A Large-Deviation Analysis of the Maximum-Likelihood Learning of Markov Tree Structures

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

***Abstract*—The problem of maximum-likelihood (ML) estimation of discrete tree-structured distributions is considered. Chow and Liu established that ML-estimation reduces to the construction of a maximum-weight spanning tree using the empirical mutual information quantities as the edge weights. Using the theory of large-deviations, we analyze the exponent associated with the error probability of the event that the ML-estimate of the Markov tree structure differs from the true tree structure, given a set of independently drawn samples. By exploiting the fact that the output of ML-estimation is a tree, we establish that the error exponent is equal to the exponential rate of decay of a single dominant *crossover* event. We prove that in this dominant crossover event, a non-neighbor node pair replaces a true edge of the distribution that is along the path of edges in the true tree graph connecting the nodes in the non-neighbor pair. Using ideas from Euclidean information theory, we then analyze the scenario of ML-estimation in the *very noisy* learning regime and show that the error exponent can be approximated as a ratio, which is interpreted as the *signal-to-noise* ratio (SNR) for learning tree distributions. We show via numerical experiments that in this regime, our SNR approximation is accurate.**

***Index Terms*—Error exponent, Euclidean information theory, large-deviations principle, Markov structure, maximum-likelihood (ML) distribution estimation, tree-structured distributions.**

I. Introduction

THE estimation of a distribution from samples is a classical and an important generic problem in machine learning and statistics and is challenging for high-dimensional multivariate distributions. In this respect, graphical models [2] provide a significant simplification of joint distribution as the distribution can be factorized according to a graph defined on the set of nodes. Many specialized algorithms [3]–[9] exist for exact and approximate learning of graphical models Markov on sparse graphs.

There are many applications of learning graphical models, including clustering and dimensionality reduction. Suppose we have d genetic variables and we would like to group the ones that are similar together. Then the construction of a graphical model provides a visualization of the relationship between genes. Those genes that have high degree are highly correlated to many other genes (e.g., those in its neighborhood). The learning of a graphical model may also provide the means to judiciously remove redundant genes from the model, thus reducing the dimensionality of the data, leading to more efficient inference of the effects of the genes subsequently.

When the underlying graph is a tree, the Chow-Liu algorithm [3] provides an efficient method for the maximum-likelihood (ML) estimation of the probability distribution from a set of i.i.d. samples drawn from the distribution. By exploiting the Markov tree structure, this algorithm reduces the ML-estimation problem to solving a maximum-weight spanning tree (MWST) problem. In this case, it is known that the ML-estimator learns the distribution correctly asymptotically, and hence, is consistent [10].

While consistency is an important qualitative property for any estimator, the study of the rate of convergence, a precise quantitative property, is also of great practical interest. We are interested in the rate of convergence of the ML-estimator (Chow-Liu algorithm) for tree distributions as we increase the number of samples. Specifically, we study the rate of decay of the error probability or the error exponent of the ML-estimator in learning the *tree structure* of the unknown distribution. A larger exponent means that the error probability in structure learning decays more rapidly. In other words, we need relatively few samples to ensure that the error probability is below some fixed level $\delta > 0$. Such model are thus "easier" to learn. We address the following questions: Is there exponential decay of the probability of error in structure learning as the number of samples tends to infinity? If so, what is the exact error exponent, and how does it depend on the parameters of the distribution? Which edges of the true tree are most-likely to be in error; in other words, what is the nature of the most-likely error in the ML-estimator? We provide concrete and intuitive answers to the above questions, thereby providing insights into how the parameters of the distribution influence the error

Manuscript received May 06, 2009; revised October 19, 2010; accepted November 18, 2010. Date of current version February 18, 2011. This work was supported in part by A*STAR, Singapore, by a MURI funded through ARO Grant W911NF-06-1-0076 and by AFOSR Grant FA9550-08-1-0180 and in part by the Army Research Office MURI Program under award W911NF-08-1-0238. The material in this paper was presented in part at the International Symposium on Information Theory (ISIT), Seoul, Korea, June 2009. V. Y. F. Tan performed this work while at MIT.

V. Y. F. Tan is with the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, WI 53706 USA (e-mail: vtan@wisc.edu).

A. Anandkumar is with the Center for Pervasive Communications and Computing, Electrical Engineering and Computer Science Department, University of California, Irvine 92697 USA (e-mail: a.anandkumar@uci.edu).

L. Tong is with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853 USA (e-mail: ltong@ece.cornell.edu).

A. S. Willsky is with the Stochastic Systems Group, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: willsky@mit.edu).

Communicated by A. Krzyzak, Associate Editor for Pattern Recognition, Statistical Learning, and Inference.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2011.2104513

exponent associated with learning the structure of discrete tree distributions.

A. Main Contributions

There are three main contributions in this paper. First, using the large-deviation principle (LDP) [11] we prove that the most-likely error in ML-estimation is a tree which differs from the true tree by a single edge. Second, again using the LDP, we derive the exact error exponent for ML-estimation of tree structures. Third, we provide a succinct and intuitive closed-form approximation for the error exponent which is tight in the *very noisy* learning regime, where the individual samples are not too informative about the tree structure. The approximate error exponent has a very intuitive explanation as the *signal-to-noise ratio* (SNR) for learning.

We analyze the *error exponent* (also called the inaccuracy rate) for the estimation of the structure of the unknown tree distribution. For the error event that the structure of the ML-estimator $\mathcal{E}_{\mathrm{ML}}$ given n samples differs from the true tree structure $\mathcal{E}_P$ of the unknown distribution P, the error exponent is given by

$$K_P := \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}(\{\mathcal{E}_{\mathrm{ML}} \neq \mathcal{E}_P\}). \tag{1}$$

To the best of our knowledge, error-exponent analysis for tree-structure learning has not been considered before (See Section I-B for a brief survey of the existing literature on learning graphical models from data).

Finding the error exponent K_P in (1) is not straightforward since in general, one has to find the *dominant* error event with the *slowest* rate of decay among all possible error events [11, Ch. 1]. For learning the structure of trees, there are a total of $d^{d-2}-1$ possible error events,[1] where d is the dimension (number of variables or nodes) of the unknown tree distribution P. Thus, in principle, one has to consider the information projection [13] of P on all these error trees. This rules out brute-force information projection approaches for finding the error exponent in (1), especially for high-dimensional data.

In contrast, we establish that the search for the dominant error event for learning the structure of the tree can be limited to a polynomial-time search space (in d). Furthermore, we establish that this dominant error event of the ML-estimator is given by a tree which differs from the true tree by only a single edge. We provide a polynomial algorithm with $\mathcal{O}(\mathrm{diam}(T_P)d^2)$ complexity to find the error exponent in (1), where $\mathrm{diam}(T_P)$ is the diameter of the tree T_P. We heavily exploit the mechanism of the ML Chow-Liu algorithm [3] for tree learning to establish these results, and specifically, the fact that the ML-estimator tree distribution depends *only* on the relative order of the empirical mutual information quantities between all the node pairs (and not their absolute values).

Although we provide a computationally-efficient way to compute the error exponent in (1), it is not available in closed-form. In Section VI, we use Euclidean information theory [14], [15] to obtain an approximate error exponent in closed-form, which can be interpreted as the signal-to-noise ratio (SNR) for tree structure learning. Numerical simulations on various discrete graphical models verify that the approximation is tight in the very noisy regime.

In Section VII, we extend our results to the case when the true distribution P is not a tree. In this case, given samples drawn independently from P, we intend to learn the *optimal reverse information projection* P^* onto the set of trees. Importantly, if P is not a tree, there may be several trees that are optimal projections [10] and this requires careful consideration of the error events. We derive the error exponent even in this scenario.

B. Related Work

The seminal work by Chow and Liu in [3] focused on learning tree models from data samples. The authors showed that the learning of the optimal tree distribution essentially decouples into two distinct steps: (i) a structure learning step and (ii) a parameter learning step. The structure learning step, which is the focus on this paper, can be performed efficiently using a max-weight spanning tree algorithm with the empirical mutual information quantities as the edge weights. The parameter learning step is a maximum-likelihood estimation procedure where the parameters of the learned model are equal to those of the empirical distribution. Chow and Wagner [10], in a follow-up paper, studied the consistency properties of the Chow-Liu algorithm for learning trees. They concluded that if the true distribution is Markov on a unique tree structure, then the Chow-Liu learning algorithm is asymptotically consistent. This implies that as the number of samples tends to infinity, the probability that the learned structure differs from the (unique) true structure tends to zero.

Unfortunately, it is known that the exact learning of general graphical models is NP-hard [16], but there have been several works to learn approximate models. For example, Chechetka and Guestrin [4] developed good approximations for learning thin junction trees [17] (junction trees where the sizes of the maximal cliques are small). Heckerman [18] proposed learning the structure of Bayesian networks by using the Bayesian Information Criterion [19] (BIC) to penalize more complex models and by putting priors on various structures. Other authors used the maximum entropy principle or (sparsity-enforcing) ℓ_1 regularization as approximate graphical model learning techniques. In particular, Dudik *et al.* [9] and Lee *et al.* [6] provide strong consistency guarantees on the learned distribution in terms of the log-likelihood of the samples. Johnson *et al.* [7] also used a similar technique known as maximum entropy relaxation (MER) to learn discrete and Gaussian graphical models. Wainwright *et al.* [5] proposed a regularization method for learning the graph structure based on ℓ_1 logistic regression and provided strong theoretical guarantees for learning the correct structure as the number of samples, the number of variables, and the neighborhood size grow. In a similar work, Meinshausen and Buehlmann [8] considered learning the structure of arbitrary Gaussian models using the Lasso [20]. They show that the error probability of learning the wrong structure, under some mild technical conditions on the neighborhood size, decays exponentially even when the size of the graph d grows with the number of samples n. However, the rate of decay is not

[1] Since the ML output $\mathcal{E}_{\mathrm{ML}}$ and the true structure $\mathcal{E}_P$ are both spanning trees over d nodes and since there are d^{d-2} possible spanning trees [12], we have $d^{d-2}-1$ number of possible error events.

provided explicitly. Zuk *et al.* [21] provided bounds on the limit inferior and limit superior of the error rate for learning the structure of Bayesian networks but, in contrast to our work, these bounds are not asymptotically tight. In addition, the work in Zuk *et al.* [21] is intimately tied to the BIC [19], whereas our analysis is for the Chow-Liu ML tree learning algorithm [3]. A modification of the Chow-Liu learning algorithm has also been applied to learning the structure of latent trees where only a subset of variables are observed [22].

There have also been a series of papers [23]–[26] that quantify the deviation of the empirical information-theoretic quantities from their true values by employing techniques from large-deviations theory. Some ideas from these papers will turn out to be important in the subsequent development because we exploit conditions under which the empirical mutual information quantities do not differ "too much" from their nominal values. This will ensure that structure learning succeeds with high probability.

C. Paper Outline

This paper is organized as follows: In Sections II and III, we state the system model and the problem statement and provide the necessary preliminaries on undirected graphical models and the Chow-Liu algorithm [3] for learning tree distributions. In Section IV, we derive an analytical expression for the crossover rate of two node pairs. We then relate the crossover rates to the overall error exponent in Section V. We also discuss some connections of the problem we solve here with robust hypothesis testing. In Section VI, we leverage on ideas in Euclidean information theory to state sufficient conditions that allow approximations of the crossover rate and the error exponent. We obtain an intuitively appealing closed-form expression. By redefining the error event, we extend our results to the case when the true distribution is not a tree in Section VII. We compare the true and approximate crossover rates by performing numerical experiments for a given graphical model in Section VIII. Perspectives and possible extensions are discussed in Section IX.

II. System Model and Problem Statement

A. Graphical Models

An undirected graphical model [2] is a probability distribution that factorizes according to the structure of an underlying undirected graph. More explicitly, a vector of random variables $\mathbf{x} := [x_1, \ldots, x_d]^T$ is said to be *Markov* on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V} = \{1, \ldots, d\}$ and edge set $\mathcal{E} \subset \binom{\mathcal{V}}{2}$ if

$$P(x_i | x_{\mathcal{V}\setminus\{i\}}) = P(x_i | x_{\mathrm{nbd}(i)}), \quad \forall\, i \in \mathcal{V} \tag{2}$$

where $\mathrm{nbd}(i)$ is the set of neighbors of i in $\mathcal{G}$, i.e., $\mathrm{nbd}(i) := \{j \in \mathcal{V} : (i,j) \in \mathcal{E}\}$. Equation (2) is called the (local) Markov property and states that if random variable x_i is conditioned on its neighboring random variables, then x_i is independent of the rest of the variables in the graph.

In this paper, we assume that each random variable $x_i \in \mathcal{X}$, and we also assume that $\mathcal{X} = \{1, \ldots, |\mathcal{X}|\}$ is a *known finite* set.[2] Hence, the joint distribution $P \in \mathcal{P}(\mathcal{X}^d)$, where $\mathcal{P}(\mathcal{X}^d)$ is the probability simplex of all distributions supported on $\mathcal{X}^d$.

Except for Section VII, we limit our analysis in this paper to the set of strictly positive[3] graphical models P, in which the graph of P is a tree on the d nodes, denoted $T_P = (\mathcal{V}, \mathcal{E}_P)$. Thus, T_P is an undirected, acyclic and connected graph with vertex set $\mathcal{V} = \{1, \ldots, d\}$ and edge set $\mathcal{E}_P$, with $d-1$ edges. Let $\mathcal{T}^d$ be the set of *spanning trees* on d nodes, and hence, $T_P \in \mathcal{T}^d$. Tree distributions possess the following factorization property [2]

$$P(\mathbf{x}) = \prod_{i \in \mathcal{V}} P_i(x_i) \prod_{(i,j) \in \mathcal{E}_P} \frac{P_{i,j}(x_i, x_j)}{P_i(x_i) P_j(x_j)} \tag{3}$$

where P_i and $P_{i,j}$ are the marginals on node $i \in \mathcal{V}$ and edge $(i,j) \in \mathcal{E}_P$ respectively. Since T_P is spanning, $P_{i,j} \neq P_i P_j$ for all $(i,j) \in \mathcal{E}_P$. Hence, there is a substantial simplification of the joint distribution which arises from the Markov tree dependence. In particular, the distribution is completely specified by the set of edges $\mathcal{E}_P$ and pairwise marginals $P_{i,j}$ on the edges of the tree $(i,j) \in \mathcal{E}_P$. In Section VII, we extend our analysis to general distributions which are not necessarily Markov on a tree.

B. Problem Statement

In this paper, we consider a learning problem, where we are given a set of n i.i.d. d-dimensional samples $\mathbf{x}^n := \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ from an unknown distribution $P \in \mathcal{P}(\mathcal{X}^d)$, which is Markov with respect to a tree $T_P \in \mathcal{T}^d$. Each sample or observation $\mathbf{x}_k := [x_{k,1}, \ldots, x_{k,d}]^T$ is a vector of d dimensions where each entry can only take on one of a finite number of values in the alphabet $\mathcal{X}$.

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

$$P_{\mathrm{ML}} := \operatorname*{arg\,max}_{Q \in \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)} \sum_{k=1}^{n} \log Q(\mathbf{x}_k) \tag{4}$$

where $\mathcal{D}(\mathcal{X}^d, \mathcal{T}^d) \subset \mathcal{P}(\mathcal{X}^d)$ is defined as the set of all tree distributions on the alphabet $\mathcal{X}^d$ over d nodes.

In 1968, Chow and Liu showed that the above ML-estimate P_{ML} can be found efficiently via a MWST algorithm [3], and is described in Section III. We denote the tree graph of the ML-estimate P_{ML} by $T_{\mathrm{ML}} = (\mathcal{V}, \mathcal{E}_{\mathrm{ML}})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}_{\mathrm{ML}}$.

Given a tree distribution P, define the probability of the error event that the set of edges is *not* estimated correctly by the ML-estimator as

$$\mathcal{A}_n := \{\mathcal{E}_{\mathrm{ML}} \neq \mathcal{E}_P\} \tag{5}$$

We denote $\mathbb{P} := P^n$ as the n-fold product probability measure of the n samples $\mathbf{x}^n$ which are drawn i.i.d. from P. In this paper,

[2]The analysis of learning the structure of jointly Gaussian variables where $\mathcal{X} = \mathbb{R}$ is deferred to a companion paper [27]. The error exponent analysis carries over to the case where $\mathcal{X}$ is a countably infinite set.

[3]A distribution P is said to be strictly positive if $P(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}^d$.

we are interested in studying the *rate* or *error exponent* [4] K_P at which the above error probability exponentially decays with the number of samples n, given by

$$K_P := \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}(\mathcal{A}_n) \tag{6}$$

whenever the limit exists. Indeed, we will prove that the limit in (6) exists in the sequel. With the $\doteq$ notation,[5] (6) can be written as

$$\mathbb{P}(\mathcal{A}_n) \doteq \exp(-nK_P). \tag{7}$$

A positive error exponent $(K_P > 0)$ implies an exponential decay of error probability in ML structure learning, and we will establish necessary and sufficient conditions to ensure this.

Note that we are only interested in quantifying the probability of the error in learning the *structure* of P in (5). We are not concerned about the parameters that define the ML tree distribution P_{ML}. Since there are only finitely many (but a super-exponential number of) structures, this is in fact akin to an ML problem where the parameter space is discrete and finite [31]. Thus, under some mild technical conditions, we can expect exponential decay in the probability of error as mentioned in [31]. Otherwise, we can only expect convergence with rate $\mathcal{O}_p(1/\sqrt{n})$ for estimation of parameters that belong to a continuous parameter space [32]. In this work, we quantify the error exponent for learning tree structures using the ML learning procedure precisely.

III. Maximum-Likelihood Learning of Tree Distributions From Samples

In this section, we review the classical Chow-Liu algorithm [3] for learning the ML tree distribution P_{ML} given a set of n samples $\mathbf{x}^n$ drawn i.i.d. from a tree distribution P. Recall the ML-estimation problem in (4), where $\mathcal{E}_{\mathrm{ML}}$ denotes the set of edges of the tree T_{ML} on which P_{ML} is tree-dependent. Note that since P_{ML} is tree-dependent, from (3), we have the result that it is completely specified by the structure $\mathcal{E}_{\mathrm{ML}}$ and consistent pairwise marginals $P_{\mathrm{ML}}(x_i, x_j)$ on its edges $(i,j) \in \mathcal{E}_{\mathrm{ML}}$.

