

# How Do the Structure and the Parameters of Gaussian Tree Models Affect Structure Learning?

Vincent Y. F. Tan, Animashree Anandkumar and Alan S. Willsky

Stochastic Systems Group,  
Laboratory for Information and Decision Systems,  
Massachusetts Institute of Technology.  
Email: {vtan,animakum,willsky}@mit.edu

**Abstract**—The problem of learning tree-structured Gaussian graphical models from 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 universally, the extremal tree structures which maximize and minimize the error exponent are the star and the Markov chain for any fixed set of correlation coefficients on the edges of the tree. 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**—Structure learning, Gaussian graphical models, Large deviations, Error exponents, Tree distributions, Euclidean information theory.

## 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 two key questions in this paper. (i) Can we characterize the error exponent for structure learning by the ML algorithm for a 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? 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 non-convex 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 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 Gaussian models (Proposition 7), where  $d$  is the number of nodes.

This work is supported in part by a AFOSR through Grant FA9550-08-1-1080, in part by a MURI funded through ARO Grant W911NF-06-1-0076 and in part under a MURI through AFOSR Grant FA9550-06-1-0324. Vincent Tan is also funded by A\*STAR, Singapore.

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* and *Markov chain* graphs maximize and minimize the error exponent respectively (Theorem 8). Therefore, extremal tree structures in terms of the diameter are *also* extremal trees for learning Gaussian tree distributions. These results are also *universal* in the sense that they hold for every set of correlation coefficients. 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 non-neighbor pairs.

## 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  $\ell_1$  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 Zuk et al. [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 Zuk et al. [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.

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.

Because of space constraints, all the results in this short paper are stated without proof. The reader may refer to the full version of this paper (found at <http://arxiv.org/abs/0909.5216>) for the complete details.

## II. PRELIMINARIES AND PROBLEM STATEMENT

### A. Basics of Undirected Gaussian Graphical Models

*Undirected graphical models* or *Markov random fields*<sup>1</sup> (MRFs) are probability distributions that factorize according to given undirected graphs [3]. In this paper, we focus solely on

<sup>1</sup>In this paper, we use the terms “graphical models” and “Markov random fields” interchangeably.

*spanning trees* (*i.e.*, undirected, acyclic, connected graphs). A  $d$ -dimensional random vector  $\mathbf{x} = [x_1, \dots, x_d]^T \in \mathbb{R}^d$  is said to be *Markov* on a spanning tree  $T_p = (\mathcal{V}, \mathcal{E}_p)$  with vertex (or node) set  $\mathcal{V} = \{1, \dots, d\}$  and edge set  $\mathcal{E}_p \subset \binom{\mathcal{V}}{2}$  if its distribution  $p(\mathbf{x})$  satisfies the (local) Markov property:

$$p(x_i | x_{\mathcal{V} \setminus \{i\}}) = p(x_i | x_{\text{nbnd}(i)}), \quad \forall i \in \mathcal{V}, \quad (1)$$

where  $\text{nbnd}(i) := \{j \in \mathcal{V} : (i, j) \in \mathcal{E}_p\}$  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  $\mathcal{E}_p$ :

$$p(\mathbf{x}) = \prod_{i \in \mathcal{V}} p_i(x_i) \prod_{(i,j) \in \mathcal{E}_p} \frac{p_{i,j}(x_i, x_j)}{p_i(x_i)p_j(x_j)}. \quad (2)$$

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 *Gaussian-Markov random field* (GMRF) with known zero mean<sup>2</sup> and unknown positive definite covariance matrix  $\Sigma \succ 0$ . Thus,  $p(\mathbf{x})$  can be written as

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} \mathbf{x}^T \Sigma^{-1} \mathbf{x}\right). \quad (3)$$

We also use the notation  $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \Sigma)$  as a shorthand for (3). For Gaussian graphical models, it is known that the fill-pattern of the inverse covariance matrix  $\Sigma^{-1}$  encodes the structure of  $p(\mathbf{x})$  [3], *i.e.*,  $\Sigma^{-1}(i, j) = 0$  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  $\Sigma^{-1}$ ), we are provided with a set of  $d$ -dimensional samples  $\mathbf{x}^n := \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  drawn from  $p$ , where  $\mathbf{x}_k := [x_{k,1}, \dots, x_{k,d}]^T \in \mathbb{R}^d$ .

