilde{J}(\rho_j, \rho_j \rho_i), \widetilde{J}(\rho_k, \rho_k \rho_j)\}$ since $|\rho_i| \geq |\rho_j| \geq$
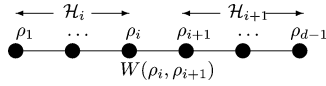
Fig. 13. Illustration of the proof of Theorem 8. Let $|\rho_1| \geq \cdots \geq |\rho_{d-1}|$. The figure shows the chain H^*_{chain} (in the line graph domain) where the correlation coefficients $\{\rho_i\}$ are placed in decreasing order.
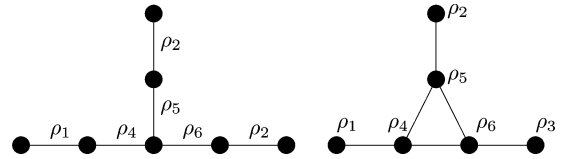

Fig. 14. A 7-node tree T and its line graph $H = \mathcal{L}(T)$ are shown in the left and right figures, respectively. In this case, $H \setminus H^*_{\text{chain}} = \{(1,4),(2,5),(4,6),(3,6)\}$ and $H^*_{\text{chain}} \setminus H = \{(1,2),(2,3),(3,4)\}$. Equation (49) holds because from (46) $W(\rho_1, \rho_4) \leq W(\rho_1, \rho_2)$, $W(\rho_2, \rho_5) \leq W(\rho_2, \rho_3)$, etc. and also if $a_i \leq b_i$ for $i \in \mathcal{I}$ (for finite $\mathcal{I}$), then $\min_{i \in \mathcal{I}} a_i \leq \min_{i \in \mathcal{I}} b_i$.

$|\rho_k|$. By the same argument, the left-hand side of (45), equals $\widetilde{J}(\rho_k, \rho_k \rho_i)$. Now we have

$$\widetilde{J}(\rho_k, \rho_k \rho_i) \leq \widetilde{J}(\rho_j, \rho_j \rho_i), \quad \widetilde{J}(\rho_k, \rho_k \rho_i) \leq \widetilde{J}(\rho_k, \rho_k \rho_j) \tag{47}$$

where the first and second inequalities follow from Lemmas 5(c) and 5(b), respectively. This establishes (45). Claim (d) follows by applying Claim (c) recursively. ∎

Proof: Proof of $T_{p_{\max}(\boldsymbol{\rho})} = T_{\text{chain}}(d)$: Assume, without loss of generality, that $|\rho_{e_1}| \geq \cdots \geq |\rho_{e_{d-1}}|$ and we also abbreviate ρ_{e_i} as ρ_i for all $i = 1, \ldots, d-1$. We use the idea of line graphs introduced in Section VI-A and Lemma 14. Recall that $\mathcal{L}(\mathcal{T}^d)$ is the set of line graphs of spanning trees with d nodes. From (28), the line graph for the structure of the best distribution $p_{\max,\boldsymbol{\rho}}$ for learning in (25) is

$$H_{\max,\boldsymbol{\rho}} := \arg\max_{H \in \mathcal{L}(\mathcal{T}^d)} \min_{(i,j) \in H} W(\rho_i, \rho_j). \tag{48}$$

We now argue that the length $d - 1$ chain H^*_{chain} (in the line graph domain) with correlation coefficients $\{\rho_i\}_{i=1}^{d-1}$ arranged in decreasing order on the nodes (see Fig. 13) is the line graph that optimizes (48). Note that the edge weights of H^*_{chain} are given by $W(\rho_i, \rho_{i+1})$ for $1 \leq i \leq d-2$. Consider any other line graph $H \in \mathcal{L}(\mathcal{T}^d)$. Then we claim that

$$\min_{(i,j) \in H \setminus H^*_{\text{chain}}} W(\rho_i, \rho_j) \leq \min_{(i,j) \in H^*_{\text{chain}} \setminus H} W(\rho_i, \rho_j). \tag{49}$$

To prove (49), note that any edge $(i, j) \in H^*_{\text{chain}} \setminus H$ is *consecutive*, i.e., of the form $(i, i+1)$. Fix any such $(i, i+1)$. Define the two subchains of H^*_{chain} as $\mathcal{H}_i := \{(1, 2), \ldots, (i-1, i)\}$ and $\mathcal{H}_{i+1} := \{(i+1, i+2), \ldots, (d-2, d-1)\}$ (see Fig. 13). Also, let $\mathcal{V}(\mathcal{H}_i) := \{1, \ldots, i\}$ and $\mathcal{V}(\mathcal{H}_{i+1}) := \{i+1, \ldots, d-1\}$ be the nodes in subchains $\mathcal{H}_i$ and $\mathcal{H}_{i+1}$, respectively. Because $(i, i+1) \notin H$, there is a set of edges (called cut set edges) $\mathcal{S}_i := \{(j, k) \in H : j \in \mathcal{V}(\mathcal{H}_i), k \in \mathcal{V}(\mathcal{H}_{i+1})\}$ to ensure that the line graph H remains connected.[11] The edge weight of each cut set edge $(j, k) \in \mathcal{S}_i$ satisfies $W(\rho_j, \rho_k) \leq W(\rho_i, \rho_{i+1})$ by (46) because $|j - k| \geq 2$ and $j \leq i$ and $k \geq i+1$. By considering all cut set edges $(j, k) \in \mathcal{S}_i$ for fixed i and subsequently all $(i, i+1) \in H^*_{\text{chain}} \setminus H$, we establish (49). It follows that