In order to obtain the ML-estimator, we need the notion of a *type* or *empirical distribution* of P, given $\mathbf{x}^n$, defined as

$$\widehat{P}(\mathbf{x}; \mathbf{x}^n) := \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\{\mathbf{x}_k = \mathbf{x}\} \tag{8}$$

where $\mathbb{I}\{\mathbf{x}_k = \mathbf{x}\} = 1$ if $\mathbf{x}_k = \mathbf{x}$ and equals 0 otherwise. For convenience, in the rest of the paper, we will denote the empirical distribution by $\widehat{P}(\mathbf{x})$ instead of $\widehat{P}(\mathbf{x}; \mathbf{x}^n)$.

Fact 1: The ML-estimator in (4) is equivalent to the following optimization problem:

$$P_{\mathrm{ML}} = \arg\min_{Q\in\mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)} D(\widehat{P}\|Q) \tag{9}$$

where $\widehat{P}$ is the empirical distribution of $\mathbf{x}^n$, given by (8). In (9), $D(\widehat{P}\|Q) = \sum_{\mathbf{x}\in\mathcal{X}^d} \widehat{P}(\mathbf{x}) \log \frac{\widehat{P}(\mathbf{x})}{Q(\mathbf{x})}$ denotes the Kullback-Leibler divergence (or relative entropy) [30, Ch. 1] between the probability distributions $\widehat{P}, Q \in \mathcal{P}(\mathcal{X}^d)$.

Proof: By the definition of the KL-divergence, we have

$$nD(\widehat{P}\|Q) = -nH(\widehat{P}) - n\sum_{\mathbf{x}\in\mathcal{X}^d} \widehat{P}(\mathbf{x}) \log Q(\mathbf{x}) \tag{10}$$

$$= -nH(\widehat{P}) - \sum_{k=1}^{n} \log Q(\mathbf{x}_k) \tag{11}$$

where we use the fact that the empirical distribution $\widehat{P}$ in (8) assigns a probability mass of $1/n$ to each sample $\mathbf{x}_k$. ■

The minimization over the second variable in (9) is also known as the *reverse I-projection* [13], [33] of $\widehat{P}$ onto the set of tree distributions $\mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)$. We now state the main result of the Chow-Liu tree learning algorithm [3]. In this paper, with a slight abuse of notation, we denote the mutual information $I(x_i; x_j)$ between two random variables x_i and x_j corresponding to nodes i and j as

$$I(P_{i,j}) := \sum_{(x_i,x_j)\in\mathcal{X}^2} P_{i,j}(x_i, x_j) \log \frac{P_{i,j}(x_i, x_j)}{P_i(x_i)P_j(x_j)}. \tag{12}$$

Note that the definition above uses only the marginal of P restricted to (x_i, x_j). If $e = (i,j)$, then we will also denote the mutual information as $I(P_e) = I(P_{i,j})$.

Theorem 1 (Chow-Liu Tree Learning [3]): The structure and parameters of the ML-estimate P_{ML} in (4) are given by

$$\mathcal{E}_{\mathrm{ML}} = \arg\max_{\mathcal{E}_Q : Q\in\mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)} \sum_{e\in\mathcal{E}_Q} I(\widehat{P}_e) \tag{13}$$

$$P_{\mathrm{ML}}(x_i, x_j) = \widehat{P}_{i,j}(x_i, x_j), \qquad \forall (i,j) \in \mathcal{E}_{\mathrm{ML}} \tag{14}$$

where $\widehat{P}$ is the empirical distribution in (8) given the data $\mathbf{x}^n$, and $I(\widehat{P}_e) = I(\widehat{P}_{i,j})$ is the empirical mutual information of random variables x_i and x_j, which is a function of the empirical distribution $\widehat{P}_e$.

[4] In the maximum-likelihood estimation literature (e.g., [28], [29]) if the limit in (6) exists, K_P is also typically known as the inaccuracy rate. We will be using the terms rate, error exponent and inaccuracy rate interchangeably in the sequel. All these terms refer to K_P.

[5] The $\doteq$ notation (used in [30]) denotes equality to the first order in the exponent. For two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n \doteq b_n$ if and only if $\lim_{n\to\infty} \frac{1}{n}\log(a_n/b_n) = 0$.

Proof: For a fixed tree distribution $Q \in \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)$, Q admits the factorization in (3), and we have

$$\begin{aligned} &D(\widehat{P}\|Q) + H(\widehat{P}) \\ &\quad = -\sum_{\mathbf{x}\in\mathcal{X}^d} \widehat{P}(\mathbf{x}) \log \left[\prod_{i\in\mathcal{V}} Q_i(x_i) \prod_{(i,j)\in\mathcal{E}_Q} \frac{Q_{i,j}(x_i,x_j)}{Q_i(x_i)Q_j(x_j)} \right] \quad (15) \\ &\quad = -\sum_{i\in\mathcal{V}} \sum_{x_i\in\mathcal{X}} \widehat{P}_i(x_i) \log Q_i(x_i) \\ &\qquad - \sum_{(i,j)\in\mathcal{E}_Q} \sum_{(x_i,x_j)\in\mathcal{X}^2} \widehat{P}_{i,j}(x_i,x_j) \log \frac{Q_{i,j}(x_i,x_j)}{Q_i(x_i)Q_j(x_j)}. \quad (16) \end{aligned}$$

For a fixed structure $\mathcal{E}_Q$, it can be shown [3] that the above quantity is minimized when the pairwise marginals over the edges of $\mathcal{E}_Q$ are set to that of $\widehat{P}$, i.e., for all $Q \in \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)$

$$\begin{aligned} &D(\widehat{P}\|Q) + H(\widehat{P}) \\ &\quad \geq -\sum_{i\in\mathcal{V}} \sum_{x_i\in\mathcal{X}} \widehat{P}_i(x_i) \log \widehat{P}_i(x_i) \\ &\qquad - \sum_{(i,j)\in\mathcal{E}_Q} \sum_{(x_i,x_j)\in\mathcal{X}^2} \widehat{P}_{i,j}(x_i,x_j) \log \frac{\widehat{P}_{i,j}(x_i,x_j)}{\widehat{P}_i(x_i)\widehat{P}_j(x_j)}. \quad (17) \\ &\quad = \sum_{i\in\mathcal{V}} H(\widehat{P}_i) - \sum_{(i,j)\in\mathcal{E}_Q} I(\widehat{P}_e). \quad (18) \end{aligned}$$

The first term in (18) is a constant with respect to Q. Furthermore, since $\mathcal{E}_Q$ is the edge set of the tree distribution $Q \in \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)$, the optimization for the ML tree distribution P_{ML} reduces to the MWST search for the optimal edge set as in (13). ■

Hence, the optimal tree probability distribution P_{ML} is the reverse I-projection of $\widehat{P}$ onto the optimal tree structure given by (13). Thus, the optimization problem in (9) essentially reduces to a search for the *structure* of P_{ML}. The structure of P_{ML} completely determines its distribution, since the parameters are given by the empirical distribution in (14). To solve (13), we use the samples $\mathbf{x}^n$ to compute the empirical distribution $\widehat{P}$ using (8), then use $\widehat{P}$ to compute $I(\widehat{P}_e)$, for each node pair $e \in \binom{\mathcal{V}}{2}$. Subsequently, we use the set of empirical mutual information quantities $\left\{ I(\widehat{P}_e) : e \in \binom{\mathcal{V}}{2} \right\}$ as the edge weights for the MWST problem.[6]

We see that the Chow-Liu MWST spanning tree algorithm is an efficient way of solving the ML-estimation problem, especially when the dimension d is large. This is because there are d^{d-2} possible spanning trees over d nodes [12] ruling out the possibility for performing an exhaustive search for the optimal tree structure. In contrast, the MWST can be found, say using Kruskal's algorithm [34], [35] or Prim's algorithm [36], in $\mathcal{O}(d^2 \log d)$ time.

[6]If we use the true mutual information quantities as inputs to the MWST, then the true edge set $\mathcal{E}_P$ is the output.

IV. LDP for Empirical Mutual Information

The goal of this paper is to characterize the error exponent for ML tree learning K_P in (6). As a first step, we consider a simpler event, which may potentially lead to an error in ML-estimation. In this section, we derive the LDP rate for this event, and in Section V, we use the result to derive K_P, the exponent associated to the error event $\mathcal{A}_n$ defined in (5).

Since the ML-estimate uses the empirical mutual information quantities as the edge weights for the MWST algorithm, the relative values of the empirical mutual information quantities have an impact on the accuracy of ML-estimation. In other words, if the order of these empirical quantities is different from the true order then it can potentially lead to an error in the estimated edge set. Hence, it is crucial to study the probability of the event that the empirical mutual information quantities of any two node pairs is different from the true order.

Formally, let us consider two distinct node pairs with no common nodes, $e, e' \in \binom{\mathcal{V}}{2}$ with unknown distribution $P_{e,e'} \in \mathcal{P}(\mathcal{X}^4)$, where the notation $P_{e,e'}$ denotes the marginal of the tree-structured graphical model P on the nodes in the set $\{e, e'\}$. Similarly, P_e is the marginal of P on edge e. Assume that the order of the true mutual information quantities follow $I(P_e) > I(P_{e'})$. A *crossover event*[7] occurs if the corresponding empirical mutual information quantities are of the reverse order, given by

$$\mathcal{C}_{e,e'} := \left\{ I(\widehat{P}_e) \leq I(\widehat{P}_{e'}) \right\}. \quad (19)$$

As the number of samples $n \to \infty$, the empirical quantities approach the true ones in probability, and hence, the probability of the above event decays to zero. When the decay is exponential, we have a LDP for the above event, and we term its rate as the crossover rate for empirical mutual information quantities, defined as

$$J_{e,e'} := \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(\mathcal{C}_{e,e'}\right) \quad (20)$$

assuming the limit in (20) exists. Indeed, we show in the proof of Theorem 2 that the limit exists. Intuitively (and as seen in our numerical simulations in Section VIII), if the difference between the true mutual information quantities $I(P_e) - I(P_{e'})$ is large (i.e., $I(P_e) \gg I(P_{e'})$), we expect the probability of the crossover event $\mathcal{C}_{e,e'}$ to be small. Thus, the rate of decay would be faster, and hence, we expect the crossover rate $J_{e,e'}$ to be large. In the following, we see that $J_{e,e'}$ depends not only on the difference of mutual information quantities $I(P_e) - I(P_{e'})$, but also on the *distribution* $P_{e,e'}$ of the variables on node pairs e and e', since the distribution $P_{e,e'}$ influences the accuracy of estimating them.

Theorem 2 (Crossover Rate for Empirical MIs): The crossover rate for a pair of empirical mutual information quantities in (20) is given by

$$J_{e,e'} = \inf_{Q\in\mathcal{P}(\mathcal{X}^4)} \{ D(Q\|P_{e,e'}) : I(Q_{e'}) = I(Q_e) \} \quad (21)$$

[7]The event $\mathcal{C}_{e,e'}$ in (19) depends on the number of samples n but we suppress this dependence for convenience.

where $Q_e, Q_{e'} \in \mathcal{P}(\mathcal{X}^2)$ are marginals of Q over node pairs e and e', which do not share common nodes, i.e.,

$$Q_e(x_e) := \sum_{x_{e'} \in \mathcal{X}^2} Q(x_e, x_{e'}) \tag{22a}$$

$$Q_{e'}(x_{e'}) := \sum_{x_e \in \mathcal{X}^2} Q(x_e, x_{e'}). \tag{22b}$$

The infimum in (21) is attained by some distribution $Q^*_{e,e'} \in \mathcal{P}(\mathcal{X}^4)$ satisfying $I\left(Q^*_{e'}\right) = I\left(Q^*_e\right)$ and $J_{e,e'} > 0$.

Proof: (*Sketch*) The proof hinges on Sanov's theorem [30, Ch. 11] and the contraction principle in large-deviations [11, Sec. III.5]. The existence of the minimizer follows from the compactness of the constraint set and Weierstrass' extreme value theorem [37, Theorem 4.16]. The rate $J_{e,e'}$ is strictly positive since we assumed, *a-priori*, that the two node pairs e and e' satisfy $I(P_e) > I(P_{e'})$. As a result, $Q^*_{e,e'} \neq P_{e,e'}$ and $D\left(Q^*_{e,e'} \| P_{e,e'}\right) > 0$. See Appendix A for the details. ∎

In the above theorem, which is analogous to Theorem 3.3 in [25], we derived the crossover rate $J_{e,e'}$ as a constrained minimization over a submanifold of distributions in $\mathcal{P}(\mathcal{X}^4)$ (See Fig. 5), and also proved the existence of an optimizing distribution $Q^*_{e,e'}$. However, it is not easy to further simplify the rate expression in (21) since the optimization is nonconvex.

Importantly, this means that it is not clear how the parameters of the distribution $P_{e,e'}$ affect the rate $J_{e,e'}$; hence, (21) is not intuitive to aid in understanding the relative ease or difficulty in estimating particular tree-structured distributions. In Section VI, we assume that P satisfies some (so-called very noisy learning) conditions and use Euclidean information theory [14], [15] to approximate the rate in (21) in order to gain insights as to how the distribution parameters affect the crossover rate $J_{e,e'}$ and ultimately, the error exponent K_P for learning the tree structure.

Remark 1: Theorem 2 specifies the crossover rate $J_{e,e'}$ when the two node pairs e and e' do not have any common nodes. If e and e' share one node, then the distribution $P_{e,e'} \in \mathcal{P}(\mathcal{X}^3)$ and here, the crossover rate for empirical mutual information is

$$J_{e,e'} = \inf_{Q \in \mathcal{P}(\mathcal{X}^3)} \{D(Q \| P_{e,e'}) : I(Q_{e'}) = I(Q_e)\}. \tag{23}$$

In Section VI, we obtain an approximate closed-form expression for $J_{e,e'}$. The expression, provided in Theorem 8, does not depend on whether e and e' share a node.

Example: Symmetric Star Graph: It is now instructive to study a simple example to see how the overall error exponent K_P for structure learning in (6) depends on the set of crossover rates $\left\{J_{e,e'} : e, e' \in \binom{\mathcal{V}}{2}\right\}$. We consider a graphical model P with an associated tree $T_P = (\mathcal{V}, \mathcal{E}_P)$ which is a d-order star with a central node 1 and outer nodes $2, \ldots, d$, as shown in Fig. 1. The edge set is given by $\mathcal{E}_P = \{(1, i) : i = 2, \ldots, d\}$.

We assign the joint distributions $Q_a, Q_b \in \mathcal{P}(\mathcal{X}^2)$ and $Q_{a,b} \in \mathcal{P}(\mathcal{X}^4)$ to the variables in this graph in the following specific way:

1) $P_{1,i} \equiv Q_a$ for all $2 \leq i \leq d$.
2) $P_{i,j} \equiv Q_b$ for all $2 \leq i, j \leq d, i \neq j$.

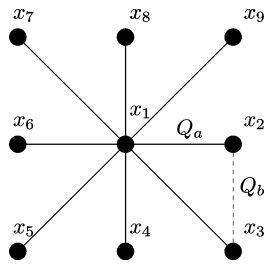

Fig. 1. Star graph with $d = 9$. Q_a is the joint distribution on any pair of variables that form an edge e.g., x_1 and x_2. Q_b is the joint distribution on any pair of variables that do not form an edge e.g., x_2 and x_3. By symmetry, all crossover rates are equal.

3) $P_{1,i,j,k} \equiv Q_{a,b}$ for all $2 \leq i, j, k \leq d, i \neq j \neq k$.

Thus, we have identical pairwise distributions $P_{1,i} \equiv Q_a$ of the central node 1 and any other node i, and also identical pairwise distributions $P_{i,j} \equiv Q_b$ of any two distinct outer nodes i and j. Furthermore, assume that $I(Q_a) > I(Q_b) > 0$. Note that the distribution $Q_{a,b} \in \mathcal{P}(\mathcal{X}^4)$ completely specifies the above graphical model with a star graph. Also, from the above specifications, we see that Q_a and Q_b are the marginal distributions of $Q_{a,b}$ with respect to to node pairs $(1, i)$ and (j, k) respectively i.e.,

$$Q_a(x_1, x_i) = \sum_{(x_j, x_k) \in \mathcal{X}^2} P_{1,i,j,k}(x_1, x_i, x_j, x_k) \tag{24a}$$

$$Q_b(x_j, x_k) = \sum_{(x_1, x_i) \in \mathcal{X}^2} P_{1,i,j,k}(x_1, x_i, x_j, x_k). \tag{24b}$$

Note that each crossover event between any nonedge e' (necessarily of length 2) and an edge e along its path results in an error in the learned structure since it leads to e' being declared an edge instead of e. Due to the symmetry, all such crossover rates between pairs e and e' are equal. By the "worst-exponent-wins" rule [11, Ch. 1], it is more likely to have a single crossover event than multiple ones. Hence, the error exponent is equal to the crossover rate between an edge and a non-neighbor pair in the symmetric star graph. We state this formally in the following proposition.