### 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 [14] between  $p_1$  and  $p_2$ , the estimation of the structure of  $p$  is given by the optimization problem<sup>3</sup>

$$\mathcal{E}_{\text{cl}}(\mathbf{x}^n) := \underset{\mathcal{E}_q: q \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d)}{\text{argmin}} D(\hat{p} || q), \quad (4)$$

where  $\hat{p}(\mathbf{x}) := \mathcal{N}(\mathbf{x}; \mathbf{0}, \hat{\Sigma})$  and  $\hat{\Sigma} := n^{-1} \sum_{k=1}^n \mathbf{x}_k \mathbf{x}_k^T$  is the *empirical covariance matrix* (or *sample covariance matrix*). Given  $\hat{p}$ , and exploiting the fact that  $q$  in (4) factorizes according to a tree as in (2), Chow and Liu [5] showed that

<sup>2</sup>Our results also extend to the scenario where the mean of the Gaussian is unknown and has to be estimated from the samples.

<sup>3</sup>Note that it is unnecessary to impose the Gaussianity constraint on  $q$  in (4). 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.

the optimization for the optimal edge set in (4) can be reduced to a MWST problem:

$$\mathcal{E}_{\text{cl}}(\mathbf{x}^n) = \underset{\mathcal{E}_q: q \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d)}{\text{argmax}} \sum_{e \in \mathcal{E}_q} I(\hat{p}_e), \quad (5)$$

where the edge weights are the *empirical mutual information quantities* [14] given by<sup>4</sup>

$$I(\hat{p}_e) = -\frac{1}{2} \log(1 - \hat{\rho}_e^2), \quad (6)$$

and where the *empirical correlation coefficients* are given by  $\hat{\rho}_e = \hat{\rho}_{i,j} := \widehat{\Sigma}(i,j) / (\widehat{\Sigma}(i,i)\widehat{\Sigma}(j,j))^{1/2}$ . Note that in (5), the estimated edge set  $\mathcal{E}_{\text{cl}}(\mathbf{x}^n)$  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 (5) produces a spanning tree as all the mutual information quantities in (6) will be non-zero. If  $T_p$  were allowed to be a *forest* (a tree that is not connected), the estimation of  $\mathcal{E}_p$  will be inconsistent because the learned edge set will definitely 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

$$\mathcal{A}_n := \{\mathbf{x}^n : \mathcal{E}_{\text{cl}}(\mathbf{x}^n) \neq \mathcal{E}_p\}, \quad (7)$$

where  $\mathcal{E}_{\text{cl}}(\mathbf{x}^n)$  is the edge set of the Chow-Liu ML estimator in (4). 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

$$K_p := \lim_{n \rightarrow \infty} -\frac{1}{n} \log \mathbb{P}(\mathcal{A}_n), \quad (8)$$

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 in (8) exists in Section III. 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*

$$\mathcal{C}_{e,e'} := \{\mathbf{x}^n : I(\hat{p}_e) \leq I(\hat{p}_{e'})\}. \quad (9)$$

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

<sup>4</sup>Our notation for the mutual information between two random variables differs from the conventional one in [14].

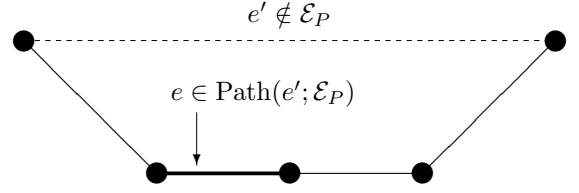


Fig. 1. If the error event occurs during the learning process, an edge  $e \in \text{Path}(e'; \mathcal{E}_p)$  is replaced by a non-edge  $e' \notin \mathcal{E}_p$  in the original model. We identify the crossover event that has the minimum rate  $J_{e,e'}$  and its rate is  $K_p$ .

*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

$$J_{e,e'} := \lim_{n \rightarrow \infty} -\frac{1}{n} \log \mathbb{P}(\mathcal{C}_{e,e'}). \quad (10)$$

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

$$J_{e,e'} = \inf_{q \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^m)} \left\{ D(q || p_{e,e'}) : I(q_e) = I(q_{e'}) \right\}. \quad (11)$$

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

The proof involves an application of Sanov's Theorem [15, Ch. 3], and the contraction principle [16, Ch. 4] in large deviations theory, together with the maximum entropy principle [14, Ch. 12]. We remark that the proof is different from the corresponding result in [7].