$$\min_{(i,j) \in H} W(\rho_i, \rho_j) \leq \min_{(i,j) \in H^*_{\text{chain}}} W(\rho_i, \rho_j) \tag{50}$$

because the other edges in H and H^*_{chain} in (49) are common. See Fig. 14 for an example to illustrate (49).

Since the chain line graph H^*_{chain} achieves the maximum bottleneck edge weight, it is the optimal line graph, i.e., $H_{\max,\boldsymbol{\rho}} = H^*_{\text{chain}}$. Furthermore, since the line graph of a chain

is a chain, the best structure $T_{p_{\max}(\rho)}$ is also a chain and we have established (30). The best distribution is given by the chain with the correlations placed in decreasing order, establishing Corollary 10. ∎

References

[1] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 1988.

[2] D. Geiger and D. Heckerman, "Learning Gaussian networks," in *Uncertainty in Artificial Intelligence (UAI)*. San Francisco, CA: Morgan Kaufmann, 1994.

[3] S. Lauritzen, *Graphical Models*. Oxford, U.K.: Oxford Univ. Press, 1996.

[4] M. J. Wainwright and M. I. Jordan, *Graphical Models, Exponential Families, and Variational Inference, vol. of Foundations and Trends in Machine Learning*. Boston, MA: Now, 2008.

[5] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inf. Theory*, vol. 14, no. 3, pp. 462–467, May 1968.

[6] C. K. Chow and T. Wagner, "Consistency of an estimate of tree-dependent probability distributions," *IEEE Trans. Inf. Theory*, vol. 19, no. 3, pp. 369–371, May 1973.

[7] V. Y. F. Tan, A. Anandkumar, L. Tong, and A. S. Willsky, "A large-deviation analysis for the maximum likelihood learning of tree structures," in *Proc. IEEE Int. Symp. Inf. Theory*, Seoul, Korea, Jul. 2009 [Online]. Available: http://arxiv.org/abs/0905.0940

[8] D. M. Chickering, "Learning equivalence classes of Bayesian network structures," *J. Mach. Learn. Res.*, vol. 2, pp. 445–498, 2002.

[9] M. Dudik, S. J. Phillips, and R. E. Schapire, "Performance guarantees for regularized maximum entropy density estimation," in *Proc. Conf. Learn. Theory (COLT)*, 2004.

[10] N. Meinshausen and P. Buehlmann, "High dimensional graphs and variable selection with the Lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436–1462, 2006.

[11] M. J. Wainwright, P. Ravikumar, and J. D. Lafferty, "High-dimensional graphical model selection using l_1-regularized logistic regression," in *Neural Information Processing Systems (NIPS)*. Cambridge, MA: MIT Press, 2006.

[12] N. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," in *Proc. IEEE Int. Symp. on Inf. Theory*, Toronto, Canada, Jul. 2008.

[13] O. Zuk, S. Margel, and E. Domany, "On the number of samples needed to learn the correct structure of a Bayesian network," in *Uncertainly in Artificial Intelligence (UAI)*. Arlington, VA: AUAI Press, 2006.

[14] A. Montanari and J. A. Pereira, "Which graphical models are difficult to learn," in *Neural Information Processing Systems (NIPS)*. Cambridge, MA: MIT Press, 2009.

[15] A. Dempster, "Covariance selection," *Biometrics*, vol. 28, pp. 157–175, 1972.

[11]The line graph $H = \mathcal{L}(G)$ of a connected graph G is connected. In addition, any $H \in \mathcal{L}(\mathcal{T}^d)$ must be a claw-free, block graph [24, Th. 8.5].

[16] A. d'Aspremont, O. Banerjee, and L. El Ghaoui, "First-order methods for sparse covariance selection," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 1, pp. 56–66, Feb. 2008.

[17] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electron. J. Statist.*, vol. 2, pp. 494–515, 2008.

[18] S. Borade and L. Zheng, "Euclidean information theory," in *Proc. Allerton Conf.*, Monticello, IL, Sep. 2007.

[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley-Intersci., 2006.