Proposition 3 (Error Exponent for Symmetric Star Graph): For the symmetric graphical model with star graph and $Q_{a,b}$ as described above, the error exponent for structure learning K_P in (6), is equal to the crossover rate between an edge and a non-neighbor node pair

$$K_P = J_{e,e'}, \quad \text{for any} \quad e \in \mathcal{E}_P, e' \notin \mathcal{E}_P \tag{25}$$

where from (21), the crossover rate is given by

$$J_{e,e'} = \inf_{R_{1,2,3,4} \in \mathcal{P}(\mathcal{X}^4)} \{D(R_{1,2,3,4} \| Q_{a,b}) : I(R_{1,2}) = I(R_{3,4})\} \tag{26}$$

with $R_{1,2}$ and $R_{3,4}$ as the marginals of $R_{1,2,3,4}$, e.g.,

$$R_{1,2}(x_1, x_2) = \sum_{(x_3, x_4) \in \mathcal{X}^2} R_{1,2,3,4}(x_1, x_2, x_3, x_4). \tag{27}$$

Proof: Since there are only two distinct distributions Q_a (which corresponds to a true edge) and Q_b (which corresponds

to a nonedge), there is only *one* unique rate $J_{e,e'}$, namely the expression in (21) with $P_{e,e'}$ replaced by $Q_{a,b}$. If the event $\mathcal{C}_{e,e'}$, in (19), occurs, an error definitely occurs. This corresponds to the case where *any one* edge $e \in \mathcal{E}_P$ is replaced by *any other* node pair e' not in $\mathcal{E}_P$.[8] ∎

Hence, we have derived the error exponent for learning a symmetric star graph through the crossover rate $J_{e,e'}$ between any node pair e which is an edge in the star graph and another node pair e' which is not an edge.

The symmetric star graph possesses symmetry in the distributions, and hence, it is easy to relate K_P to a sole crossover rate. In general, it is not straightforward to derive the error exponent K_P from the set of crossover rates $\{J_{e,e'}\}$ since they may not all be equal and more importantly, crossover events for different node pairs affect the learned structure $\mathcal{E}_{\mathrm{ML}}$ in a complex manner. In Section V, we provide an exact expression for K_P by identifying the (sole) crossover event related to a dominant error tree. Finally, we remark that the crossover event $\mathcal{C}_{e,e'}$ is related to the notion of neighborhood selection in the graphical model learning literature [5], [8].

V. Error Exponent for Structure Learning

The analysis in the previous section characterized the rate $J_{e,e'}$ for the crossover event $\mathcal{C}_{e,e'}$ between two empirical mutual information pairs. In this section, we connect these set of rate functions $\{J_{e,e'}\}$ to the quantity of interest, viz., the error exponent for ML-estimation of edge set K_P in (6).

Recall that the event $\mathcal{C}_{e,e'}$ denotes an error in estimating the order of mutual information quantities. However, such events $\mathcal{C}_{e,e'}$ need not necessarily lead to the error event $\mathcal{A}_n$ in (5) that the ML-estimate of the edge set $\mathcal{E}_{\mathrm{ML}}$ is different from the true set $\mathcal{E}_P$. This is because the ML-estimate $\mathcal{E}_{\mathrm{ML}}$ is a tree and this global constraint implies that certain crossover events can be ignored. In the sequel, we will identify useful crossover events through the notion of a dominant error tree.

A. Dominant Error Tree

We can decompose the error event for structure estimation $\mathcal{A}_n$ in (5) into a set of mutually-exclusive events

$$\mathbb{P}(\mathcal{A}_n) = \mathbb{P}\left(\bigcup_{T \in \mathcal{T}^d \setminus \{T_P\}} \mathcal{U}_n(T) \right) = \sum_{T \in \mathcal{T}^d \setminus \{T_P\}} \mathbb{P}(\mathcal{U}_n(T)) \tag{28}$$

where each $\mathcal{U}_n(T)$ denotes the event that the graph of the ML-estimate T_{ML} is a tree T different from the true tree T_P. In other words

$$\mathcal{U}_n(T) := \begin{cases} \{T_{\mathrm{ML}} = T\}, & \text{if } T \in \mathcal{T}^d \setminus \{T_P\} \\ \emptyset, & \text{if } T = T_P. \end{cases} \tag{29}$$

Note that $\mathcal{U}_n(T) \cap \mathcal{U}_n(T') = \emptyset$ whenever $T \neq T'$. The large-deviation rate or the exponent for each error event $\mathcal{U}_n(T)$ is

$$\Upsilon(T) := \lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}(\mathcal{U}_n(T)) \tag{30}$$

whenever the limit exists. Among all the error events $\mathcal{U}_n(T)$, we identify the dominant one with the slowest rate of decay.

Definition 1 (Dominant Error Tree): A dominant error tree $T_P^* = (\mathcal{V}, \mathcal{E}_P^*)$ is a spanning tree given by[9]

$$T_P^* := \operatorname*{arg\,max}_{T \in \mathcal{T}^d \setminus \{T_P\}} \Upsilon(T). \tag{31}$$

Roughly speaking, a dominant error tree is the tree that is the most-likely asymptotic output of the ML-estimator in the event of an error. Hence, it belongs to the set $\mathcal{T}^d \setminus \{T_P\}$. In the following, we note that the error exponent in (6) is equal to the exponent of the dominant error tree.

Proposition 4 (Dominant Error Tree & Error Exponent): The error exponent K_P for structure learning is equal to the exponent $\Upsilon(T_P^*)$ of the dominant error tree T_P^*

$$K_P = \Upsilon(T_P^*). \tag{32}$$

Proof: From (30), we can write

$$\mathbb{P}(\mathcal{U}_n(T)) \doteq \exp(-n\Upsilon(T)), \quad \forall T \in \mathcal{T}^d \setminus \{T_P\}. \tag{33}$$

Now from (28), we have

$$\mathbb{P}(\mathcal{A}_n) \doteq \sum_{T \in \mathcal{T}^d \setminus \{T_P\}} \exp(-n\Upsilon(T)) \doteq \exp(-n\Upsilon(T_P^*)) \tag{34}$$

from the "worst-exponent-wins" principle [11, Ch. 1] and the definition of the dominant error tree T_P^* in (31). ∎

Thus, by identifying a dominant error tree T_P^*, we can find the error exponent $K_P = \Upsilon(T_P^*)$. To this end, we revisit the crossover events $\mathcal{C}_{e,e'}$ in (19), studied in the previous section. Consider a non-neighbor node pair e' with respect to $\mathcal{E}_P$ and the unique path of edges in $\mathcal{E}_P$ connecting the two nodes, which we denote as $\mathrm{Path}(e'; \mathcal{E}_P)$. See Fig. 2, where we define the notion of the path given a nonedge e'. Note that e' and $\mathrm{Path}(e'; \mathcal{E}_P)$ necessarily form a cycle; if we replace any edge $e \in \mathcal{E}_P$ along the path of the non-neighbor node pair e', the resulting edge set $\mathcal{E}_P \setminus \{e\} \cup \{e'\}$ is still a spanning tree. Hence, all such replacements are feasible outputs of the ML-estimation in the event of an error. As a result, all such crossover events $\mathcal{C}_{e,e'}$ need to be considered for the error event for structure learning $\mathcal{A}_n$ in (5). However, for the error exponent K_P, again by the "worst-exponent-wins" principle, we only need to consider the crossover event between each non-neighbor node pair e' and its dominant replacement edge $r(e') \in \mathcal{E}_P$ defined below.

Definition 2 (Dominant Replacement Edge): For each non-neighbor node pair $e' \notin \mathcal{E}_P$, its dominant replacement edge $r(e') \in \mathcal{E}_P$ is defined as the edge in the unique path along $\mathcal{E}_P$ connecting the nodes in e' having the minimum crossover rate

$$r(e') := \operatorname*{arg\,min}_{e \in \mathrm{Path}(e'; \mathcal{E}_P)} J_{e,e'} \tag{35}$$

where the crossover rate $J_{e,e'}$ is given by (21).

[8] Also see Theorem 5 and its proof for the argument that the dominant error tree differs from the true tree by a single edge.

[9] We will use the notation $\arg\min$ extensively in the sequel. It is to be understood that if there is no unique minimum (e.g., in (31)), then we arbitrarily choose one of the minimizing solutions.

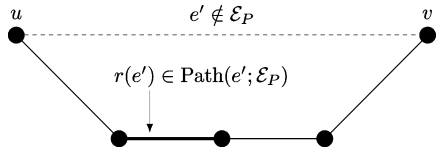

Fig. 2. Path associated to the nonedge $e' = (u, v) \notin \mathcal{E}_P$, denoted $\mathrm{Path}(e'; \mathcal{E}_P) \subset \mathcal{E}_P$, is the set of edges along the unique path linking the end points of $e' = (u, v)$. The edge $r(e')$ is the dominant replacement edge associated to $e' \notin \mathcal{E}_P$.

We are now ready to characterize the error exponent K_P in terms of the crossover rate between non-neighbor node pairs and their dominant replacement edges.

Theorem 5 (Error Exponent as a Single Crossover Event): The error exponent for ML-tree estimation in (6) is given by

$$K_P = J_{r(e^*),e^*} = \min_{e' \notin \mathcal{E}_P} \min_{e \in \mathrm{Path}(e'; \mathcal{E}_P)} J_{e,e'} \tag{36}$$

where $r(e^*)$ is the dominant replacement edge, defined in (35), associated to $e^* \notin \mathcal{E}_P$ and e^* is the optimizing non-neighbor node pair

$$e^* := \arg\min_{e' \notin \mathcal{E}_P} J_{r(e'),e'}. \tag{37}$$

The dominant error tree $T_P^* = (\mathcal{V}, \mathcal{E}_P^*)$ in (31) has edge set

$$\mathcal{E}_P^* = \mathcal{E}_P \cup \{e^*\} \setminus \{r(e^*)\}. \tag{38}$$

In fact, we also have the following (finite-sample) upper bound on the error probability:

$$\mathbb{P}(\mathcal{A}_n) \leq \frac{(d-1)^2(d-2)}{2} \binom{n+1+|\mathcal{X}|^4}{n+1} \exp(-nK_P) \tag{39}$$

for all $n \in \mathbb{N}$.

Proof: (Sketch) The edge set of the dominant error tree $\mathcal{E}_P^*$ differs from $\mathcal{E}_P$ in exactly one edge (See Appendix B). This is because if $\mathcal{E}_P^*$ were to differ from $\mathcal{E}_P$ in strictly more than one edge, the resulting error exponent would not be the minimum, hence contradicting Proposition 4. To identify the dominant error tree, we use the union bound as in (28) and the "worst-exponent-wins" principle [11, Ch. 1], to conclude that the rate that dominates is the minimum $J_{r(e'),e'}$ over all possible non-neighbor node pairs $e' \notin \mathcal{E}_P$. See Appendix B for the details. ∎

The above theorem relates the set of crossover rates $\{J_{e,e'}\}$, which we characterized in the previous section, to the overall error exponent K_P, defined in (6). Note that the result in (36) and also the existence of the limit in (6) means that the error probability is tight to first order in the exponent in the sense that $\mathbb{P}(\mathcal{A}_n) \doteq \exp(-nK_P)$. This is in contrast to the work in [21], where bounds on the upper and lower limit on the sequence $-\frac{1}{n}\log \mathbb{P}(\mathcal{A}_n)$ were established.[10] We numerically compute the error exponent K_P for different discrete distributions in Section VIII.

[10]However, in [21], the authors analyzed the learning of general (non-tree) Bayesian networks.

From (36), we see that if at least one of the crossover rates $J_{e,e'}$ in the minimization is zero, the overall error exponent K_P is zero. This observation is important for the derivation of necessary and sufficient conditions for K_P to be positive, and hence, for the error probability to decay exponentially in the number of samples n.

B. Conditions for Exponential Decay

We now provide necessary and sufficient conditions that ensure that K_P is strictly positive. This is obviously of crucial importance since if $K_P > 0$, this implies exponential decay of the desired probability of error $\mathbb{P}(\mathcal{A}_n)$, where the error event $\mathcal{A}_n$ is defined in (5).

Theorem 6 (Equivalent Conditions for Exponential Decay): Assume that T_P, the original structure is acyclic (i.e., it may not be connected). Then, the following three statements are equivalent.

(a) The probability of error $\mathbb{P}(\mathcal{A}_n)$ decays exponentially i.e.,

$$K_P > 0. \tag{40}$$

(b) The mutual information quantities satisfy

$$I(P_{e'}) < I(P_e), \quad \forall\, e \in \mathrm{Path}\,(e'; \mathcal{E}_P)\,, e' \notin \mathcal{E}_P. \tag{41}$$

(c) T_P is not a proper forest.[11]

Proof: (*Sketch*) We first show that (a) ⇔ (b).

(⇒) We assume statement (a) is true i.e., $K_P > 0$ and prove that statement (b) is true. Suppose, to the contrary, that $I(P_{e'}) = I(P_e)$ for some $e \in \mathrm{Path}\,(e'; \mathcal{E}_P)$ and some $e' \notin \mathcal{E}_P$. Then $J_{r(e'),e'} = 0$, where $r(e')$ is the replacement edge associated to e'. By (36), $K_P = 0$, which is a contradiction.

(⇐) We now prove that statement (a) is true assuming statement (b) is true i.e., $I(P_{e'}) < I(P_e)$ for all $e \in \mathrm{Path}\,(e'; \mathcal{E}_P)$ and $e' \notin \mathcal{E}_P$. By Theorem 2, the crossover rate $J_{r(e'),e'}$ in (21) is positive for all $e' \notin \mathcal{E}_P$. From (36), $K_P > 0$ since there are only finitely many e'; hence, the minimum in (37) is attained at some nonzero value, i.e., $K_P = \min_{e' \notin \mathcal{E}_P} J_{r(e'),e'} > 0$.

Statement (c) is equivalent to statement (b). The proof of this claim makes use of the positivity condition that $P(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}^d$ and the fact that if variables x_1, x_2 and x_3 form Markov chains $x_1 - x_2 - x_3$ and $x_1 - x_3 - x_2$, then x_1 is necessarily *jointly independent* of (x_2, x_3). Since this proof is rather lengthy, we refer the reader to Appendix C for the details. ∎

Condition (b) states that, for each nonedge e', we need $I(P_{e'})$ to be strictly smaller than the mutual information of its dominant replacement edge $I(P_{r(e')})$. Condition (c) is a more intuitive condition for exponential decay of the probability of error $\mathbb{P}(\mathcal{A}_n)$. This is an important result since it says that for *any* nondegenerate tree distribution in which all the pairwise joint distributions are not product distributions (i.e., not a proper forest), then we have exponential decay in the error probability. The learning of proper forests is discussed in a companion paper [38].

[11]A proper forest on d nodes is an undirected, acyclic graph that has (strictly) fewer than $d - 1$ edges.

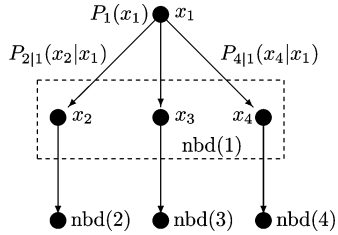

Fig. 3. Illustration for Example 1.

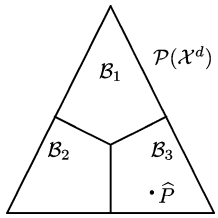

Fig. 4. Partitions of the simplex associated to our learning problem are given by $\mathcal{B}_i$, defined in (44). In this example, the type $\widehat{P}$ belongs to $\mathcal{B}_3$ so the tree associated to partition $\mathcal{B}_3$ is favored.

In the following example, we describe a simple random process for constructing a distribution P such that all three conditions in Theorem 6 are satisfied with probability one (w.p. 1). See Fig. 3.

Example 1: Suppose the structure of P, a spanning tree distribution with graph $T_P = (\mathcal{V}, \mathcal{E}_P)$, is fixed and $\mathcal{X} = \{0,1\}$. Now, we assign the parameters of P using the following procedure. Let x_1 be the root node. Then randomly draw the parameter of the Bernoulli distribution $P_1(x_1)$ from a uniform distribution on $[0,1]$ i.e., $P_1(x_1 = 0) = \theta_{x_1^0}$ and $\theta_{x_1^0} \sim \mathcal{U}[0,1]$. Next let $\mathrm{nbd}(1)$ be the set of neighbors of x_1. Regard the set of variables $\{x_j : j \in \mathrm{nbd}(1)\}$ as the children[12] of x_1. For each $j \in \mathrm{nbd}(1)$, sample both $P(x_j = 0|x_1 = 0) = \theta_{x_j^0|x_1^0}$ as well as $P(x_j = 0|x_1 = 1) = \theta_{x_j^0|x_1^1}$ from independent uniform distributions on $[0,1]$ i.e., $\theta_{x_j^0|x_1^0} \sim \mathcal{U}[0,1]$ and $\theta_{x_j^0|x_1^1} \sim \mathcal{U}[0,1]$. Repeat this procedure for all children of x_1. Then repeat the process for all other children. This construction results in a joint distribution $P(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}^d$ w.p. 1. In this case, by continuity, all mutual informations are distinct w.p. 1, the graph is not a proper forest w.p. 1 and the rate $K_P > 0$ w.p. 1.

This example demonstrates that $\mathbb{P}(\mathcal{A}_n)$ decays exponentially for *almost every* tree distribution.

C. Computational Complexity

Finally, we provide an upper bound on the computational complexity to compute K_P in (36). Our upper bound on the computational complexity depends on the *diameter* of the tree $T_P = (\mathcal{V}, \mathcal{E}_P)$ which is defined as

$$\mathrm{diam}(T_P) := \max_{u,v \in \mathcal{V}} L(u,v) \tag{42}$$

where $L(u,v)$ is the length (number of hops) of the unique path between nodes u and v. For example, $L(u,v) = 4$ for the nonedge $e' = (u,v)$ in the subtree in Fig. 2.

Theorem 7 (Computational Complexity for K_P): The number of computations of $J_{e,e'}$ to compute K_P, denoted $N(T_P)$, satisfies

$$N(T_P) \leq \frac{1}{2}\mathrm{diam}(T_P)(d-1)(d-2). \tag{43}$$

Proof: Given a non-neighbor node pair $e' \notin \mathcal{E}_P$, we perform a maximum of $\mathrm{diam}(T_P)$ calculations to determine the dominant replacement edge $r(e')$ from (35). Combining this with the fact that there are a total of $\left|\binom{\mathcal{V}}{2} \setminus \mathcal{E}_P\right| = \binom{d}{2} - (d-1) = \frac{1}{2}(d-1)(d-2)$ node pairs not in $\mathcal{E}_P$, we obtain the upper bound. ■

Thus, if the diameter of the tree $\mathrm{diam}(T_P)$ is relatively low and independent of number of nodes d, the complexity is quadratic in d. For instance, for a star graph, the diameter $\mathrm{diam}(T_P) = 2$. For a balanced tree,[13] $\mathrm{diam}(T_P) = \mathcal{O}(\log d)$; hence, the number of computations is $\mathcal{O}(d^2 \log d)$.