Theorem 1 says that in order to compute the crossover rate  $J_{e,e'}$ , we can restrict our attention to a problem that only involves 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 (8). 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, 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 [7]):* The error exponent for structure learning of tree-structured Gaussian

graphical models, defined in (8), is given as

$$K_p = \min_{e' \notin \mathcal{E}_p} \min_{e \in \text{Path}(e'; \mathcal{E}_p)} J_{e,e'}, \quad (12)$$

where  $\text{Path}(e'; \mathcal{E}_p) \subset \mathcal{E}_p$  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 the 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 (12), we need to compute at most  $\zeta(T_p)(d-1)(d-2)/2$  crossover rates, where  $\zeta(T_p)$  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 [7], [17] (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  $\Sigma$  has full rank and  $T_p$  is not a forest (as was assumed in Section II).

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 (11), can only be found numerically, since the optimization is non-convex (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. Firstly, 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. Secondly, it allows us to approximate the intractable problem in (11) 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 (17) and (10) 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 non-convex optimization problem in (11) to a convex program.

*Definition 1 ( $\epsilon$ -Very Noisy Condition):* The joint pdf  $p_{e,e'}$  on node pairs  $e$  and  $e'$  is said to satisfy the  $\epsilon$ -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  $\epsilon$  and  $\rho_e$ , there exists a  $\delta = \delta(\epsilon, \rho_e) > 0$  such that  $|I(p_e) - I(p_{e'})| < \delta$ , which means that if  $\epsilon$  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  $I(\hat{p}_e)$  and  $I(\hat{p}_{e'})$  may be incorrect. Thus, if  $\epsilon$  is small, we are in the very noisy learning regime, where learning is difficult.

To perform our analysis, we recall from Verdu [19, 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

$$D(\mathcal{N}(\mathbf{0}, \Sigma_{e,e'} + \Delta_{e,e'}) || \mathcal{N}(\mathbf{0}, \Sigma_{e,e'})) \leq \frac{\|\Sigma_{e,e'}^{-1} \Delta_{e,e'}\|_F^2}{4}, \quad (13)$$

where  $\|\mathbf{M}\|_F$  is the Frobenius norm of the matrix  $\mathbf{M}$ . Furthermore, the inequality in (13) is tight when the perturbation matrix  $\Delta_{e,e'}$  is small. More precisely, as the ratio of the singular values  $\frac{\sigma_{\max}(\Delta_{e,e'})}{\sigma_{\min}(\Sigma_{e,e'})}$  tends to zero, the inequality in (13) becomes tight. To convexify the problem, we also perform a linearization of the nonlinear constraint set in (11) 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 (11) as the affine subspace

$$\begin{aligned} L_{\Delta}(p_{e,e'}) &:= \{\Delta_{e,e'} \in \mathbb{R}^{m \times m} : I(\Sigma_e) + \langle \nabla_{\Sigma_e} I(\Sigma_e), \Delta_{e,e'} \rangle \\ &= I(\Sigma_{e'}) + \langle \nabla_{\Sigma_{e'}} I(\Sigma_{e'}), \Delta_{e,e'} \rangle\}, \end{aligned} \quad (14)$$

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 (13) over the affine subspace  $L_{\Delta}(p_{e,e'})$  defined in (14):

$$\tilde{J}_{e,e'} := \min_{\Delta_{e,e'} \in L_{\Delta}(p_{e,e'})} \frac{1}{4} \|\Sigma_{e,e'}^{-1} \Delta_{e,e'}\|_F^2. \quad (15)$$

One can view (15) as a *convexified* version of the original optimization problem in (11). It turns out that 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 useful information-theoretic quantity before stating the Euclidean approximation.

*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

$$s_{i,j}(x_i, x_j) := \log \frac{p_{i,j}(x_i, x_j)}{p_i(x_i)p_j(x_j)}. \quad (16)$$

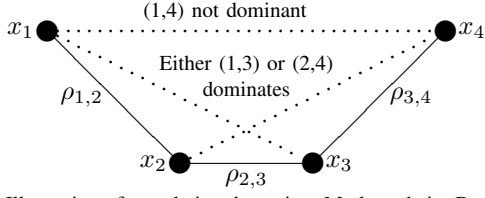


Fig. 2. Illustration of correlation decay in a Markov chain. By Lemma 5(b), only the node pairs (1, 3) and (2, 4) need to be considered for computing the error exponent  $\tilde{K}_p$ . By correlation decay, the node pair (1, 4) will not be mistaken as a true edge by the estimator because its distance, which is equal to 3, is longer than either (1, 3) or (2, 4), whose distances are equal to 2.

