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

Vincent Y. F. Tan*, Animashree Anandkumar*†, Lang Tong† and Alan S. Willsky*

* Stochastic Systems Group, LIDS, MIT, Cambridge, MA 02139, Email: {vtan,animakum,willsky}@mit.edu
† ECE Dept., Cornell University, Ithaca, NY 14853, Email: {aa332@,ltong@ece.}cornell.edu

***Abstract*—The problem of maximum-likelihood learning of the structure of an unknown discrete distribution from samples is considered when the distribution is Markov on a tree. Large-deviation analysis of the error in estimation of the set of edges of the tree is performed. Necessary and sufficient conditions are provided to ensure that this error probability decays exponentially. These conditions are based on the mutual information between each pair of variables being distinct from that of other pairs. The rate of error decay, or error exponent, is derived using the large-deviation principle. The error exponent is approximated using Euclidean information theory and is given by a ratio, to be interpreted as the signal-to-noise ratio (SNR) for learning. Numerical experiments show the SNR approximation is accurate.**

***Index Terms*—Large-deviations, Tree structure learning, Error exponents, Euclidean Information Theory.**

I. Introduction

The estimation of a distribution from samples is a classical problem in machine learning and is challenging for high-dimensional multivariate distributions. In this respect, graphical models [1] provide a significant simplification of the joint distribution, and incorporate a Markov structure in terms of a graph defined on the set of nodes. Many specialized algorithms exist for learning graphical models with sparse graphs.

A special scenario is when the graph is a tree. In this case, the classical Chow-Liu algorithm [2] finds the maximum-likelihood (ML) estimate of the probability distribution by exploiting the Markov structure to learn the tree edges in terms of a maximum-weight spanning tree (MWST). The ML estimator learns the distribution correctly as we obtain more learning samples (consistency). These learning samples are drawn independently from the distribution.

In this paper, we study the performance of ML estimator in terms of the nature of convergence with increasing sample size. Specifically, we are interested in the event that the set of edges of the ML tree is not the true set of edges. We address the following questions: Is there exponential decay of error in structure learning as the number of samples goes to infinity? If so, what is the rate of decay of the error probability? How does the rate depend on the parameters of the distribution? Our analysis and answers to these questions provide us insights into the distributions and the edges where we are more likely to make an error when learning the tree structure.

Summary of Main Results and Related Work

Learning the structure of graphical models is an extensively studied problem (e.g. [2]–[5]). The previous works look at establishing consistency of the estimators, while a few prove the estimators to have exponential rate of error decay under some technical conditions in the Gaussian case [4]. However, to the best of our knowledge, none of the works quantifies the exact rate of decay of error for structure learning.

Following the seminal work of Chow and Liu in [2], a number of works have looked at learning graphical models. Using the maximum entropy principle as a learning technique, Dudik et al. [3] provides strong consistency guarantees on the learned distribution in terms of the log-likelihood of the samples. Wainwright et al. [5] also proposed a regularization method for learning the graph structure based on ℓ_1 logistic regression and provide theoretical guarantees for learning the correct structure as the number of samples, the number of variables, and the neighborhood size grow. Meinshausen et al. [4] consider learning the structure for Gaussian models, and show that the error probability of getting the structure wrong decays exponentially. However, the rate is not provided.

The main contributions of this paper are three-fold. First, using the large-deviation principle (LDP) [6] 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, using ideas from Euclidean Information Theory [7], we provide a succinct and intuitively appealing 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. It corroborates the intuition that if the edges belonging to the true tree model are strongly distinguishable from the non-edges using the samples, we can expect the rate of decay of error to be large.

All the results are stated without proof. The reader may refer to http://web.mit.edu/vtan/www/isit09 for the details.

II. Problem Statement and Chow-Liu Algorithm

An *undirected graphical model* [1] is a probability distribution that factorizes according to the structure of an

The first author is supported by A*STAR, Singapore, in part by a MURI funded through ARO Grant W911NF-06-1-0076 and in part under a MURI through AFOSR Grant FA9550-06-1-0324. The second author is supported by the IBM Ph.D Fellowship for the year 2008-09 and is currently a visiting student at MIT, Cambridge, MA 02139. This work is also supported in part through collaborative participation in Communications and Networks Consortium sponsored by the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0011 and by the Army Research Office under Grant ARO-W911NF-06-1-0346.