D. Relation of the Maximum-Likelihood Structure Learning Problem to Robust Hypothesis Testing

We now take a short detour and discuss the relation between the analysis of the learning problem and robust hypothesis testing, which was first considered by Huber and Strassen in [39]. Subsequent work was done in [40]–[42] albeit for differently defined uncertainty classes known as moment classes.

We hereby consider an alternative but related problem. Let $T_1, \ldots, T_M$ be the $M = d^{d-2}$ trees with d nodes. Also let $\mathcal{Q}_1, \ldots, \mathcal{Q}_M \subset \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)$ be the subsets of tree-structured graphical models Markov on $T_1, \ldots, T_M$ respectively. The structure learning problem is similar to the M-ary hypothesis testing problem between the uncertainty classes of distributions $\mathcal{Q}_1, \ldots, \mathcal{Q}_M$. The uncertainty class $\mathcal{Q}_i$ denotes the set of tree-structured graphical models with different *parameters* (marginal $\{P_i : i \in \mathcal{V}\}$ and pairwise distributions $\{P_{i,j} : (i,j) \in \mathcal{E}_P\}$) but Markov on the same tree T_i.

In addition, we note that the probability simplex $\mathcal{P}(\mathcal{X}^d)$ can be partitioned into M subsets[14] $\mathcal{B}_1, \ldots, \mathcal{B}_M \subset \mathcal{P}(\mathcal{X}^d)$ where each $\mathcal{B}_i, i = 1, \ldots, M$ is defined as

$$\mathcal{B}_i := \bigcup_{P' \in \mathcal{Q}_i} \left\{ Q : D(P'\|Q) \leq \min_{R \in \cup_{j \neq i} \mathcal{Q}_i} D(P'\|R) \right\}. \tag{44}$$

See Fig. 4. According to the ML criterion in (9), if the type $\widehat{P}$ belongs to $\mathcal{B}_i$, then the i-th tree is favored.

In [43], a subset of the authors of this paper considered the Neyman-Pearson setup of a robust binary hypothesis testing problem where the null hypothesis corresponds to the true tree model P and the (composite) alternative hypothesis corresponds to the set of distributions Markov on some erroneous tree $T_Q \neq T_P$. The false-alarm probability was constrained to be smaller

[12]Let x_1 be the root of the tree. In general, the children of a node $x_k (k \neq 1)$ is the set of nodes connected to x_k that are further away from the root than x_k.

[13]A balanced tree is one where no leaf is much farther away from the root than any other leaf. The length of the longest direct path between any pair of nodes is $\mathcal{O}(\log d)$.

[14]From the definition in (44), we see that the relative interior of the subsets are pairwise disjoint. We discuss the scenario when P lies on the boundaries of these subsets in Section VII.

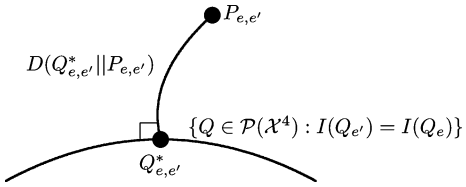

Fig. 5. A geometric interpretation of (21) where $P_{e,e'}$ is projected onto the submanifold of probability distributions $\{Q \in \mathcal{P}(\mathcal{X}^4) : I(Q_{e'}) = I(Q_e)\}$.

than $\alpha > 0$ and optimized for worst-case type-II (missed detection) error exponent using the Chernoff-Stein Lemma [30, Ch. 12]. It was established that the worst-case error exponent can be expressed in closed-form in terms of the mutual information of so-called *bottleneck edges*, i.e., the edge and nonedge pair that have the smallest mutual information difference. However, in general, for the binary hypothesis testing problem, the error event *does not* decompose into a union of local events. This is in contrast to error exponent for learning the ML tree K_P, which can be computed by considering local crossover events $\mathcal{C}_{e,e'}$, defined in (19).

Note that $\{\widehat{P} \in \mathcal{B}_i\}$ corresponds to a *global event* since each $\mathcal{B}_i \subset \mathcal{P}(\mathcal{X}^d)$. The large-deviation analysis techniques we utilized to obtain the error exponent K_P in Theorem 5 show that such global error events can be also decomposed into a collection of local crossover events $\mathcal{C}_{e,e'}$. These local events depend only on the type *restricted* to pairs of nodes e and e' and are more intuitive for assessing (and analyzing) when and how an error can occur during the Chow-Liu learning process.

VI. Euclidean Approximations

In order to gain more insight into the error exponent, we make use of *Euclidean approximations* [15] of information-theoretic quantities to obtain an approximate but closed-form solution to (21), which is nonconvex and hard to solve exactly. In addition, we note that the dominant error event results from an edge and a nonedge that satisfy the conditions for which the Euclidean approximation is valid, i.e., the very-noisy condition given later in Definition 4. This justifies our approach we adopt in this section. Our use of Euclidean approximations for various information-theoretic quantities is akin to various problems considered in other contexts in information theory [14], [15], [44].

We first approximate the crossover rate $J_{e,e'}$ for any two node pairs e and e', which do not share a common node. The joint distribution on e and e', namely $P_{e,e'}$ belongs to the set $\mathcal{P}(\mathcal{X}^4)$. Intuitively, the crossover rate $J_{e,e'}$ should depend on the "separation" of the mutual information values $I(P_e)$ and $I(P_{e'})$, and also on the uncertainty of the difference between mutual information estimates $I(\widehat{P}_e)$ and $I(\widehat{P}_{e'})$. We will see that the approximate rate also depends on these mutual information quantities given by a simple expression which can be regarded as the signal-to-noise ratio (SNR) for learning.

Roughly speaking, our strategy is to "convexify" the objective and the constraints in (21). See Figs. 5 and 6. To do so, we recall that if P and Q are two discrete distributions with the same support $\mathcal{Y}$, and they are close entry-wise, the KL divergence can be approximated [15] as

$$\begin{aligned} &D(Q\|P) \\ &= -\sum_{a\in\mathcal{Y}} Q(a)\log\frac{P(a)}{Q(a)} && (45)\\ &= -\sum_{a\in\mathcal{Y}} Q(a)\log\left[1+\left(\frac{P(a)-Q(a)}{Q(a)}\right)\right] && (46)\\ &= \frac{1}{2}\sum_{a\in\mathcal{Y}}\frac{(Q(a)-P(a))^2}{Q(a)} + o\left(\|Q-P\|_\infty^2\right) && (47)\\ &= \frac{1}{2}\|Q-P\|_Q^2 + o\left(\|Q-P\|_\infty^2\right) && (48)\end{aligned}$$

where $\|y\|_w^2$ denotes the weighted squared norm of y, i.e., $\|y\|_w^2 := \sum_i y_i^2/w_i$. The equality in (47) holds because $\log(1+t) = \sum_{i=1}^{\infty}(-1)^{i+1}t^i/i$ for $t \in (-1, 1]$. The difference between the divergence and the Euclidean approximation becomes tight as $\epsilon = \|P-Q\|_\infty \to 0$. Moreover, it remains tight even if the subscript Q in (48) is changed to a distribution Q' in the vicinity of Q [15]. That is, the difference between $\|Q-P\|_Q$ and $\|Q-P\|_{Q'}$ is negligible compared to either term when $Q' \approx Q$. Using this fact and the assumption that P and Q are two discrete distributions that are close entry-wise

$$D(Q\|P) \approx \frac{1}{2}\|Q-P\|_P^2. \quad (49)$$

In fact, it is also known [15] that if $\|P-Q\|_\infty < \epsilon$ for some $\epsilon > 0$, we also have $D(P\|Q) \approx D(Q\|P)$.

In the following, to make our statements precise, we will use the notation $\alpha_1 \approx_\delta \alpha_2$ to denote that two real numbers α_1 and α_2 are in the δ neighborhood of each other, i.e., $|\alpha_1 - \alpha_2| < \delta$.[15] We will also need the following notion of information density to state our approximation for $J_{e,e'}$.

Definition 3 (Information Density): Given a pairwise joint distribution $P_{i,j}$ on $\mathcal{X}^2$ with marginals P_i and P_j, the *information density* [45], [46] function, denoted by $s_{i,j} : \mathcal{X}^2 \to \mathbb{R}$, is defined as

$$s_{i,j}(x_i, x_j) := \log\frac{P_{i,j}(x_i,x_j)}{P_i(x_i)P_j(x_j)}, \quad \forall\,(x_i,x_j) \in \mathcal{X}^2. \quad (50)$$

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

Recall that we also assumed in Section II that T_P is a spanning tree, which implies that for all node pairs (i, j), $P_{i,j}$ is *not* a product distribution, i.e., $P_{i,j} \neq P_iP_j$, because if it were, then T_P would be disconnected. We now define a condition for which our approximation holds.

[15] In the following, we will also have continuity statements where given $\epsilon > 0$ and $\alpha_1 \approx_\epsilon \alpha_2$, implies that there exists some $\delta = \delta(\epsilon) > 0$ such that $\beta_1 \approx_\delta \beta_2$. We will be casual about specifying what the δ's are.

Definition 4 (ϵ-Very Noisy Condition): We say that $P_{e,e'} \in \mathcal{P}(\mathcal{X}^4)$, the joint distribution on node pairs e and e', satisfies the ϵ-very noisy condition if

$$\|P_e - P_{e'}\|_\infty := \max_{(x_i,x_j)\in\mathcal{X}^2} |P_e(x_i,x_j) - P_{e'}(x_i,x_j)| < \epsilon. \tag{51}$$

This condition is needed because if (51) holds, then by continuity of the mutual information, there exists a $\delta > 0$ such that $I(P_e) \approx_\delta I(P_{e'})$, which means that the mutual information quantities are difficult to distinguish and the approximation in (48) is accurate.[16] Note that proximity of the mutual information values is not sufficient for the approximation to hold since we have seen from Theorem 2 that $J_{e,e'}$ depends not only on the mutual information quantities but on the entire joint distribution $P_{e,e'}$.

We now define the approximate crossover rate on disjoint node pairs e and e' as

$$\widetilde{J}_{e,e'} := \inf\left\{\frac{1}{2}\|Q - P_{e,e'}\|^2_{P_{e,e'}} : Q \in \mathcal{L}(P_{e,e'})\right\} \tag{52}$$

where the (linearized) constraint set is

$$\begin{aligned}\mathcal{L}(P_{e,e'}) := \Big\{Q \in \mathcal{P}(\mathcal{X}^4) : I(P_e) + \langle \nabla_{P_e} I(P_e), Q - P_{e,e'}\rangle \\ = I(P_{e'}) + \left\langle \nabla_{P_{e'}} I(P_{e'}), Q - P_{e,e'}\right\rangle\Big\}\end{aligned} \tag{53}$$

where $\nabla_{P_e} I(P_e)$ is the gradient vector of the mutual information with respect to the joint distribution P_e, regarded as a length-$|\mathcal{X}|^2$ vector. We also define the approximate error exponent as

$$\widetilde{K}_P := \min_{e'\notin\mathcal{E}_P} \min_{e\in \mathrm{Path}(e';\mathcal{E}_P)} \widetilde{J}_{e,e'}. \tag{54}$$

We now provide the expression for the approximate crossover rate $\widetilde{J}_{e,e'}$ and also state the conditions under which the approximation is asymptotically accurate in ϵ.[17]

Theorem 8 (Euclidean Approximation of $J_{e,e'}$): The approximate crossover rate for the empirical mutual information quantities, defined in (52), is given by

$$\widetilde{J}_{e,e'} = \frac{(\mathbb{E}[s_{e'} - s_e])^2}{2\mathrm{Var}(s_{e'} - s_e)} = \frac{(I(P_{e'}) - I(P_e))^2}{2\mathrm{Var}(s_{e'} - s_e)} \tag{55}$$

where s_e is the information density defined in (50) and the expectation and variance are both with respect to $P_{e,e'}$. Furthermore, the approximation (55) is asymptotically accurate, i.e., as $\epsilon \to 0$ (in the definition of ϵ-very noisy condition), we have that $\widetilde{J}_{e,e'} \to J_{e,e'}$.

[16]Here and in the following, we do not specify the exact value of δ but we simply note that as $\epsilon \to 0$, the approximation in (49) becomes tighter.

[17]We say that a collection of approximations $\{\widetilde{\theta}(\epsilon) : \epsilon > 0\}$ of a true parameter θ is asymptotically accurate in ϵ (or simply asymptotically accurate) if the approximations converge to θ as $\epsilon \to 0$, i.e., $\lim_{\epsilon\to 0} \widetilde{\theta}(\epsilon) = \theta$.

Proof: (*Sketch*) (52) and (53) together define a least squares problem. Upon simiplification of the solution, we obtain (55). See Appendix D for the details. ∎

We also have an additional result for the Euclidean approximation for the overall error exponent K_P. The proof is clear from the definition of $\widetilde{K}_P$ in (54) and the continuity of the min function.

Corollary 9 (Euclidean Approximation of K_P): The approximate error exponent $\widetilde{K}_P$ is asymptotically accurate if all joint distributions in the set $\{P_{e,e'} : e \in \mathrm{Path}(e;\mathcal{E}_P), e' \notin \mathcal{E}_P\}$ satisfy the ϵ-very noisy condition.

Hence, the expressions for the crossover rate $J_{e,e'}$ and the error exponent K_P are vastly simplified under the ϵ-very noisy condition on the joint distributions $P_{e,e'}$. The approximate crossover rate $\widetilde{J}_{e,e'}$ in (55) has a very intuitive meaning. It is proportional to the square of the difference between the mutual information quantities of P_e and $P_{e'}$. This corresponds exactly to our initial intuition—that if $I(P_e)$ and $I(P_{e'})$ are well separated $(I(P_e) \gg I(P_{e'}))$ then the crossover rate has to be large. $\widetilde{J}_{e,e'}$ is also weighted by the precision (inverse variance) of $(s_{e'} - s_e)$. If this variance is large then we are uncertain about the estimate $I(\widehat{P}_e) - I(\widehat{P}_{e'})$, and crossovers are more likely, thereby reducing the crossover rate $\widetilde{J}_{e,e'}$.

We now comment on our assumption of $P_{e,e'}$ satisfying the ϵ-very noisy condition, under which the approximation is tight as seen in Theorem 8. When $P_{e,e'}$ is ϵ-very noisy, then we have $I(P_e) \approx_\delta I(P_{e'})$, which implies that the optimal solution $Q^*_{e,e'} \approx_{\delta'} P_{e,e'}$. When e is an edge and e' is a non-neighbor node pair, this implies that it is very hard to distinguish the relative magnitudes of the empiricals $I(\widehat{P}_e)$ and $I(\widehat{P}_{e'})$. Hence, the particular problem of learning the distribution $P_{e,e'}$ from samples is *very noisy*. Under these conditions, the approximation in (55) is accurate.

In summary, our approximation in (55) takes into account not only the absolute difference between the mutual information quantities $I(P_e)$ and $I(P_{e'})$, but also the uncertainty in learning them. The expression in (55) is, in fact, the SNR for the estimation of the difference between empirical mutual information quantities. This answers one of the fundamental questions we posed in the introduction, viz., that we are now able to distinguish between distributions that are "easy" to learn and those that are "difficult" by computing the set of SNR quantities $\{\widetilde{J}_{e,e'}\}$ in (55).

VII. Extensions to Non-Tree Distributions

In all the preceding sections, we dealt exclusively with the case where the true distribution P is Markov on a tree. In this section, we extend the preceding large-deviation analysis to deal with distributions P that may not be tree-structured but in which we estimate a tree distribution from the given set of samples $\mathbf{x}^n$, using the Chow-Liu ML-estimation procedure. Since the Chow-Liu procedure outputs a tree, it is not possible to learn the structure of P correctly. Hence, it will be necessary to redefine the error event.

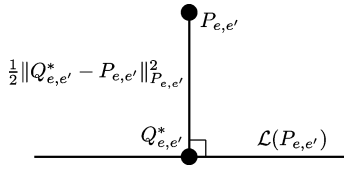

Fig. 6. Convexifying the objective results in a least-squares problem. The objective is converted into a quadratic as in (52) and the linearized constraint set $\mathcal{L}(P_{e,e'})$ is given (53).

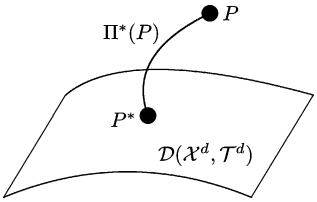

Fig. 7. Reverse I-projection [13] of P onto the set of tree distributions $\mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)$ given by (56).

TABLE I
TABLE OF PROBABILITY VALUES FOR EXAMPLE 2

x_1	x_2	x_3	Distribution $P(\mathbf{x})$
0	0	0	$(1/2-\xi)(1/2-\kappa)$
0	0	1	$(1/2+\xi)(1/2-\kappa)$
0	1	0	$(1/3+\xi)\kappa$
0	1	1	$(2/3-\xi)\kappa$
1	0	0	$(2/3-\xi)\kappa$
1	0	1	$(1/3+\xi)\kappa$
1	1	0	$(1/2-\xi)(1/2-\kappa)$
1	1	1	$(1/2+\xi)(1/2-\kappa)$

When P is not a tree distribution, we analyze the properties of the optimal *reverse I-projection* [13] of P onto the set of tree distributions, given by the optimization problem[18]

$$\Pi^*(P) := \min_{Q \in \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)} D(P\|Q). \tag{56}$$

$\Pi^*(P)$ is the KL-divergence of P to the closest element in $\mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)$. See Fig. 7. As Chow and Wagner [10] noted, if P is not a tree, there may be several trees optimizing (56).[19] We denote the set of optimal projections as $\mathcal{P}^*(P)$, given by

$$\mathcal{P}^*(P) := \{Q \in \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d) : D(P\|Q) = \Pi^*(P)\}. \tag{57}$$

We now illustrate that $\mathcal{P}^*(P)$ may have more than one element with the following example.