Hence, for each node pair  $e = (i, j)$ , the information density  $s_e$  can also be regarded as random variable whose expectation is simply the mutual information of  $x_i$  and  $x_j$ , i.e.,  $\mathbb{E}[s_e] = I(p_e)$ .

*Theorem 4 (Euclidean Approx. of Crossover Rate):*

The approximate crossover rate for the empirical mutual information quantities, defined in (15), is given by

$$\tilde{J}_{e,e'} = \frac{(\mathbb{E}[s_{e'} - s_e])^2}{2 \text{Var}(s_{e'} - s_e)} = \frac{(I(p_{e'}) - I(p_e))^2}{2 \text{Var}(s_{e'} - s_e)}, \quad (17)$$

where  $s_e$  is the information density, defined in (16), and the expectation and variance are with respect to the density  $p_{e,e'}$ . In addition, the approximate error exponent corresponding to  $\tilde{J}_{e,e'}$  in (15) is given by

$$\tilde{K}_p = \min_{e' \in \mathcal{E}_p} \min_{e \in \text{Path}(e'; \mathcal{E}_p)} \tilde{J}_{e,e'}. \quad (18)$$

We have obtained a closed-form expression for the approximate crossover rate  $\tilde{J}_{e,e'}$  in (17). 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  $\tilde{J}_{e,e'}$  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  $I(\hat{p}_e) - I(\hat{p}_{e'})$ , 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 (17) to significantly reduce the complexity in finding the error exponent  $\tilde{K}_p$  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  $\tilde{K}_p$  in (18), we have to consider all possible replacements between the non-edges (1, 4), (1, 3) and (2, 4) and the true edges along the unique 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  $\tilde{K}_p$  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  $\tilde{J}_{e,e'}$ . The key to this result is

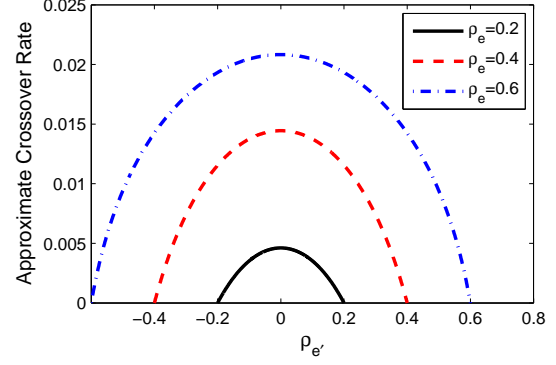


Fig. 3. Illustration of the properties of  $\tilde{J}(\rho_e, \rho_{e'})$  in Lemma 5.

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:

$$\rho_{e'} = \prod_{e \in \text{Path}(e'; \mathcal{E}_p)} \rho_e, \quad \forall e' \notin \mathcal{E}_p. \quad (19)$$

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  $\tilde{J}_{e,e'}$  in (17) depends *only* on the correlation coefficients  $\rho_e$  and  $\rho_{e'}$  and not the variances  $\sigma_i^2$ . 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  $\tilde{J}_{e,e'} = \tilde{J}(\rho_e, \rho_{e'})$ . Finally define  $\rho_{\text{crit}} := 0.63055$ .

*Lemma 5 (Monotonicity of  $\tilde{J}(\rho_e, \rho_{e'})$ ):*  $\tilde{J}(\rho_e, \rho_{e'})$ , derived in (17), has the following properties:

- (a)  $\tilde{J}(\rho_e, \rho_{e'})$  is an even function of both  $\rho_e$  and  $\rho_{e'}$ .
- (b)  $\tilde{J}(\rho_e, \rho_{e'})$  is monotonically *decreasing* in  $|\rho_{e'}|$  for fixed  $\rho_e \in (-1, 1)$ .

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

Our intuition about correlation decay is substantiated by Lemma 5(b), which implies that for the example in Fig. 2,  $\tilde{J}(\rho_{2,3}, \rho_{1,3}) \leq \tilde{J}(\rho_{2,3}, \rho_{1,4})$ , since  $|\rho_{1,4}| \leq |\rho_{1,3}|$  due to Markov property on the chain (19). Therefore,  $\tilde{J}(\rho_{2,3}, \rho_{1,4})$  can be ignored in the minimization to find  $\tilde{K}_p$  in (18).