underlying undirected graph. More explicitly, a vector of random variables $\mathbf{x} = (x_1, \ldots, x_d)$ is said to be *Markov* on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{1, \ldots, d\}$ and $\mathcal{E} \subset \binom{\mathcal{V}}{2}$ if $P(x_i | x_{\mathcal{V} \setminus \{i\}}) = P(x_i | x_{\mathcal{N}(i)})$ where $\mathcal{N}(i)$ is the set of neighbors of x_i. In this paper, we are given a set of d-dimensional samples (or observations) $\mathbf{x}^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ where each $\mathbf{x}_k \in \mathcal{X}^d$ and $\mathcal{X}$ is a finite set. Each element of the set of samples $\mathbf{x}_k$ is drawn independently from some unknown, positive everywhere distribution P whose support is $\mathcal{X}^d$, that is $P \in \mathcal{P}(\mathcal{X}^d)$, the set of all distributions supported on $\mathcal{X}^d$. We further assume that the graph of P, denoted $T_P = (\mathcal{V}, \mathcal{E}_P)$, belongs to the set of spanning trees[1] on d nodes $\mathcal{T}^d$. The set of d-dimensional tree distributions is denoted $\mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)$. Tree distributions possess the following factorization property [1]

$$P(\mathbf{x}) = \prod_{i \in \mathcal{V}} P_i(x_i) \prod_{(i,j) \in \mathcal{E}_P} \frac{P_{i,j}(x_i, x_j)}{P_i(x_i) P_j(x_j)}, \tag{1}$$

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. Given $\mathbf{x}^n$, we define the *empirical distribution* of P to be $\widehat{P}(\mathbf{x}) = N(\mathbf{x}|\mathbf{x}^n)/n$, where $N(\mathbf{x}|\mathbf{x}^n)$ is the number of times $\mathbf{x} \in \mathcal{X}^d$ occurred in $\mathbf{x}^n$.

Using the samples $\mathbf{x}^n$, we can use the Chow-Liu algorithm [2], reviewed in Section II-A, to compute P_{ML}, the ML tree-structured distribution with edge set $\mathcal{E}_{\text{ML}}$. It is known [8] that as $n \to \infty$, $\mathcal{E}_{\text{ML}}$ approaches $\mathcal{E}_P$, the true tree structure. But at what rate does this happen for a given tree distribution P? Is the error decay exponential? In this paper, we use the Large-Deviation Principle (LDP) [6] to quantify the exponential rate at which the probability of the error event

$$\mathcal{A}_n := \{\mathcal{E}_{\text{ML}} \neq \mathcal{E}_P\} \tag{2}$$

decays to zero. In other words, given a distribution P and $\mathbb{P} := P^n$, we are interested to study the *rate* K_P given by,

$$K_P := \lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}(\mathcal{A}_n), \tag{3}$$

whenever the limit exists. K_P is the *error exponent* for the event $\mathcal{A}_n$ as defined in (2). When $K_P > 0$, there is exponential decay of error probability in structure learning, and we would like to provide conditions to ensure this.

In addition, we define the set of disjoint events $\mathcal{U}_n(T)$ that the graph of the ML estimate P_{ML} of the tree distribution is a tree T different from the true tree T_P, *i.e.*,

$$\mathcal{U}_n(T) \ := \ \begin{cases} \{T_{\text{ML}} = T\}, & \text{if } \ T \in \mathcal{T}^d \setminus \{T_P\}, \\ \emptyset, & \text{if } \ T = T_P. \end{cases} \tag{4}$$

From (2), $\mathcal{A}_n = \bigcup_{T \in \mathcal{T}^d \setminus \{T_P\}} \mathcal{U}_n(T)$. We also define the exponent for each error event $\mathcal{U}_n(T)$ as

$$\Upsilon(T) := \lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}\left(\mathcal{U}_n(T)\right). \tag{5}$$

Definition The *dominant error tree* $T_P^* = (\mathcal{V}, \mathcal{E}_P^*)$ is the spanning tree given by

$$T_P^* = \operatorname*{argmin}_{T \in \mathcal{T}^d \setminus \{T_P\}} \ \Upsilon(T). \tag{6}$$

In section IV, we characterize the dominant error tree and its relation to the error exponent K_P in (3).

A. Maximum-Likelihood Learning of Tree Distributions

In this section, we review the Chow-Liu algorithm [2] for learning the ML tree distribution P_{ML} given a set of n samples $\mathbf{x}^n$ drawn independently from a tree distribution P. The Chow-Liu algorithm solves the following ML problem:

$$P_{\text{ML}} = \operatorname*{argmax}_{Q \in \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)} \sum_{k=1}^{n} \log Q(\mathbf{x}_k) = \operatorname*{argmin}_{Q \in \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)} D(\widehat{P} || Q). \tag{7}$$

For the above problem, the ML tree model will *always* be the projection of the empirical distribution $\widehat{P}$ onto some tree structure [2], *i.e.*, $P_{\text{ML}}(x_i) = \widehat{P}_i(x_i)$ for all $i \in \mathcal{V}$ and $P_{\text{ML}}(x_i, x_j) = \widehat{P}_{i,j}(x_i, x_j)$ for all $(i,j) \in \mathcal{E}_{\text{ML}}$. Thus, the optimization problem in (7) reduces to a search for the tree structure of P_{ML}. By using the fact that the graph of Q is a tree, the KL divergence in (7) decomposes into a sum of mutual information quantities. Hence, if $\mathcal{E}_Q$ is the edge set of the tree distribution Q, the optimization for the structure of P_{ML} is