[20] J.-D. Deuschel and D. W. Stroock, *Large Deviations*. Providence, RI: Amer. Math. Soc., 2000.

[21] F. Den Hollander, *Large Deviations (Fields Institute Monographs)*. Providence, RI: Amer. Math. Soc., 2000.

[22] D. B. West, *Introduction to Graph Theory*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 2000.

[23] S. Verdu, "Spectral efficiency in the wideband regime," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, Jun. 2002.

[24] F. Harary, *Graph Theory*. Reading, MA: Addison-Wesley, 1972.

[25] S. Shen, "Large deviation for the empirical correlation coefficient of two Gaussian random variables," *Acta Math. Scientia*, vol. 27, no. 4, pp. 821–828, Oct. 2007.

[26] M. Fazel, H. Hindi, and S. P. Boyd, "Log-det heuristic for matrix rank minimization with applications with applications to Hankel and Euclidean distance metrics," in *Proc. Amer. Control Conf.*, 2003.

Animashree Anandkumar (S'02–M'09) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Madras, in 2004, and the Ph.D. degree in electrical engineering from Cornell University, Ithaca, NY, in 2009.

She is currently a Postdoctoral Researcher with the Stochastic Systems Group, the Massachusetts Institute of Technology (MIT), Cambridge. Her research interests are in the area of statistical-signal processing, information theory and networking with a focus on distributed inference, learning and fusion in graphical models.

Dr. Anandkumar received the 2008 IEEE Signal Processing Society (SPS) Young Author award for her paper coauthored with L. Tong, which appeared in the IEEE TRANSACTIONS ON SIGNAL PROCESSING. She is the recipient of the Fran Allen IBM Ph.D. fellowship 2008–2009, presented annually to one female Ph.D. student in conjunction with the IBM Ph.D. Fellowship Award. She was named a finalist for the Google Anita-Borg Scholarship 2007–2008 and also received the Student Paper Award at the 2006 International Conference on Acoustic, Speech, and Signal Processing (ICASSP). She has served as a reviewer for IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON INFORMATION THEORY, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and IEEE SIGNAL PROCESSING LETTERS.

Vincent Y. F. Tan (S'07) received the B.A. (with first-class honors) and the M.Eng. (with distinction) degrees in Electrical and Information Sciences Tripos (EIST) from Sidney Sussex College, Cambridge University, U.K.

He is currently pursuing the Ph.D. degree in electrical engineering with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge. He was a Research Engineer with the Defence Science Organization National Laboratories, Singapore, during 2005–2006; Research Officer with the Institute for Infocomm Research, Singapore, during 2006–2007; Teaching Assistant with the National University of Singapore in 2006; and Research Intern with Microsoft Research in 2008 and 2009. His research interests include statistical signal processing, probabilistic graphical models, machine learning, and information theory.

Mr. Tan received the Public Service Commission Overseas Merit Scholarship in 2001 and the National Science Scholarship from the Agency for Science Technology and Research (A*STAR) in 2006. In 2005, he received the Charles Lamb Prize, a Cambridge University Engineering Department prize awarded annually to the candidate who demonstrates the greatest proficiency in the EIST.

Alan S. Willsky (S'70–M'73–SM'82–F'86) received the S.B. degree in 1969 and the Ph.D. degree in 1973 from the Department of Aeronautics and Astronautics, Massachusetts Institute of Technology (MIT), Cambridge.

He joined the MIT faculty in 1973 and is the Edwin Sibley Webster Professor of Electrical Engineering and the Director of the Laboratory for Information and Decision Systems. He was a founder of Alphatech, Inc. and Chief Scientific Consultant, a role in which he continues at BAE Systems Advanced Information Technologies. From 1998 to 2002, he served on the U.S. Air Force Scientific Advisory Board. He has delivered numerous keynote addresses and is coauthor of "Signals and Systems" (Englewood Cliffs, NJ: Prentice-Hall, 1996). His research interests are in the development and application of advanced methods of estimation, machine learning, and statistical signal and image processing.

Dr. Willsky received several awards including the 1975 American Automatic Control Council Donald P. Eckman Award, the 1979 ASCE Alfred Noble Prize, the 1980 IEEE Browder J. Thompson Memorial Award, the IEEE Control Systems Society Distinguished Member Award in 1988, the 2004 IEEE Donald G. Fink Prize Paper Award, and Doctorat Honoris Causa from Universit de Rennes in 2005. He and his students, colleagues, and Postdoctoral Associates have also received a variety of Best Paper Awards at various conferences and for papers in journals, including the 2001 IEEE Conference on Computer Vision and Pattern Recognition, the 2003 Spring Meeting of the American Geophysical Union, the 2004 Neural Information Processing Symposium, Fusion 2005, and the 2008 award from the journal *Signal Processing* for Outstanding Paper in 2007.