From Lemma 5(b) (and the above motivating example in Fig. 2), finding the approximate error exponent  $\tilde{K}_p$  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  $\tilde{J}(\rho_e, \rho_{e'})$  for *adjacent edges*.

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

$$\tilde{K}_p = \min_{e_i, e_j \in \mathcal{E}_p, e_i \sim e_j} W(\rho_{e_i}, \rho_{e_j}), \quad (20)$$

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

and the weights are defined as

$$W(\rho_{e_1}, \rho_{e_2}) := \min \left\{ \tilde{J}(\rho_{e_1}, \rho_{e_1} \rho_{e_2}), \tilde{J}(\rho_{e_2}, \rho_{e_1} \rho_{e_2}) \right\}. \quad (21)$$

If we carry out the computations in (20) independently, the complexity is  $\mathcal{O}(d \deg_{\max})$ , where  $\deg_{\max}$  is the maximum degree 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 (18) is used. We can, in fact, do better and reduce the number of computations to  $\mathcal{O}(d)$  by rewriting (20) in a different form, as stated below.

*Proposition 7 (Complexity in computing  $\tilde{K}_p$ ):* The approximate error exponent  $\tilde{K}_p$ , derived in (18), can be computed in linear time ( $d - 1$  operations) as

$$\tilde{K}_p = \min_{e \in \mathcal{E}_p} \tilde{J}(\rho_e, \rho_e \rho_e^*), \quad (22)$$

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

$$\rho_e^* := \max\{|\rho_{\tilde{e}}| : \tilde{e} \in \mathcal{E}_p, \tilde{e} \sim e\}. \quad (23)$$

The computation of  $\tilde{K}_p$  is reduced significantly from  $\mathcal{O}(\zeta(T_p)d^2)$  in (12) to  $\mathcal{O}(d)$ . Thus, there is a further reduction in the complexity to estimate the error exponent  $\tilde{K}_p$  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  $\tilde{K}_p$  using the concept of correlation decay and the properties of the crossover rate  $\tilde{J}_{e,e'}$  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  $\tilde{K}_p$ ) 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  $\zeta(T_p)$ , and show that the extremal trees, in terms of their diameter, are also extremal trees for learning.

From the Markov property in (19), 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 (since all variables have unit variance) and they completely characterize  $p$ . Note that this parameterization neatly decouples the structure from the correlation coefficients. We use this fact in the subsequent sections to study the influence of changing the structure  $T_p$  while keeping the set of correlation coefficients on the edges fixed.<sup>5</sup> Before doing so, we review a basic graph-theoretic notion.

<sup>5</sup>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 (26).

*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  $\zeta : \mathcal{T}^d \rightarrow \{2, \dots, d - 1\}$  as

$$T_{\max}(d) := \operatorname{argmax}_{T \in \mathcal{T}^d} \zeta(T), \quad T_{\min}(d) := \operatorname{argmin}_{T \in \mathcal{T}^d} \zeta(T), \quad (24)$$

Then it is clear that the two extremal structures, the *chain* (where nodes are arranged in a path) and the *star* (where there is one central node) have the largest and smallest diameters respectively, i.e.,

$$T_{\max}(d) = T_{\text{chain}}(d), \quad T_{\min}(d) = T_{\text{star}}(d). \quad (25)$$

### A. Formulation: Extremal Structures for Learning

We now formulate the problem of finding the best and worst tree structures. Let  $\boldsymbol{\rho} := [\rho_1, \rho_2, \dots, \rho_{d-1}]$  be a *fixed* vector of feasible<sup>6</sup> correlation coefficients, i.e.,  $\rho_i \in (-1, 1) \setminus \{0\}$  for all  $i$ . For a tree, it follows from (19) that if  $\rho_i$ '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  $\mathbf{\Pi}_{d-1}$  to be the group of permutations of order  $d - 1$ , hence elements in  $\mathbf{\Pi}_{d-1}$  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,

$$\begin{aligned} \mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d; \boldsymbol{\rho}) := \{ & p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \boldsymbol{\Sigma}) \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d) : \\ & \boldsymbol{\Sigma}(i, i) = 1, \forall i \in \mathcal{V}, \exists \boldsymbol{\pi}_p \in \mathbf{\Pi}_{d-1} : \boldsymbol{\sigma}_{\mathcal{E}_p} = \boldsymbol{\pi}_p(\boldsymbol{\rho}) \}, \end{aligned} \quad (26)$$