Example 2: Consider the parameterized discrete probability distribution $P \in \mathcal{P}(\{0,1\}^3)$ shown in Table I where $\xi \in (0, 1/3)$ and $\kappa \in (0, 1/2)$ are constants.

Proposition 10 (Nonuniqueness of Projection): For sufficiently small κ, the Chow-Liu MWST algorithm (using either Kruskal's [35] or Prim's [36] procedure) will first include the edge (1, 2). Then, it will arbitrarily choose between the two remaining edges (2, 3) or (1, 3).

[18]The minimum in the optimization problem in (56) is attained because the KL-divergence is continuous and the set of tree distributions $\mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)$ is compact.

[19]This is a technical condition of theoretical interest in this section. In fact, it can be shown that the set of distributions such that there is more than one tree optimizing (56) has ($\mathbb{R}^{|\mathcal{X}|^d} - 1$-Lebesgue) measure zero in $\mathcal{P}(\mathcal{X}^d)$.

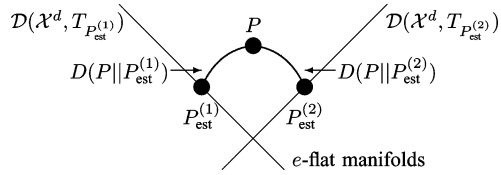

Fig. 8. Each tree defines an e-flat submanifold [47], [48] of probability distributions. These are the two lines as shown in the figure. If the KL-divergences $D(P\|P^{(1)}_{\text{est}})$ and $D(P\|P^{(2)}_{\text{est}})$ are equal, then $P^{(1)}_{\text{est}}$ and $P^{(2)}_{\text{est}}$ do not have the same structure but both are optimal with respect to the optimization problem in (56). An example of such a distribution P is provided in Example 2.

The proof of this proposition is provided in Appendix E where we show that $I(P_{1,2}) > I(P_{2,3}) = I(P_{1,3})$ for sufficiently small κ. Thus, the optimal tree structure P^* is not unique. This in fact corresponds to the case where P belongs to the boundary of some set $\mathcal{B}_i \subset \mathcal{P}(\mathcal{X}^d)$ defined in (44). See Fig. 8 for an information geometric interpretation.

Every tree distribution in $\mathcal{P}^*(P)$ has the maximum sum mutual information weight. More precisely, we have

$$\sum_{e \in \mathcal{E}_Q} I(Q_e) = \max_{Q' \in \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)} \sum_{e \in \mathcal{E}_{Q'}} I(Q'_e), \forall\, Q \in \mathcal{P}^*(P). \tag{58}$$

Given (58), we note that when we use a MWST algorithm to find the optimal solution to the problem in (56), ties will be encountered during the greedy addition of edges, as demonstrated in Example 2. Upon breaking the ties arbitrarily, we obtain some distribution $Q \in \mathcal{P}^*(P)$. We now provide a sequence of useful definitions that lead to definition of a new error event for which we can perform large-deviation analysis.

We denote the set of tree structures[20] corresponding to the distributions in $\mathcal{P}^*(P)$ as

$$\mathcal{T}_{\mathcal{P}^*(P)} := \left\{T_Q \in \mathcal{T}^d : Q \in \mathcal{P}^*(P)\right\} \tag{59}$$

and term it as the set of optimal tree projections. A similar definition applies to the edge sets of optimal tree projections

$$\mathcal{E}_{\mathcal{P}^*(P)} := \left\{\mathcal{E}_Q : T_Q = (\mathcal{V}, \mathcal{E}_Q) \in \mathcal{T}^d, Q \in \mathcal{P}^*(P)\right\}. \tag{60}$$

Since the distribution P is unknown, our goal is to estimate the optimal tree-projection P_{est} using the empirical distribution $\widehat{P}$, where P_{est} is given by

$$P_{\text{est}} := \operatorname*{arg\,min}_{Q \in \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)} D(\widehat{P}\|Q). \tag{61}$$

If there are many distributions Q, we arbitrarily pick one of them. We will see that by redefining the error event, we will have still a LDP. Finding the reverse I-projection P_{est} can be solved efficiently (in time $\mathcal{O}(d^2 \log d)$) using the Chow-Liu algorithm [3] as described in Section III.

[20]In fact, each tree defines a so-called *e-flat submanifold* [47], [48] in the set of probability distributions on $\mathcal{X}^d$ and P_{est} lies in both submanifolds. The so-called *m-geodesic* connects P to any of its optimal projection $P_{\text{est}} \in \mathcal{P}^*(P)$.

We define $T_{P_{\text{est}}} = (\mathcal{V}, \mathcal{E}_{P_{\text{est}}})$ as the graph of P_{est}, which is the learned tree and redefine the new *error event* as

$$\mathcal{A}_n(\mathcal{P}^*(P)) := \left\{\mathcal{E}_{P_{\text{est}}} \notin \mathcal{E}_{\mathcal{P}^*(P)}\right\}. \tag{62}$$

Note that this new error event essentially reduces to the original error event $\mathcal{A}_n = \mathcal{A}_n(\{P\})$ in (5) if $\mathcal{T}_{\mathcal{P}^*(P)}$ contains only one member. So if the learned structure belongs to $\mathcal{E}_{\mathcal{P}^*(P)}$, there is no error, otherwise an error is declared. We would like to analyze the decay of the error probability of $\mathcal{A}_n(\mathcal{P}^*(P))$ as defined in (62), i.e., find the new *error exponent*

$$K_{\mathcal{P}^*(P)} := \lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}\left(\mathcal{A}_n\left(\mathcal{P}^*(P)\right)\right). \tag{63}$$

It turns out that the analysis of the new event $\mathcal{A}_n(\mathcal{P}^*(P))$ is very similar to the analysis performed in Section V. We redefine the notion of a dominant replacement edge and the computation of the new rate $K_{\mathcal{P}^*(P)}$ then follows automatically.

Definition 5 (Dominant Replacement Edge): Fix an edge set $\mathcal{E}_Q \in \mathcal{E}_{\mathcal{P}^*(P)}$. For the error event $\mathcal{A}_n(\mathcal{P}^*(P))$ defined in (62), given a non-neighbor node pair $e' \notin \mathcal{E}_Q$, its dominant replacement edge $r(e';\mathcal{E}_Q)$ with respect to $\mathcal{E}_Q$, is given by

$$r(e';\mathcal{E}_Q) := \operatorname*{arg\,min}_{\substack{e\in \mathrm{Path}(e';\mathcal{E}_Q)\\ \mathcal{E}_Q\cup\{e'\}\setminus\{e\}\notin\mathcal{E}_{\mathcal{P}^*(P)}}} J_{e,e'} \tag{64}$$

if there exists an edge $e \in \mathrm{Path}(e';\mathcal{E}_Q)$ such that $\mathcal{E}_Q \cup \{e'\} \setminus \{e\} \notin \mathcal{E}_{\mathcal{P}^*(P)}$. Otherwise $r(e';\mathcal{E}_Q) = \emptyset$. $J_{e,e'}$ is the crossover rate of mutual information quantities defined in (20). If $r(e';\mathcal{E}_Q)$ exists, the corresponding crossover rate is

$$J_{r(e';\mathcal{E}_Q),e'} = \min_{\substack{e\in \mathrm{Path}(e';\mathcal{E}_Q)\\ \mathcal{E}_Q\cup\{e'\}\setminus\{e\}\notin\mathcal{E}_{\mathcal{P}^*(P)}}} J_{e,e'} \tag{65}$$

otherwise $J_{\emptyset,e'} = +\infty$.

In (64), we are basically fixing an edge set $\mathcal{E}_Q \in \mathcal{E}_{\mathcal{P}^*(P)}$ and excluding the trees with $e \in \mathrm{Path}(e';\mathcal{E}_Q)$ replaced by e' if it belongs to the set of optimal tree projections $\mathcal{T}_{\mathcal{P}^*(P)}$. We further remark that in (64), $r(e')$ may not necessarily exist. Indeed, this occurs if every tree with $e \in \mathrm{Path}(e';\mathcal{E}_Q)$ replaced by e' belongs to the set of optimal tree projections. This is, however, *not* an error by the definition of the error event in (62); hence, we set $J_{\emptyset,e'} = +\infty$. In addition, we define the *dominant nonedge* associated to edge set $\mathcal{E}_Q \in \mathcal{E}_{\mathcal{P}^*(P)}$ as

$$e^*(\mathcal{E}_Q) := \operatorname*{arg\,min}_{e'\notin\mathcal{E}_Q} \min_{\substack{e\in \mathrm{Path}(e';\mathcal{E}_Q)\\ \mathcal{E}_Q\cup\{e'\}\setminus\{e\}\notin\mathcal{E}_{\mathcal{P}^*(P)}}} J_{e,e'}. \tag{66}$$

Also, the *dominant structure* in the set of optimal tree projections is defined as

$$\mathcal{E}_{P^*} := \operatorname*{arg\,min}_{\mathcal{E}_Q\in\mathcal{E}_{\mathcal{P}^*(P)}} J_{r(e^*(\mathcal{E}_Q);\mathcal{E}_Q),e^*(\mathcal{E}_Q)} \tag{67}$$

where the crossover rate $J_{r(e';\mathcal{E}_Q),e'}$ is defined in (65) and the dominant nonedge $e^*(\mathcal{E}_Q)$ associated to $\mathcal{E}_Q$ is defined in (66). Equipped with these definitions, we are now ready to state the generalization of Theorem 5.

Theorem 11 (Dominant Error Tree): For the error event $\mathcal{A}_n(\mathcal{P}^*(P))$ defined in (62), a dominant error tree (which may not be unique) has edge set given by

$$\mathcal{E}_{P^*} \cup \{e^*(\mathcal{E}_{P^*})\} \setminus \{r(e^*(\mathcal{E}_{P^*});\mathcal{E}_{P^*})\} \tag{68}$$

where $e^*(\mathcal{E}_{P^*})$ is the dominant nonedge associated to the dominant structure $\mathcal{E}_{P^*} \in \mathcal{E}_{\mathcal{P}^*(P)}$ and is defined by (66) and (67). Furthermore, the error exponent $K_{\mathcal{P}^*(P)}$, defined in (63) is given as

$$K_{\mathcal{P}^*(P)} = \min_{\mathcal{E}_Q\in\mathcal{E}_{\mathcal{P}^*(P)}} \min_{e'\notin\mathcal{E}_Q} \min_{\substack{e\in \mathrm{Path}(e';\mathcal{E}_Q)\\ \mathcal{E}_Q\cup\{e'\}\setminus\{e\}\notin\mathcal{E}_{\mathcal{P}^*(P)}}} J_{e,e'}. \tag{69}$$

Proof: The proof of this theorem follows directly by identifying the dominant error tree belonging to the set $\mathcal{T}^d \setminus \mathcal{T}_{\mathcal{P}^*(P)}$. By further applying the result in Proposition 4 and Theorem 5, we obtain the result via the "worst-exponent-wins" [11, Ch. 1] principle by minimizing over all trees in the set of optimal projections $\mathcal{E}_{\mathcal{P}^*(P)}$ in (69). ∎

This theorem now allows us to analyze the more general error event $\mathcal{A}_n(\mathcal{P}^*(P))$, which includes $\mathcal{A}_n$ in (5) as a special case if the set of optimal tree projections $\mathcal{T}_{\mathcal{P}^*(P)}$ in (59) is a singleton.

VIII. Numerical Experiments

In this section, we perform a series of numerical experiments with the following three objectives:

1) In Section VIII-A, we study the accuracy of the Euclidean approximations (Theorem 8). We do this by analyzing under which regimes the approximate crossover rate $\widetilde{J}_{e,e'}$ in (55) is close to the true crossover rate $J_{e,e'}$ in (21).
2) Since the LDP and error exponent analysis are asymptotic theories, in Section VIII-B we use simulations to study the behavior of the actual crossover rate, given a finite number of samples n. In particular, we study how fast the crossover rate, obtained from simulations, converges to the true crossover rate. To do so, we generate a number of samples from the true distribution and use the Chow-Liu algorithm to learn trees structures. Then we compare the result to the true structure and finally compute the error probability.
3) In Section VIII-C, we address the issue of the learner not having access to the true distribution, but nonetheless wanting to compute an estimate of the crossover rate. The learner only has the samples $\mathbf{x}^n$ or equivalently, the empirical distribution $\widehat{P}$. However, in all the preceding analysis, to compute the true crossover rate $J_{e,e'}$ and the overall error exponent K_P, we used the true distribution P and solved the constrained optimization problem in (21). Alternatively we computed the approximation in (55), which is also a function of the true distribution. However, in practice, it is also useful to compute an online estimate of the crossover rate by using the empirical distribution in place of the true distribution in the constrained optimization problem in (21). This is an estimate of the rate that the learner can compute given the samples. We call this the *empirical rate* and formally define it in Section VIII-C.

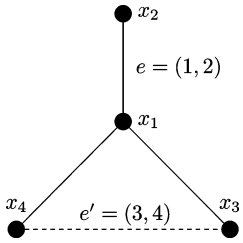

Fig. 9. Graphical model used for our numerical experiments. The true model is a symmetric star (cf. Section IV) in which the mutual information quantities satisfy $I(P_{1,2}) = I(P_{1,3}) = I(P_{1,4})$ and by construction, $I(P_{e'}) < I(P_{1,2})$ for any nonedge e'. Besides, the mutual information quantities on the nonedges are equal, for example, $I(P_{2,3}) = I(P_{3,4})$..

We perform convergence analysis of the empirical rate and also numerically verify the rate of convergence to the true crossover rate.

In the following, we will be performing numerical experiments for the undirected graphical model with four nodes as shown in Fig. 9. We parameterize the distribution with $d = 4$ variables with a single parameter $\gamma > 0$ and let $\mathcal{X} = \{0, 1\}$, i.e., all the variables are binary. For the parameters, we set $P_1(x_1 = 0) = 1/3$ and

$$P_{i|1}(x_i = 0|x_1 = 0) = \frac{1}{2} + \gamma, \quad i = 2, 3, 4 \tag{70a}$$

$$P_{i|1}(x_i = 0|x_1 = 1) = \frac{1}{2} - \gamma, \quad i = 2, 3, 4. \tag{70b}$$

With this parameterization, we see that if γ is small, the mutual information $I(P_{1,i})$ for $i = 2, 3, 4$ is also small. In fact if $\gamma = 0$, x_1 is independent of x_i for $i = 2, 3, 4$ and as a result, $I(P_{1,i}) = 0$. Conversely, if γ is large, the mutual information $I(P_{1,i})$ increases as the dependence of the outer nodes with the central node increases. Thus, we can vary the size of the mutual information along the edges by varying γ. By symmetry, there is only one crossover rate, and hence, this crossover rate is also the error exponent for the error event $\mathcal{A}_n$ in (5). This is exactly the same as the symmetric star graph as described in Section IV.

A. Accuracy of Euclidean Approximations

We first study the accuracy of the Euclidean approximations used to derive the result in Theorem 8. We denote the *true rate* as the crossover rate resulting from the nonconvex optimization problem (21) and the *approximate rate* as the crossover rate computed using the approximation in (55).

We vary γ from 0 to 0.2 and plot both the true and approximate rates against the difference between the mutual informations $I(P_e) - I(P_{e'})$ in Fig. 10, where e denotes any edge and e' denotes any nonedge in the model. The nonconvex optimization problem was performed using the Matlab function `fmincon` in the optimization toolbox. We used several different feasible starting points and chose the best optimal objective value to avoid problems with local minima. We first note from Fig. 10 that both rates increase as $I(P_e) - I(P_{e'})$ 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 very close. This is in line with the assumptions for Theorem 8. Recall

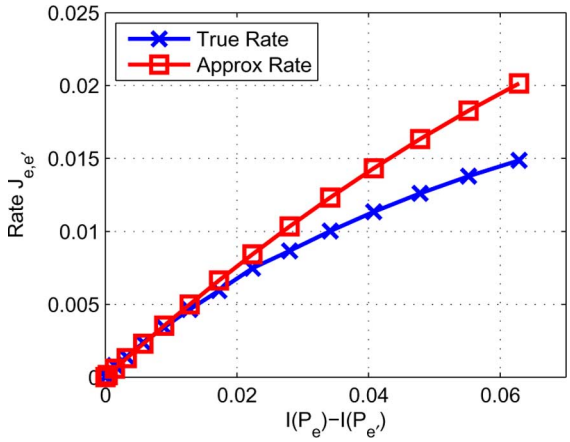

Fig. 10. Comparison of true and approximate rates.

that if $P_{e,e'}$ satisfies the ϵ-very noisy condition (for some small ϵ), then the mutual information quantities $I(P_e)$ and $I(P_{e'})$ are close and consequently the true and approximate crossover rates are also close. When the difference between the mutual informations increases, the true and approximate rate separate from each other.

B. Comparison of True Crossover Rate to the Rate Obtained From Simulations

In this section, we compare the true crossover rate in (21) to the rate we obtain when we learn tree structures using Chow-Liu with i.i.d. samples drawn from P, which we define as the *simulated rate*. We fixed $\gamma > 0$ in (70) then for each n, we estimated the probability of error using the Chow-Liu algorithm as described in Section III. We state the procedure precisely in the following steps.

1) Fix $n \in \mathbb{N}$ and sample n i.i.d. observations $\mathbf{x}^n$ from P.
2) Compute the empirical distribution $\widehat{P}$ and the set of empirical mutual information quantities $\left\{I(\widehat{P}_e) : e \in \binom{\mathcal{V}}{2}\right\}$.
3) Learn the Chow-Liu tree $\mathcal{E}_{\mathrm{ML}}$ using a MWST algorithm with $\left\{I(\widehat{P}_e) : e \in \binom{\mathcal{V}}{2}\right\}$ as the edge weights.
4) If $\mathcal{E}_{\mathrm{ML}}$ is not equal to $\mathcal{E}_P$, then we declare an error.
5) Repeat steps 1–4 a total of $M \in \mathbb{N}$ times and estimate the probability of error $\mathbb{P}(\mathcal{A}_n) = \#\mathrm{errors}/M$ and the error exponent $-(1/n)\log \mathbb{P}(\mathcal{A}_n)$, which is the simulated rate.

If the probability of error $\mathbb{P}(\mathcal{A}_n)$ is very small, then the number of runs M to estimate $\mathbb{P}(\mathcal{A}_n)$ has to be fairly large. This is often the case in error exponent analysis as the sample size needs to be substantial to estimate very small error probabilities.

In Fig. 11, we plot the true rate, the approximate rate and the simulated rate when $\gamma = 0.01$ (and $M = 10^7$) and $\gamma = 0.2$ (and $M = 5 \times 10^8$). Note that, in the former case, the true rate is higher than the approximate rate and in the latter case, the reverse is true. When γ is large ($\gamma = 0.2$), there are large differences in the true tree models. Thus, we expect that the error probabilities to be very small, and hence, M has to be large in order to estimate the error probability correctly but n does not have to be too large for the simulated rate to converge to the true rate. On the other hand, when γ is small ($\gamma = 0.01$), there are only subtle differences in the graphical models; hence, we need

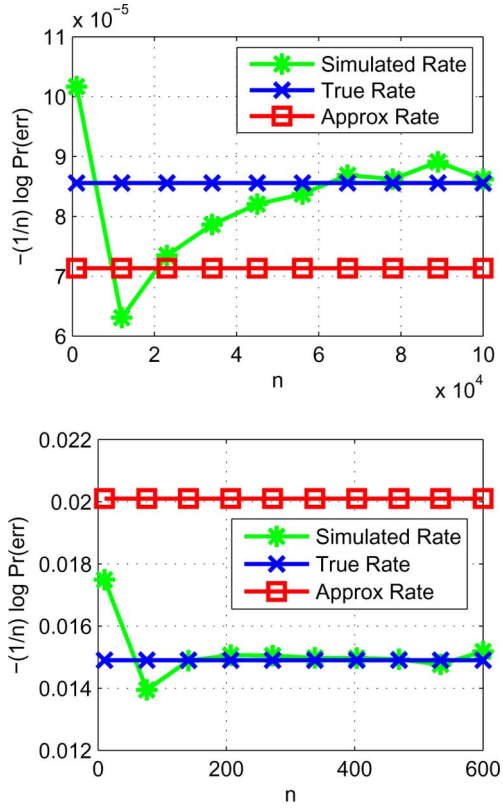

Fig. 11. Comparison of true, approximate and simulated rates with $\gamma = 0.01$ (top) and $\gamma = 0.2$ (bottom). Here the number of runs $M = 10^7$ for $\gamma = 0.01$ and $M = 5 \times 10^8$ for $\gamma = 0.2$. The probability of error is computed dividing the total number of errors by the total number of runs.

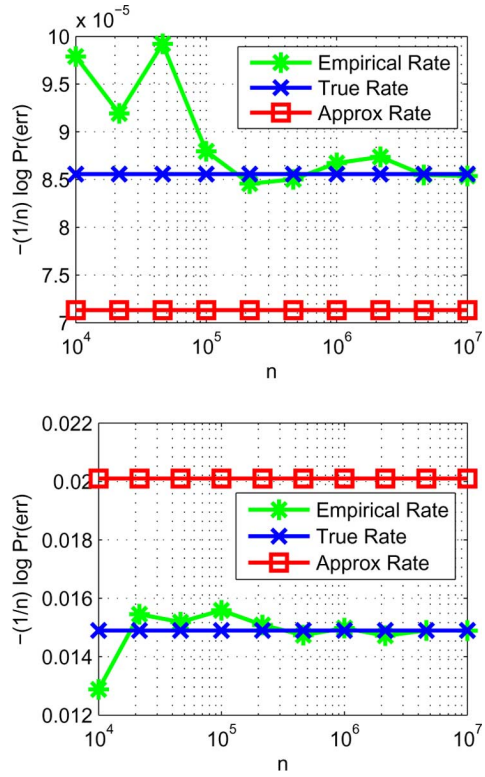

Fig. 12. Comparison of True, Approximate and Empirical Rates with $\gamma = 0.01$ (top) and $\gamma = 0.2$ (bottom). Here n is the number of observations used to estimate the empirical distribution.

a larger number of samples n for the simulated rate to converge to its true value, but M does not have to be large since the error probabilities are not small. The above observations are in line with our intuition.

C. Comparison of True Crossover Rate to Rate Obtained From the Empirical Distribution

In this subsection, we compare the true rate to the *empirical rate*, which is defined as

$$\widehat{J}_{e,e'} := \inf_{Q \in \mathcal{P}(\mathcal{X}^4)} \left\{ D(Q \| \widehat{P}_{e,e'}) : I(Q_{e'}) = I(Q_e) \right\}. \quad (71)$$

The empirical rate $\widehat{J}_{e,e'} = \widehat{J}_{e,e'}(\widehat{P}_{e,e'})$ is a function of the empirical distribution $\widehat{P}_{e,e'}$. This rate is computable by a learner, who does not have access to the true distribution P. The learner only has access to a finite number of samples $\mathbf{x}^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Given $\mathbf{x}^n$, the learner can compute the empirical probability $\widehat{P}_{e,e'}$ and perform the optimization in (71). This is an estimate of the true crossover rate. A natural question to ask is the following: Does the empirical rate $\widehat{J}_{e,e'}$ converge to the true crossover rate $J_{e,e'}$ as $n \to \infty$? The next theorem answers this question in the affirmative.

Theorem 12 (Crossover Rate Consistency): The empirical crossover rate $\widehat{J}_{e,e'}$ in (71) converges almost surely to the true crossover rate $J_{e,e'}$ in (21), i.e.,

$$\mathbb{P}\left(\lim_{n \to \infty} \widehat{J}_{e,e'} = J_{e,e'} \right) = 1. \quad (72)$$

Proof: (*Sketch*) The proof of this theorem follows from the continuity of $\widehat{J}_{e,e'}$ in the empirical distribution $\widehat{P}_{e,e'}$ and the continuous mapping theorem by Mann and Wald [49]. See Appendix F for the details. ■

We conclude that the learning of the rate from samples is strongly consistent. Now we perform simulations to determine how many samples are required for the empirical rate to converge to the true rate.

We set $\gamma = 0.01$ and $\gamma = 0.2$ in (70). We then drew n i.i.d. samples from P and computed the empirical distribution $\widehat{P}_{e,e'}$. Next, we solved the optimization problem in (71) using the `fmincon` function in Matlab, using different initializations and compared the empirical rate to the true rate. We repeated this for several values of n and the results are displayed in Fig. 12. We see that for $\gamma = 0.01$, approximately $n = 8 \times 10^6$ samples are required for the empirical distribution to be close enough to the true distribution so that the empirical rate converges to the true rate.

IX. Conclusion, Extensions and Open Problems

In this paper, we presented a solution to the problem of finding the error exponent for tree structure learning by extensively using tools from large-deviations theory combined with facts about tree graphs. We quantified the error exponent for learning the structure and exploited the structure of the true tree to identify the dominant tree in the set of erroneous trees. We also drew insights from the approximate crossover rate, which can be interpreted as the SNR for learning. These two

main results in Theorems 5 and 8 provide the intuition as to how errors occur for learning discrete tree distributions via the Chow-Liu algorithm.

In a companion paper [27], we develop counterparts to the results here for the Gaussian case. Many of the results carry through but thanks to the special structure that Gaussian distributions possess, we are also able to identify which structures (among the class of trees) are easier to learn and which are harder to learn given a fixed set of correlation coefficients on the edges. Using Euclidean information theory, we show that if the parameters on the edges are fixed, the star is the most difficult to learn (requiring many more samples to ensure $\mathbb{P}(\mathcal{A}_n) \leq \delta$) while the Markov chain is the easiest. The results in this paper have also been extended to learning high-dimensional general acyclic models (forests) [38], where d grows with n and typically the growth of d is much faster than that of n.

There are many open problems resulting from this paper. One of these involves studying the optimality of the error exponent associated to the ML Chow-Liu algorithm K_P, i.e., whether the rate established in Theorem 5 is the best (largest) among all consistent estimators of the edge set. Also, since large-deviation rates may not be indicative of the true error probability $\mathbb{P}(\mathcal{A}_n)$, results from weak convergence theory [50] may potentially be applicable to provide better approximations to $\mathbb{P}(\mathcal{A}_n)$.

Appendix A
Proof of Theorem 2

Proof: We divide the proof of this theorem into three steps. Steps 1 and 2 prove the expression in (21). Step 3 proves the existence of the optimizer.

Step 1: First, we note from Sanov's Theorem [30, Ch. 11] that the empirical joint distribution on edges e and e' satisfies

$$\lim_{n\to\infty} -\frac{1}{n}\log \mathbb{P}(\widehat{P}_{e,e'} \in \mathcal{F}) = \inf\{D(Q\|P_{e,e'}) : Q \in \mathcal{F}\} \tag{73}$$

for any set $\mathcal{F} \subset \mathcal{P}(\mathcal{X}^4)$ that equals the closure of its interior, i.e., $\mathcal{F} = \mathrm{cl}(\mathrm{int}(\mathcal{F}))$. We now have a LDP for the sequence of probability measures $\widehat{P}_{e,e'}$, the empirical distribution on (e, e'). Assuming that e and e' do not share a common node, $\widehat{P}_{e,e'} \in \mathcal{P}(\mathcal{X}^4)$ is a probability distribution over four variables (the variables in the node pairs e and e'). We now define the function $h : \mathcal{P}(\mathcal{X}^4) \to \mathbb{R}$ as

$$h(Q) := I(Q_{e'}) - I(Q_e). \tag{74}$$

Since $Q_e = \sum_{x_{e'}} Q$, defined in (22) is continuous in Q and the mutual information $I(Q_e)$ is also continuous in Q_e, we conclude that h is indeed continuous, since it is the composition of continuous functions. By applying the contraction principle [11] to the sequence of probability measures $\widehat{P}_{e,e'}$ and the continuous map h, we obtain a corresponding LDP for the new sequence of probability measures $h(\widehat{P}_{e,e'}) = I(\widehat{P}_{e'}) - I(\widehat{P}_e)$, where the rate is given by

$$J_{e,e'} = \inf_{Q\in\mathcal{P}(\mathcal{X}^4)}\{D(Q\|P_{e,e'}) : h(Q) \geq 0\} \tag{75}$$

$$= \inf_{Q\in\mathcal{P}(\mathcal{X}^4)}\{D(Q\|P_{e,e'}) : I(Q_{e'}) \geq I(Q_e)\}. \tag{76}$$

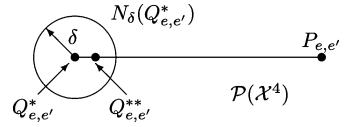

Fig. 13. Illustration of Step 2 of the proof of Theorem 2.

We now claim that the limit in (20) exists. From Sanov's theorem [30, Ch. 11], it suffices to show that the constraint set $\mathcal{F} := \{I(Q_{e'}) \geq I(Q_e)\}$ in (76) is a regular closed set, i.e., it satisfies $\mathcal{F} = \mathrm{cl}(\mathrm{int}(\mathcal{F}))$. This is true because there are no isolated points in $\mathcal{F}$, and, thus, the interior is nonempty. Hence, there exists a sequence of distributions $\{Q_n\}_{n=1}^{\infty} \subset \mathrm{int}(\mathcal{F})$ such that $\lim_{n\to\infty} D(Q_n\|P_{e,e'}) = D(Q^*\|P_{e,e'})$, which proves the existence of the limit in (20).

Step 2: We now show that the optimal solution $Q^*_{e,e'}$, if it exists (as will be shown in Step 3), must satisfy $I(Q^*_e) = I(Q^*_{e'})$. Suppose, to the contrary, that $Q^*_{e,e'}$ with objective value $D\left(Q^*_{e,e'}\|P_{e,e'}\right)$ is such that $I(Q^*_{e'}) > I(Q^*_e)$. Then $h\left(Q^*_{e,e'}\right) > 0$, where h, as shown above, is continuous. Thus, there exists a $\delta > 0$ such that the δ-neighborhood

$$N_\delta\left(Q^*_{e,e'}\right) := \left\{R : \left\|R - Q^*_{e,e'}\right\|_\infty < \delta\right\} \tag{77}$$

satisfies $h\left(N_\delta\left(Q^*_{e,e'}\right)\right) \subset (0,\infty)$ [37, Ch. 2]. Consider the new distribution (See Fig. 13)

$$Q^{**}_{e,e'} = Q^*_{e,e'} + \frac{\delta}{2}\left(P_{e,e'} - Q^*_{e,e'}\right) \tag{78}$$

$$= \left(1 - \frac{\delta}{2}\right) Q^*_{e,e'} + \frac{\delta}{2} P_{e,e'}. \tag{79}$$

Note that $Q^{**}_{e,e'}$ belongs to $N_\delta(Q^*_{e,e'})$, and hence, is a feasible solution of (76). We now prove that $D\left(Q^{**}_{e,e'}\|P_{e,e'}\right) < D\left(Q^*_{e,e'}\|P_{e,e'}\right)$, which contradicts the optimality of $Q^*_{e,e'}$.

$$D(Q^{**}_{e,e'}\|P_{e,e'})$$

$$= D\left(\left(1 - \frac{\delta}{2}\right) Q^*_{e,e'} + \frac{\delta}{2} P_{e,e'}\middle\|P_{e,e'}\right) \tag{80}$$

$$\leq \left(1 - \frac{\delta}{2}\right) D\left(Q^*_{e,e'}\|P_{e,e'}\right) + \frac{\delta}{2} D\left(P_{e,e'}\|P_{e,e'}\right) \tag{81}$$

$$= \left(1 - \frac{\delta}{2}\right) D(Q^*_{e,e'}\|P_{e,e'}) \tag{82}$$

$$< D\left(Q^*_{e,e'}\|P_{e,e'}\right) \tag{83}$$

where (81) is due to the convexity of the KL-divergence in the first variable [30, Ch. 2], (82) is because $D(P_{e,e'}\|P_{e,e'}) = 0$ and (83) is because $\delta > 0$. Thus, we conclude that the optimal solution must satisfy $I(Q^*_e) = I(Q^*_{e'})$ and the crossover rate can be stated as (21).

Step 3: Now, we prove the existence of the minimizer $Q^*_{e,e'}$, which will allow us to replace the $\inf$ in (21) with min. First, we note that $D(Q\|P_{e,e'})$ is continuous in both variables, and hence continuous and the first variable Q. It remains to show that the constraint set

$$\Lambda := \left\{Q \in \mathcal{P}(\mathcal{X}^4) : I(Q_{e'}) = I(Q_e)\right\} \tag{84}$$

is compact, since it is clearly nonempty (the uniform distribution belongs to Λ). Then we can conclude, by Weierstrass' extreme value theorem [37, Theorem 4.16], that the minimizer $Q^*_{e,e'} \in \Lambda$ exists. By the Heine-Borel theorem [37, Theorem 2.41], it suffices to show that Λ is bounded and closed. Clearly Λ is bounded since $\mathcal{P}(\mathcal{X}^4)$ is a bounded set. Now, $\Lambda = h^{-1}(\{0\})$ where h is defined in (74). Since h is continuous and $\{0\}$ is closed, Λ is closed [37, Theorem 4.8]. Hence, that Λ is compact. We also need to use the fact that Λ is compact in the proof of Theorem 12. ∎

Appendix B
Proof of Theorem 5

Proof: We first claim that $\mathcal{E}_P^*$, the edge set corresponding to the dominant error tree, differs from $\mathcal{E}_P$ by exactly one edge.[21] To prove this claim, assume, to the contrary, that $\mathcal{E}_P^*$ differs from $\mathcal{E}_P$ by two edges. Let $\mathcal{E}_{\mathrm{ML}} = \mathcal{E}' := \mathcal{E}_P \setminus \{e_1, e_2\} \cup \{e_1', e_2'\}$, where e_1', $e_2' \notin \mathcal{E}_P$ are the two edges that have replaced e_1, $e_2 \in \mathcal{E}_P$ respectively. Since $T' = (\mathcal{V}, \mathcal{E}')$ is a tree, these edges cannot be arbitrary and specifically, $\{e_1, e_2\} \in \{\mathrm{Path}(e_1'; \mathcal{E}_P) \cup \mathrm{Path}(e_2'; \mathcal{E}_P)\}$ for the tree constraint to be satisfied. Recall that the rate of the event that the output of the ML algorithm is T' is given by $\Upsilon(T')$ in (30).

Suppose that $e_i \in \mathrm{Path}(e_i'; \mathcal{E}_P)$ for $i = 1, 2$ and $e_i \notin \mathrm{Path}(e_j'; \mathcal{E}_P)$ for $i, j = 1, 2$ and $i \neq j$. See Fig. 14. Note that the true mutual information quantities satisfy $I(P_{e_i}) > I(P_{e_i'})$. We prove this claim by contradiction that suppose $I(P_{e_i'}) \geq I(P_{e_i})$ then, $\mathcal{E}_P$ does not have maximum weight because if the nonedge e_i' replaces the true edge e_i, the resulting tree[22] would have higher weight, contradicting the optimality of the true edge set $\mathcal{E}_P$, which is the MWST with the true mutual information quantities as edge weights. More precisely, we can compute the exponent when T' is the output of the MWST algorithm

$$\Upsilon(T') = \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(\bigcap_{i=1,2} \{I(\widehat{P}_{e_i'}) \geq I(\widehat{P}_{e_i})\} \right) \tag{85}$$

$$\geq \max_{i=1,2} \lim_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(\{I(\widehat{P}_{e_i'}) \geq I(\widehat{P}_{e_i})\} \right) \tag{86}$$

$$= \max\left\{ J_{e_1,e_1'}, J_{e_2,e_2'} \right\}. \tag{87}$$

Now $J_{e_i,e_i'} = \Upsilon(T_i)$ where $T_i := (\mathcal{V}, \mathcal{E}_P \setminus \{e_i\} \cup \{e_i'\})$. From Prop. 4, the error exponent associated to the dominant error tree, i.e., $K_P = \min_{T \neq T_P} \Upsilon(T)$ and from (87), the dominant error tree cannot be T' and should differ from T_P by one and only one edge.