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<sup>7</sup> 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  $\tilde{K}_p$  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) distributions for learning. Formally, the optimization problems for the best and worst distributions for learning are given by

$$p_{\max, \boldsymbol{\rho}} := \operatorname{argmax}_{p \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d; \boldsymbol{\rho})} \tilde{K}_p, \quad (27)$$

$$p_{\min, \boldsymbol{\rho}} := \operatorname{argmin}_{p \in \mathcal{P}_{\mathcal{N}}(\mathbb{R}^d, \mathcal{T}^d; \boldsymbol{\rho})} \tilde{K}_p. \quad (28)$$

Thus,  $p_{\max, \boldsymbol{\rho}}$  corresponds to the Gaussian tree model which has the highest approximate error exponent. Also  $T_{p_{\max, \boldsymbol{\rho}}}$  and

<sup>6</sup>We do not allow any of the correlations  $\rho_i$  to be zero because otherwise,  $T_p$  would be a proper forest.

<sup>7</sup>None of the elements in  $\boldsymbol{\Sigma}$  are allowed to be zero because of the Markov property in (19) and the fact that  $\rho_i \neq 0$  for every  $i \in \mathcal{V}$ .

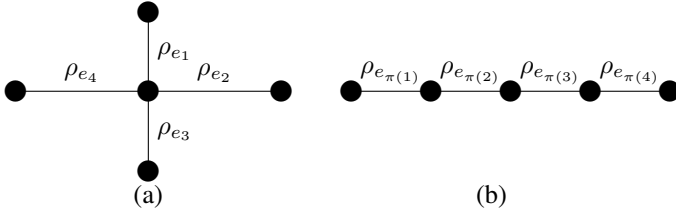


Fig. 4. Illustration for Theorem 8: The star (a) and the chain (b) minimize and maximize the approximate error exponent respectively. For the chain, the order (or permutation) of the correlation coefficients  $\pi(i)$  along the path is determined by construction.

$T_{p_{\min}, \rho}$  are the tree structures associated to  $p_{\max, \rho}$  and  $p_{\min, \rho}$  respectively.

### B. The Main Result: Best and Worst Tree Structures

We now state our main result, which is to identify the structures corresponding to the extremal distributions,  $p_{\max, \rho}$  and  $p_{\min, \rho}$  by exploiting the monotonicity of  $\tilde{J}(\rho_e, \rho_{e'})$  given in Lemma 5(b).

*Theorem 8 (Universal Extremal Tree Structures):* The tree structures of the extremal distributions that maximize and minimize the approximate error exponent  $\tilde{K}_p$  in (27) and (28) are given by

$$T_{p_{\min}, \rho} = T_{\text{star}}(d), \quad T_{p_{\max}, \rho} = T_{\text{chain}}(d), \quad (29)$$

for all feasible correlation coefficient vectors  $\rho$  with  $\rho_i \in (-1, 1) \setminus \{0\}$ . See Fig. 4.

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 is *universal* in the sense that the extremal tree structures  $T_{p_{\max}, \rho}$  and  $T_{p_{\min}, \rho}$  are *independent of* the correlation coefficients  $\rho$  and the permutation  $\pi_p$ .

## VII. NUMERICAL EXPERIMENTS

We now perform experiments with the aim of studying how various tree structures (e.g. chains and stars) influence the error exponents (Theorem 8).

In Fig. 5, we simulate the error probabilities by drawing samples from three different  $d = 10$  node tree graphs – a chain, a star and a hybrid between a chain and a star as shown in Fig. 6. We then used the samples to learn the structure by solving (5). The  $d-1 = 9$  correlation coefficients were 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. 5 that for fixed  $n$ , the star and chain have the highest and lowest error probabilities respectively. The *simulated error exponents* given by  $\{-\frac{1}{n} \log \mathbb{P}(\mathcal{A}_n)\}_{n \in \mathbb{N}}$  converge to their true values as  $n \rightarrow \infty$ . The exponent associated to the star is higher than that of the chain, which is corroborated by Theorem 8. We also observe that the exponent of the hybrid is between that of the star and the chain.