The similar conclusion holds for the two other cases (i) $e_i \in \mathrm{Path}(e_i'; \mathcal{E}_P)$ for $i = 1, 2$, $e_2 \in \mathrm{Path}(e_1'; \mathcal{E}_P)$ and $e_1 \notin \mathrm{Path}(e_2'; \mathcal{E}_P)$ and (ii) $e_i \in \mathrm{Path}(e_i'; \mathcal{E}_P)$ for $i = 1, 2$, $e_1 \in \mathrm{Path}(e_2'; \mathcal{E}_P)$ and $e_2 \notin \mathrm{Path}(e_1'; \mathcal{E}_P)$. In other words, the dominant error tree differs from the true tree by one edge.

We now use the "worst-exponent-wins principle" [11, Ch. 1], to conclude that the rate that dominates is the minimum $J_{r(e'),e'}$ over all possible $e' \notin \mathcal{E}_P$, namely $J_{r(e^*),e^*}$ with e^* defined in (37). More precisely

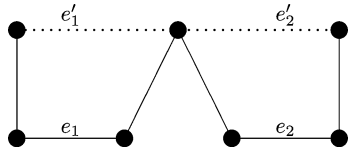

Fig. 14. Illustration of the proof of Theorem 5. The dominant error event involves only one crossover event.

$$\mathbb{P}(\mathcal{A}_n)$$

$$= \mathbb{P}\left(\bigcup_{e' \notin \mathcal{E}_P} \{e' \text{ replaces any } e \in \mathrm{Path}\,(e'; \mathcal{E}_P) \text{ in } T_{\mathrm{ML}}\} \right)$$

$$= \mathbb{P}\left(\bigcup_{e' \notin \mathcal{E}_P} \bigcup_{e \in \mathrm{Path}(e'; \mathcal{E}_P)} \{e' \text{replaces } e \text{ in } T_{\mathrm{ML}}\} \right) \tag{88}$$

$$\leq \sum_{e' \notin \mathcal{E}_P} \sum_{e \in \mathrm{Path}(e'; \mathcal{E}_P)} \mathbb{P}\left(\{e' \text{replaces } e \text{ in } T_{\mathrm{ML}}\} \right) \tag{89}$$

$$= \sum_{e' \notin \mathcal{E}_P} \sum_{e \in \mathrm{Path}(e'; \mathcal{E}_P)} \mathbb{P}(\{I(\widehat{P}_{e'}) \geq I(\widehat{P}_e)\}) \tag{90}$$

$$\doteq \sum_{e' \notin \mathcal{E}_P} \sum_{e \in \mathrm{Path}(e'; \mathcal{E}_P)} \exp(-n J_{e,e'}) \tag{91}$$

$$\doteq \exp\left(-n \min_{e' \notin \mathcal{E}_P} \min_{e \in \mathrm{Path}(e'; \mathcal{E}_P)} J_{e,e'} \right) \tag{92}$$

where (89) is from the union of events bound, (90) and (91) are from the definitions of the crossover event and rate respectively (as described in Cases 1 and 2 above) and (92) is an application of the "worst-exponent-wins" principle [11, Ch. 1].

We conclude[23] from (92) that

$$\mathbb{P}(\mathcal{A}_n) \dot{\leq} \exp(-n J_{r(e^*),e^*}) \tag{93}$$

from the definition of the dominant replacement edge $r(e')$ and the dominant nonedge e^*, defined in (35) and (37) respectively. The lower bound follows trivially from the fact that if $e^* \notin \mathcal{E}_P$ replaces $r(e^*)$, then the error $\mathcal{A}_n$ occurs. Thus, $\{e^* \text{replaces } r(e^*)\} \subset \mathcal{A}_n$ and

$$\mathbb{P}(\mathcal{A}_n) \dot{\geq} \mathbb{P}\left(\{e^* \text{ replaces } r(e^*) \text{ in } T_{\mathrm{ML}}\} \right) \tag{94}$$

$$\doteq \exp(-n J_{r(e^*),e^*}). \tag{95}$$

Hence, (93) and (95) imply that $\mathbb{P}(\mathcal{A}_n) \doteq \exp(-n J_{r(e^*),e^*})$, which proves our main result in (36).

The finite-sample result in (39) comes from the upper bound in (92) and the following two facts:

1) The exact number of n-types with alphabet $\mathcal{Y}$ is given by $\binom{n+1+|\mathcal{Y}|}{n+1}$ [51]. In particular, we have

$$\mathbb{P}(\mathcal{C}_{e,e'}) \leq \binom{n+1+|\mathcal{X}|^4}{n+1} \exp(-n J_{e,e'}) \tag{96}$$

for all $n \in \mathbb{N}$, since $\mathcal{C}_{e,e'}$ only involves the distribution $P_{e,e'} \in \mathcal{P}(\mathcal{X}^4)$. Note that the exponent 4 of $|\mathcal{X}|^4$ in (96)

[21] This is somewhat analogous to the fact that the second-best MWST differs from the MWST by exactly one edge [34].

[22] The resulting graph is indeed a tree because $\{e_i'\} \cup \mathrm{Path}(e_i'; \mathcal{E}_P)$ form a cycle so if any edge is removed, the resulting structure does not have any cycles and is connected; hence, it is a tree. See Fig. 2.

[23] The notation $a_n \dot{\leq} b_n$ means that $\limsup_{n\to\infty} \frac{1}{n} \log(a_n/b_n) \leq 0$. Similarly, $a_n \dot{\geq} b_n$ means that $\liminf_{n\to\infty} \frac{1}{n} \log(a_n/b_n) \geq 0$.

is an upper bound since if e and e' share a node $P_{e,e'} \in \mathcal{P}(\mathcal{X}^3)$.

2) The number of error events $\mathcal{C}_{e,e'}$ is at most $(d-1)^2(d-2)/2$ because there are $\binom{d}{2}-(d-1) = (d-1)(d-2)/2$ nonedges and for each nonedge, there are at most $d-1$ edges along its path.

This completes the proof. ∎

Appendix C
Proof of Theorem 6

Statement (a) $\Leftrightarrow$ statement (b) was proven in full after the theorem was stated. Here we provide the proof that (b) $\Leftrightarrow$ (c). Recall that statement (c) says that T_P is not a proper forest. We first begin with a preliminary lemma.

Lemma 13: Suppose x, y, z are three random variables taking on values on finite sets $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ respectively. Assume that $P(x,y,z) > 0$ everywhere. Then $x - y - z$ and $x - z - y$ are Markov chains if and only if x is jointly independent of y, z.

Proof: ($\Rightarrow$) That $x - y - z$ is a Markov chain implies that

$$P(z|y,x) = P(z|y) \tag{97}$$

or alternatively

$$P(x,y,z) = P(x,y)\frac{P(y,z)}{P(y)}. \tag{98}$$

Similarly from the fact that $x - z - y$ is a Markov chain, we have

$$P(x,y,z) = P(x,z)\frac{P(y,z)}{P(z)}. \tag{99}$$

Equating (98) and (99), and use the positivity to cancel $P(y,z)$, we arrive at

$$P(x|y) = P(x|z). \tag{100}$$

It follows that $P(x|y)$ does not depend on y, so there is some constant $C(x)$ such that $P(x|y) = C(x)$ for all $y \in \mathcal{Y}$. This immediately implies that $C(x) = P(x)$ so that $P(x|y) = P(x)$. A similar argument gives that $P(x|z) = P(x)$. Furthermore, if $x - y - z$ is a Markov chain, so is $z - y - x$, therefore

$$P(x|y,z) = P(x|y) = P(x). \tag{101}$$

The above equation says that x is jointly independent of both y and z.

($\Leftarrow$) Conversely, if x is jointly independent of both y and z, then $x - y - z$ and $x - z - y$ are Markov chains. In fact x is not connected to $y - z$. ∎

Proof: We now prove (b) $\Longleftrightarrow$ (c) using Lemma 13 and the assumption that $P(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}^d$.

($\Rightarrow$) If (b) is true then $I(P_{e'}) < I(P_e)$ for all $e \in \mathrm{Path}(e'; \mathcal{E}_P)$ and for all $e' \notin \mathcal{E}_P$. Assume, to the contrary, that T_P is a proper forest, i.e., it contains at least 2 connected components (each connected component may only have one node), say $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)$ for $i = 1, 2$. Without loss of generality, let x_1 be in component $\mathcal{G}_1$ and x_2, x_3 belong to component $\mathcal{G}_2$. Then since $\mathcal{V}_1 \cap \mathcal{V}_2 = \emptyset$ and $\mathcal{V}_1 \cup \mathcal{V}_2 = \mathcal{V}$, we have that x_1 jointly independent of x_2 and x_3. By Lemma 13, we have the following Markov chains $x_1 - x_2 - x_3$ and $x_1 - x_3 - x_2$. This implies from the Data Processing Inequality [30, Theorem 2.8.1] that $I(P_{1,2}) \geq I(P_{1,3})$ and at the same time $I(P_{1,2}) \leq I(P_{1,3})$ which means that $I(P_{1,2}) = I(P_{1,3})$. This contradicts (b) since by taking $e' = (1,2)$, the mutual informations along the path $\mathrm{Path}(e'; \mathcal{E}_P)$ are no longer distinct.

($\Leftarrow$) Now assume that (c) is true, i.e., T_P is not a proper forest. Suppose, to the contrary, (b) is not true, i.e., there exists a $e' \notin \mathcal{E}_P$ such that $I(P_{e'}) = I(P_{r(e')})$, where $r(e')$ is the replacement edge associated with the nonedge e'. Without loss of generality, let $e' = (1,2)$ and $r(e') = (3,4)$, then since T_P is not a proper forest, we have the following Markov chain $x_1 - x_3 - x_4 - x_2$. Now note that $I(P_{1,2}) = I(P_{3,4})$. In fact, because there is no loss of mutual information $I(P_{1,4}) = I(P_{3,4})$, and hence, by the Data Processing Inequality, we also have $x_3 - x_1 - x_4 - x_2$. By using Lemma 13, we have x_4 jointly independent of x_1 and x_3; hence, we have a proper forest, which is a contradiction. ∎

Appendix D
Proof of Theorem 8

Proof: The proof proceeds in several steps. See Figs. 5 and 6 for intuition behind this proof.

Step 1: Let Q be such that

$$Q(x_i, x_j, x_k, x_l) = P_{e,e'}(x_i, x_j, x_k, x_l) + \epsilon_{i,j,k,l}. \tag{102}$$

Thus, the $\epsilon_{i,j,k,l}$'s are the deviations of Q from $P_{e,e'}$. To ensure that Q is a valid distribution we require $\sum \epsilon_{i,j,k,l} = 0$. The objective in (52) can now be alternatively expressed as

$$\frac{1}{2}\boldsymbol{\epsilon}^T \mathbf{K}_{e,e'} \boldsymbol{\epsilon} = \frac{1}{2} \sum_{x_i,x_j,x_k,x_l} \frac{\epsilon_{i,j,k,l}^2}{P_{e,e'}(x_i,x_j,x_k,x_l)} \tag{103}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^{\|\mathcal{X}\|^4}$ is the vectorized version of the deviations $\epsilon_{i,j,k,l}$ and $\mathbf{K}_{e,e'}$ is a $|\mathcal{X}|^4 \times |\mathcal{X}|^4$ diagonal matrix containing the entries $1/P_{e,e'}(x_i,x_j,x_k,x_l)$ along its diagonal.

Step 2: We now perform a first-order Taylor expansion of $I(Q_e)$ in the neighborhood of $I(P_e)$.

$$I(Q_e) = I(P_e) + \boldsymbol{\epsilon}^T \nabla_{\boldsymbol{\epsilon}} I(Q_e)\Big|_{\boldsymbol{\epsilon}=0} + o(\|\boldsymbol{\epsilon}\|), \tag{104}$$

$$= I(P_e) + \boldsymbol{\epsilon}^T \mathbf{s}_e + o(\|\boldsymbol{\epsilon}\|), \tag{105}$$

where $\mathbf{s}_e$ is the length $|\mathcal{X}|^4$-vector that contains the information density values of edge e. Note that because of the assumption that P is not a proper forest, $P_{i,j} \neq P_i P_j$ for all (i,j); hence, the linear term does not vanish.[24] The constraints can now be rewritten as

$$\boldsymbol{\epsilon}^T \mathbf{1} = 0, \qquad \boldsymbol{\epsilon}^T(\mathbf{s}_{e'} - \mathbf{s}_e) = I(P_e) - I(P_{e'}). \tag{106}$$

or in matrix notation as

$$\begin{bmatrix} \mathbf{s}_{e'}^T - \mathbf{s}_e^T \\ 1^T \end{bmatrix} \boldsymbol{\epsilon} = \begin{bmatrix} I(P_e) - I(P_{e'}) \\ 0 \end{bmatrix} \tag{107}$$

[24]Indeed if P_e were a product distribution, the linear term in (105) vanishes and $I(Q_e)$ is approximately a quadratic in $\boldsymbol{\epsilon}$ (as shown in [15]).

where $\mathbf{1}$ is the length-$|\mathcal{X}|^4$ vector consisting of all ones. For convenience, we define $\mathbf{L}_{e,e'}$ to be the matrix in (107) i.e.,

$$\mathbf{L}_{e,e'} := \begin{bmatrix} \mathbf{s}_{e'}^T - \mathbf{s}_e^T \\ \mathbf{1}^T \end{bmatrix} \in \mathbb{R}^{2\times|\mathcal{X}|^4}. \tag{108}$$

Step 3: The optimization problem now reduces to minimizing (103) subject to the constraints in (107). This is a standard least-squares problem. By using the Projection Theorem in Hilbert spaces (or Lagrange multipliers), we get the solution

$$\boldsymbol{\epsilon}^* = \mathbf{K}_{e,e'}^{-1}\mathbf{L}_{e,e'}^T(\mathbf{L}_{e,e'}\mathbf{K}_{e,e'}^{-1}\mathbf{L}_{e,e'}^T)^{-1}\begin{bmatrix} I(P_e) - I(P_{e'}) \\ 0 \end{bmatrix}. \tag{109}$$

The inverse of $\mathbf{L}_{e,e'}\mathbf{K}_{e,e'}^{-1}\mathbf{L}_{e,e'}^T$ exists because we assumed T_P is not a proper forest, and hence, $P_{i,j} \neq P_iP_j$ for all $(i,j) \in \binom{\mathcal{V}}{2}$. This is a sufficient condition for the matrix $\mathbf{L}_{e,e'}$ to have full row rank and thus, $\mathbf{L}_{e,e'}\mathbf{K}_{e,e'}^{-1}\mathbf{L}_{e,e'}^T$ is invertible. Finally, we substitute $\boldsymbol{\epsilon}^*$ in (109) into (103) to obtain

$$\widetilde{J}_{e,e'} = \frac{1}{2}\left[\left(\mathbf{L}_{e,e'}\mathbf{K}_{e,e'}^{-1}\mathbf{L}_{e,e'}^T\right)^{-1}\right]_{11}(I(P_e) - I(P_{e'}))^2 \tag{110}$$

where $[\mathbf{M}]_{11}$ is the (1,1) element of the matrix $\mathbf{M}$. Define ψ to be the weighting function given by

$$\psi(P_{e,e'}) := \left[\left(\mathbf{L}_{e,e'}\mathbf{K}_{e,e'}^{-1}\mathbf{L}_{e,e'}^T\right)^{-1}\right]_{11}. \tag{111}$$

It now suffices to show that $\psi(P_{e,e'})$ is indeed the inverse variance of $s_e - s_{e'}$. We now simplify the expression for the weighting function $\psi(P_{e,e'})$ recalling how $\mathbf{L}_{e,e'}$ and $\mathbf{K}_{e,e'}$ are defined. The product of the matrices in (111) is

$$\mathbf{L}_{e,e'}\mathbf{K}_{e,e'}^{-1}\mathbf{L}_{e,e'}^T = \begin{bmatrix} \mathbb{E}[(s_{e'} - s_e)^2] & \mathbb{E}[s_{e'} - s_e] \\ \mathbb{E}[s_{e'} - s_e] & 1 \end{bmatrix} \tag{112}$$

where all expectations are with respect to the distribution $P_{e,e'}$. Note that the determinant of (112) is $\mathbb{E}[(s_{e'} - s_e)^2] - \mathbb{E}[(s_{e'} - s_e)]^2 = \mathrm{Var}(s_{e'} - s_e)$. Hence, the (1,1) element of the inverse of (112) is simply

$$\psi(P_{e,e'}) = \mathrm{Var}(s_{e'} - s_e)^{-1}. \tag{113}$$

Now, if e and e' share a node, this proof proceeds in exactly the same way. In particular, the crucial step (105) will also remain the same since the Taylor expansion does not change. This concludes the first part of the proof.