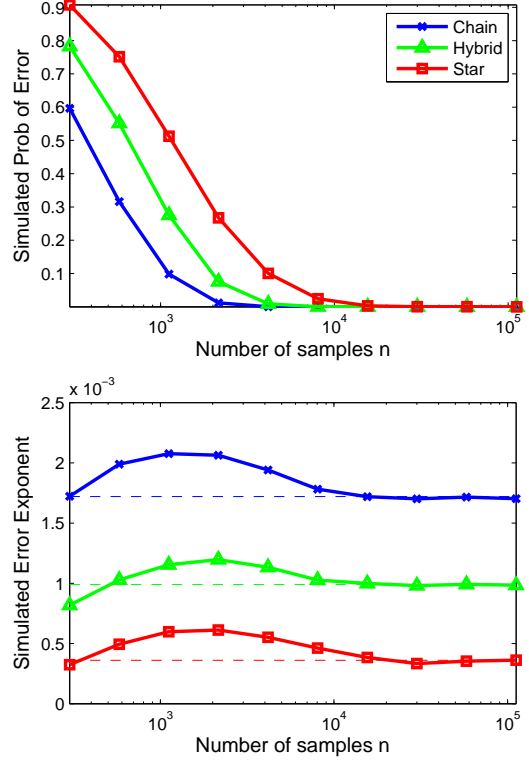


Fig. 5. Simulated error probabilities and error exponents for chain, hybrid and star graphs with fixed  $\rho$ . The dashed lines show the true error exponent  $K_p$  computed numerically using (11) and (12). Observe that the simulated error exponent converges to the true error exponent as  $n \rightarrow \infty$ . The legend applies to both plots.

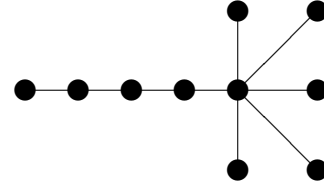


Fig. 6. The structure of a *hybrid* tree graph with  $d = 10$  nodes. This is a tree with a length- $d/2$  chain and a order  $d/2$  star attached to one of the leaf nodes of the chain.

## 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 (resp. stars) are the easiest (resp. hardest) structures to learn as they maximize (resp. 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.

In future work, we would like to find the Gaussian distributions that optimize both (27) and (28), *i.e.*, the distributions that maximize and minimize the error exponents. The difficulty in doing so stems primarily from the fact that the number of permutations of the correlation coefficients on the edges is prohibitively large. To ameliorate this problem, we seek to identify subclasses of Gaussian tree models that admit tractable algorithms for finding the extremal distributions. Finally, we would also like to analyze how adding or deleting nodes and edges from the original tree model affects the error exponent.

## 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)*, 1994.
- [3] S. Lauritzen, *Graphical Models*. Oxford University Press, USA, 1996.
- [4] M. J. Wainwright and M. I. Jordan, *Graphical Models, Exponential Families, and Variational Inference*, ser. Foundations and Trends in Machine Learning. Now Publishers Inc, 2008, vol. 1.
- [5] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. on 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. on 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. of IEEE Intl. Symp. on Inf. Theory*, Seoul, July 2009, <http://arxiv.org/abs/0905.0940>.
- [8] D. M. Chickering, "Learning equivalence classes of Bayesian network structures," *Journal of Machine Learning Research*, 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 *Conference on Learning Theory (COLT)*, 2004.
- [10] N. Meinshausen and P. Buehlmann, "High dimensional graphs and variable selection with the Lasso," *Annals of Statistics*, 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)*. MIT Press, 2006.
- [12] N. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," in *Proc. of IEEE Intl. Symp. on Inf. Theory*, Toronto, Canada, July 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 *Uncertainty in Artificial Intelligence (UAI)*, 2006.
- [14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, 2006.
- [15] J.-D. Deuschel and D. W. Stroock, *Large Deviations*. American Mathematical Society, Dec 2000.
- [16] F. D. Hollander, *Large Deviations (Fields Institute Monographs)*. American Mathematical Society, Feb 2000.
- [17] D. B. West, *Introduction to Graph Theory*, 2nd ed. Prentice Hall, 2000.
- [18] S. Borade and L. Zheng, "Euclidean Information Theory," in *Allerton Conference*, 2007.
- [19] S. Verdu, "Spectral efficiency in the wideband regime," *IEEE Trans. on Inf. Theory*, vol. 48, no. 6, Jun 2002.