Step 4: We now prove the continuity statement. The idea is that all the approximations become increasingly exact as ϵ (in the definition of the ϵ-very noisy condition) tends to zero. More concretely, for every $\delta > 0$, there exists a $\epsilon_1 > 0$ such that if $P_{e,e'}$ satisfies the ϵ_1-very noisy condition, then

$$|I(P_e) - I(P_{e'})| < \delta \tag{114}$$

since mutual information is continuous. For every $\delta > 0$, there exists a $\epsilon_2 > 0$ such that if $P_{e,e'}$ satisfies the ϵ_2-very noisy condition, then

$$\|Q_{e,e'}^* - P_{e,e'}\|_\infty < \delta \tag{115}$$

since if $P_{e,e'}$ is ϵ_2-very noisy it is close to the constraint set $\{Q : I(Q_{e'}) \geq I(Q_e)\}$, and hence, close to the optimal solution $Q_{e,e'}^*$. For every $\delta > 0$, there exists a $\epsilon_3 > 0$ such that if $P_{e,e'}$ satisfies the ϵ_3-very noisy condition, then

$$\left|D(Q_{e,e'}^*\|\|P_{e,e'}) - \frac{1}{2}\|Q_{e,e'}^* - P_{e,e'}\|_{P_{e,e'}}^2\right| < \delta \tag{116}$$

which follows from the approximation of the divergence and the continuity statement in (115). For every $\delta > 0$, there exists a $\epsilon_4 > 0$ such that if $P_{e,e'}$ satisfies the ϵ_4-very noisy condition, then

$$\left|I(P_e) - \mathbf{s}_e^T(Q_{e,e'}^* - P_{e,e'})\right| < \delta \tag{117}$$

which follows from retaining only the first term in the Taylor expansion of the mutual information in (105). Finally, for every $\delta > 0$, there exists a $\epsilon_5 > 0$ such that if $P_{e,e'}$ satisfies the ϵ_5-very noisy condition, then

$$\|\widetilde{J}_{e,e'} - J_{e,e'}\| < \delta \tag{118}$$

which follows from continuity of the objective in the constraints (117). Now choose $\epsilon = \min_{i=1,\ldots,5}\epsilon_i$ to conclude that for every $\delta > 0$, there exists a $\epsilon > 0$ such that if $P_{e,e'}$ satisfies the ϵ-very noisy condition, then (118) holds. This completes the proof. ■

Appendix E
Proof of Proposition 10

Proof: The following facts about P in Table I can be readily verified: P

1) P is positive everywhere, i.e., $P(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}^3$.
2) P is Markov on the complete graph with $d = 3$ nodes; hence, P is not a tree-structured distribution.
3) The mutual information between x_1 and x_2 as a function of κ is given by

$$I(P_{1,2}) = \log 2 + (1-2\kappa)\log(1-2\kappa) + 2\kappa\log(2\kappa).$$

Thus, $I(P_{1,2}) \to \log 2 = 0.693$ as $\kappa \to 0$.
4) For any $(\xi,\kappa) \in (0,1/3)\times(0,1/2)$, $I(P_{2,3}) = I(P_{1,3})$ and this pair of mutual information quantities can be made arbitrarily small as $\kappa \to 0$.

Thus, for sufficiently small $\kappa > 0$, $I(P_{1,2}) > I(P_{2,3}) = I(P_{1,3})$. We conclude that the Chow-Liu MWST algorithm will first pick the edge (1, 2) and then arbitrarily choose between the two remaining edges: (2, 3) or (1, 3). Thus, optimal tree structure is not unique. ■

Appendix F
Proof of Theorem 12

We first state two preliminary lemmas and prove the first one. Theorem 12 will then be an immediate consequence of these lemmas.

Lemma 14: Let X and Y be two metric spaces and let $\mathcal{K} \subset X$ be a compact set in X. Let $f : X \times Y \to \mathbb{R}$ be a

continuous real-valued function. Then the function $g : Y \to \mathbb{R}$, defined as

$$g(y) := \min_{x \in \mathcal{K}} f(x, y), \quad \forall y \in Y \tag{119}$$

is continuous on Y.

Proof: Set the minimizer in (119) to be

$$x(y) := \arg\min_{x \in \mathcal{K}} f(x, y). \tag{120}$$

The optimizer $x(y) \in \mathcal{K}$ exists since $f(x, y)$ is continuous on $\mathcal{K}$ for each $y \in Y$ and $\mathcal{K}$ is compact. This follows from Weierstrass' extreme value theorem [37, Theorem 4.16]. We want to show that for $\lim_{y' \to y} g(y') = g(y)$. In other words, we need to prove that

$$\lim_{y' \to y} f(x(y'), y') \to f(x(y), y). \tag{121}$$

Consider the difference

$$\begin{aligned} |f(x(y'), x') - f(x(y), y)| \leq |f(x(y), y) - f(x(y), y')| \\ + |f(x(y), y') - f(x(y'), y')|. \end{aligned} \tag{122}$$

The first term in (122) tends to zero as $y' \to y$ by the continuity of f so it remains to show that the second term, $B_{y'} := |f(x(y), y') - f(x(y'), y')| \to 0$, as $y' \to y$. Now, we can remove the absolute value since by the optimality of $x(y')$, $f(x(y), y') \geq f(x(y'), y')$. Hence

$$B_{y'} = f(x(y), y') - f(x(y'), y'). \tag{123}$$

Suppose, to the contrary, there exists a sequence $\{y'_n\}_{n=1}^{\infty} \subset Y$ with $y'_n \to y$ such that

$$f\left(x(y), y'_n\right) - f\left(x\left(y'_n\right), y'_n\right) > \epsilon > 0, \quad \forall n \in \mathbb{N}. \tag{124}$$

By the compactness of $\mathcal{K}$, for the sequence $\{x(y'_n)\}_{n=1}^{\infty} \subset \mathcal{K}$, there exists a subsequence $\{x(y'_{n_k})\}_{k=1}^{\infty} \subset \mathcal{K}$ whose limit is $x^* = \lim_{k \to \infty} x(y'_{n_k})$ and $x^* \in \mathcal{K}$ [37, Theorem 3.6(a)]. By the continuity of f

$$\lim_{k \to \infty} f\left(x(y), y'_{n_k}\right) = f(x(y), y) \tag{125}$$

$$\lim_{k \to \infty} f\left(x\left(y'_{n_k}\right), y'_{n_k}\right) = f(x^*, y) \tag{126}$$

since every subsequence of a convergent sequence $\{y'_n\}$ converges to the same limit y. Now (124) can be written as

$$f\left(x(y), y'_{n_k}\right) - f\left(x\left(y'_{n_k}\right), y'_{n_k}\right) > \epsilon > 0, \quad \forall k \in \mathbb{N}. \tag{127}$$

We now take the limit as $k \to \infty$ of (127). Next, we use (125) and (126) to conclude that

$$f(x(y), y) - f(x^*, y) > \epsilon \Rightarrow f(x(y), y) > f(x^*, y) + \epsilon \tag{128}$$

which contradicts the optimality of $x(y)$ in (120). Thus, $B_{y'} \to 0$ as $y' \to y$ and $\lim_{y' \to y} g(y') = g(y)$, which demonstrates the continuity of g on Y. ∎

Lemma 15 (The Continuous Mapping Theorem [49]): Let $(\Omega, \mathcal{B}(\Omega), \nu)$ be a probability space. Let the sequence of random variables $\{X_n\}_{n=1}^{\infty}$ on Ω converge ν-almost surely to X, i.e., $X_n \xrightarrow{a.s.} X$. Let $g : \Omega \to \mathbb{R}$ be a continuous function. Then $g(X_n)$ converges ν-almost surely to $g(X)$, i.e., $g(X_n) \xrightarrow{a.s.} g(X)$.

Proof: Now, using Lemmas 14 and 15, we complete the proof of Theorem 12. First we note from (71) that $\widehat{J}_{e,e'} = \widehat{J}_{e,e'}(\widehat{P}_{e,e'})$, i.e., $\widehat{J}_{e,e'}$ is a function of the empirical distribution on node pairs e and e'. Next, we note that $D(Q||P_{e,e'})$ is a continuous function in $(Q, P_{e,e'})$. If $\widehat{P}_{e,e'}$ is fixed, the expression (71) is a minimization of $D(Q||\widehat{P}_{e,e'})$, over the compact set[25] $\Lambda = \{Q \in \mathcal{P}(\mathcal{X}^4) : I(Q_{e'}) = I(Q_e)\}$; hence, Lemma 14 applies (with the identifications $f \equiv D$ and $\Lambda \equiv \mathcal{K}$) which implies that $\widehat{J}_{e,e'}$ is continuous in the empirical distribution $\widehat{P}_{e,e'}$. Since the empirical distribution $\widehat{P}_{e,e'}$ converges almost surely to $P_{e,e'}$ [30, Sec. 11.2], $\widehat{J}_{e,e'}(\widehat{P}_{e,e'})$ also converges almost surely to $J_{e,e'}$, by Lemma 15. ∎

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees and Associate Editor A. Krzyzak who have helped to improve the exposition. One reviewer, in particular, helped highlight the connection of this work with robust hypothesis testing, leading to Section V-D. The authors would also like to thank Prof. L. Zheng, M. Agrawal, and A. Olshevsky for many stimulating discussions.

REFERENCES

[1] 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. Information Theory*, Seoul, Korea, Jul. 2009, pp. 1140–1144.

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

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

[4] A. Chechetka and C. Guestrin, "Efficient principled learning of thin junction trees," presented at the Neural Inf. Process. Syst., 2007.

[5] M. J. Wainwright, P. Ravikumar, and J. D. Lafferty, "High-dimensional graphical model selection using -regularized logistic regression," in *Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2006, pp. 1465–1472.

[6] S. Lee, V. Ganapathi, and D. Koller, "Efficient structure learning of Markov networks using ℓ_1-regularization," presented at the Neural Inf. Process. Syst., 2006.

[7] J. Johnson, V. Chandrasekaran, and A. S. Willsky, "Learning Markov structure by maximum entropy relaxation," presented at the Artificial Intelligence and Statistics (AISTATS), 2007.

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

[9] M. Dudik, S. Phillips, and R. Schapire, "Performance guarantees for regularized maximum entropy density estimation," presented at the Conf. Learning Theor. (COLT), 2004.

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

[11] F. D. Hollander, *Large Deviations (Fields Institute Monographs, 14)*. Providence, RI: AMS, 2000.

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

[13] I. Csiszár and F. Matúš, "Information projections revisited," *IEEE Trans. Inf. Theory*, vol. 49, no. 6, pp. 1474–1490, Jun. 2003.

[14] S. Borade and L. Zheng, "I-projection and the geometry of error exponents," presented at the Allerton Conf. Communication, Control, Comput., 2006.

[15] S. Borade and L. Zheng, "Euclidean information theory," presented at the Allerton Conf. Communication, Control, Comput., 2007.

[25] Compactness of Λ was proven in Theorem 2 cf. Equation (84).

[16] D. Karger and N. Srebro, "Learning Markov networks: Maximum bounded tree-width graphs," in *Proc. Symp. Discrete Algorithms*, 2001, pp. 392–401.
[17] F. Bach and M. I. Jordan, "Thin junction trees," presented at the Neural Inf. Process. Syst., 2002.
[18] D. Heckerman and D. Geiger, Learning Bayesian networks, Microsoft Research, Redmond, WA, 1994, Tech. Rep. MSR-TR-95-02.
[19] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
[20] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.
[21] O. Zuk, S. Margel, and E. Domany, "On the number of samples needed to learn the correct structure of a Bayesian network," presented at the 22nd Annu. Conf. Uncertainty Artificial Intell. (UAI-06), 2006.
[22] M. J. Choi, V. Y. F. Tan, A. Anandkumar, and A. S. Willsky, "Learning latent tree graphical models," *J. Mach. Learn. Res.*, Sep. 2010, ArXiv:1009.2722, submitted for publication.
[23] L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.*, vol. 15, no. 6, pp. 1191–1253, Jun 1, 2003.
[24] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Struct. Alg.*, pp. 163–193, 2001.
[25] J.-R. Chazottes and D. Gabrielli, "Large deviations for empirical entropies of g-measures," *Nonlinearity*, vol. 18, pp. 2545–2563, Nov. 2005.
[26] M. Hutter, "Distribution of mutual information," in *Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2001, pp. 399–406.
[27] V. Y. F. Tan, A. Anandkumar, and A. S. Willsky, "Learning Gaussian tree models: Analysis of error exponents and extremal structures," *IEEE Trans. Signal Process.*, vol. 58, no. 5, pp. 2701–2714, May 2010.
[28] A. Kester and W. Kallenberg, "Large deviations of estimators," *Ann. Statist.*, pp. 648–664, 1986.
[29] R. Bahadur, S. Zabell, and J. Gupta, "Large deviations, tests, and estimates," *Asymptotic Theor. Stat. Tests Estim.*, pp. 33–64, 1980.
[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley-Interscience, 2006.
[31] K. Ryu, "Econometric analysis of mixed parameter models," *J. Econ. Theor. Econ.*, vol. 5, no. 113–124, 1999.
[32] H. L. V. Trees, *Detection, Estimation, and Modulation Theory, Part I*. Hoboken, NJ: Wiley, 1968.
[33] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Stat. Decisions, Supplementary*, no. 1, pp. 205–237, Jul. 1984.
[34] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd ed. New York: McGraw-Hill Science/Engineering/Math, 2003.
[35] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," presented at the Amer. Math. Soc., Feb. 1956.
[36] R. C. Prim, "Shortest connection networks and some generalizations," *Bell Syst. Tech. J.*, vol. 36, 1957.
[37] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York: McGraw-Hill Science/Engineering/Math, 1976.
[38] V. Y. F. Tan, A. Anandkumar, and A. S. Willsky, "Learning high-dimensional Markov forest distributions: Analysis of error rates," *J. Mach. Learning Research*, May 2010, ArXiv:1005.0766, submitted for publication.
[39] P. J. Huber and V. Strassen, "Minimax tests and the Neyman-Pearson lemma for capacities," *Ann. Statist.*, vol. 1, pp. 251–263.
[40] C. Pandit and S. P. Meyn, "Worst-case large-deviations with application to queueing and information theory," *Stochastic Processes Appl.*, vol. 116, no. 5, pp. 724–756, May 2006.
[41] O. Zeitouni and M. Gutman, "On universal hypotheses testing via large deviations," *IEEE Trans. Inf. Theory*, vol. 37, no. 2, pp. 285–290, Mar. 1991.
[42] J. Unnikrishnan, D. Huang, S. Meyn, A. Surana, and V. V. Veeravalli, "Universal and composite hypothesis testing via mismatched divergence," *IEEE Trans. Inf. Theory*, to be published.
[43] V. Y. F. Tan, A. Anandkumar, and A. S. Willsky, "Error exponents for composite hypothesis testing of Markov forest distributions," in *Proc. Int. Symp. Inf. Theor.*, Austin, TX, Jun. 2010, pp. 1613–1617.
[44] E. Abbe and L. Zheng, "Linear universal decoding for compound channels: An Euclidean geometric approach," in *Proc. Int. Symp. Inf. Theor.*, 2008, pp. 1098–1102.
[45] M. S. Pinsker, *Information and Information Stability of Random Variables*. Oakland, CA: Holden-Day, 1964.
[46] J. N. Laneman, "On the distribution of mutual information," presented at the Inf. Theor. Applicat. Workshop, 2006.
[47] S.-I. Amari and H. Nagaoka, *Methods of Information Geometry*. Providence, RI: AMS, 2000.
[48] S.-I. Amari, "Information geometry on hierarchy of probability distributions," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1701–1711, Jul. 2001.
[49] H. B. Mann and A. Wald, "On the statistical treatment of linear stochastic difference equations," *Econometrics*, vol. 11, pp. 173–220, 1943.
[50] P. Billingsley, *Weak Convergence of Measures: Applications in Probability*. Philadelphia, PA: SIAM, 1987.
[51] I. Csiszár and J. Korner, *Information Theory: Coding Theorems For Discrete Memoryless Systems*. Budapest, Hungary: Akademiai Kiado, 1997.

Vincent Y. F. Tan (S'07–M'11) received the B.A. and M.Eng. degrees in electrical engineering from Cambridge University, Cambridge, U.K., in 2005, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, in February 2011.

He was a research intern with Microsoft Research in 2008 and 2009. He is currently a postdoctoral researcher at the Department of Electrical and Computer Engineering, University of Wisconsin-Madison. His research interests include statistical signal processing, machine learning and information theory.

Dr. Tan received the Public Service Commission (PSC) Scholarship in 2001 and the National Science Scholarship (Ph.D.) 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 electrical engineering. He has served as a reviewer for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, and the *Journal of Machine Learning Research*.

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

She was with the Stochastic Systems Group, Massachusetts Institute of Technology, Cambridge, MA, as a postdoctoral researcher. She has been an assistant professor at EECS Dept. at University of California Irvine since July 2010. Her research interests are in the area of learning and inference of graphical models.

She is the recipient of the 2009 Best Thesis Award by the ACM Sigmetrics Society, the 2008 IEEE Signal Processing Society Young Author Best Paper Award, the 2008 IBM Fran Allen Ph.D. fellowship, and the student paper award at 2006 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP). She has served as a reviewer for the IEEE TRANSACTIONS ON INFORMATION THEORY, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, THE IEEE Transactions on Wireless Communications, and the IEEE SIGNAL PROCESSING LETTERS.

Lang Tong (S'87–M'91–SM'01–F'05) received the B.E. degree from Tsinghua University, Beijing, China, in 1985, and the M.S. and Ph.D. degrees in electrical engineering from the University of Notre Dame, Notre Dame, IN, in 1987 and 1991, respectively.

He is the Irwin and Joan Jacobs Professor in Engineering at Cornell University, Ithaca, NY. He was a Postdoctoral Research Affiliate at the Information Systems Laboratory, Stanford University, Stanford, CA, in 1991. He was the 2001 Cor Wit Visiting Professor at the Delft University of Technology, Delft, The Netherlands, and held visiting positions at Stanford University and the University of California, Berkeley. His research is in the general area of statistical signal processing, communications and networking, and information theory.

Dr. Tong received the 1993 Outstanding Young Author Award from the IEEE Circuits and Systems Society, the 2004 best paper award (with M. Dong) from the IEEE Signal Processing Society, and the 2004 Leonard G. Abraham Prize Paper Award from the IEEE Communications Society (with P. Venkitasubramaniam and S. Adireddy). He is also a coauthor of seven student paper awards. He received the Young Investigator Award from the Office of Naval Research. He has served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE TRANSACTIONS ON INFORMATION THEORY, and the IEEE SIGNAL PROCESSING LETTERS. He was named as a 2009–2010 Distinguished Lecturer by the IEEE Signal Processing Society.

Alan S. Willsky (S'70–M'73–SM'82–F'86) received the B.S. 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 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 the text *Signals and Systems* (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, the Doctorat Honoris Causa from Universit de Rennes in 2005, and the IEEE Signal Processing Society Technical Achievement Award in 2010. 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 the outstanding paper in the year 2007.