$$\mathcal{E}_{\text{ML}} = \operatorname*{argmax}_{\mathcal{E}_Q : Q \in \mathcal{D}(\mathcal{X}^d, \mathcal{T}^d)} \ \sum_{e \in \mathcal{E}_Q} I(\widehat{P}_e), \tag{8}$$

where $I(\widehat{P}_e)$ is the *empirical mutual information* given the data $\mathbf{x}^n$, defined for $e = (i,j)$ as:

$$I(\widehat{P}_e) := \sum_{(x_i, x_j) \in \mathcal{X}^2} \widehat{P}_{i,j}(x_i, x_j) \log \frac{\widehat{P}_{i,j}(x_i, x_j)}{\widehat{P}_i(x_i) \widehat{P}_j(x_j)}. \tag{9}$$

Note that $\mathcal{E}_P$ is the structure obtained from the MWST algorithm without any error, *i.e.*, with the true mutual informations as edge weights. To solve (8), we use the samples $\mathbf{x}^n$ to compute $I(\widehat{P}_e)$, for each node pair $e \in \binom{\mathcal{V}}{2}$ given the empirical distribution $\widehat{P}$. Subsequently, we use these as the edge weights for the MWST problem in (8). Note that the search for the MWST is not the same as that for largest $d-1$ mutual information quantities as one has to take into consideration the tree constraint. There are well-known algorithms [9] that solve the MWST problem in $\mathcal{O}(d^2 \log d)$ time.

III. LDP for Empirical Mutual Information

To compute the exponent for the error in structure learning K_P, we first consider a simpler event. Let $e, e' \in \binom{\mathcal{V}}{2}$ be any two node pairs satisfying $I(P_e) > I(P_{e'})$, where $I(P_e)$ and $I(P_{e'})$ are the true mutual informations on node pairs e and e' respectively. We are interested in the *crossover event of the empirical mutual informations* defined as:

$$\mathcal{C}_{e,e'} := \left\{ I(\widehat{P}_e) \leq I(\widehat{P}_{e'}) \right\}. \tag{10}$$

The occurrence of this event *may* potentially lead to an error in structure learning when $\mathcal{E}_{\text{ML}}$ differs from $\mathcal{E}_P$. In the next section, we will see how these crossover events relate to K_P. Now, we would like to compute the *crossover rate for empirical mutual informations* $J_{e,e'}$ as:

$$J_{e,e'} := \lim_{n \to \infty} -\frac{1}{n} \log \mathbb{P}\left(\mathcal{C}_{e,e'}\right). \tag{11}$$

[1] In a spanning tree, none of pairwise joint distributions $P_{i,j}$ are allowed to be product distributions.

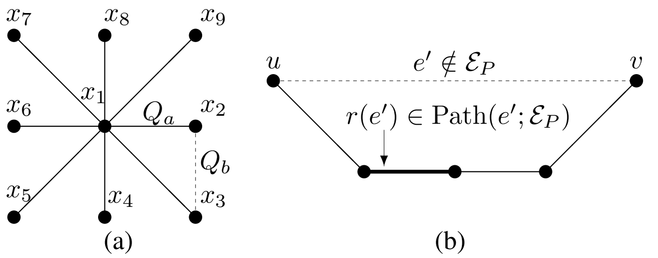
x_7
x_8
x_9
x_1
x_6
Q_a
x_2
Q_b
x_5
x_4
x_3
(a)
u
$e' \notin \mathcal{E}_P$
v
$r(e') \in \mathrm{Path}(e'; \mathcal{E}_P)$
(b)

Fig. 1. (a) The symmetric "star" graph with $d = 9$. (b) The path associated to the non-edge $e' = (u, v) \notin \mathcal{E}_P$, 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'. $r(e') \in \mathrm{Path}(e', \mathcal{E}_P)$ is the dominant replacement edge associated to $e' \notin \mathcal{E}_P$.

Intuitively, if the difference between the true mutual informations $I(P_e) - I(P_{e'})$ is large, we expect the crossover rate to be large. Consequently, the probability of the crossover event would be small. Note that $J_{e,e'}$ in (11) is not the same as K_P in (3). We will see that the rate $J_{e,e'}$ depends, not only on the mutual informations $I(P_e)$ and $I(P_{e'})$, but also on the distribution $P_{e,e'}$ of the variables on node pairs e and e'.

Theorem 1 (LDP for Empirical MIs): Let the distribution on two node pairs e and e' be $P_{e,e'} \in \mathcal{P}(\mathcal{X}^4)$.[2] The crossover rate for empirical mutual information is

$$J_{e,e'} = \inf_{Q\in\mathcal{P}(\mathcal{X}^4)} \left\{D(Q\,||\,P_{e,e'}) : I(Q_{e'}) = I(Q_e)\right\}, \tag{12}$$

where $Q_e, Q_{e'} \in \mathcal{P}(\mathcal{X}^2)$ are distributions on e, e' resp., *i.e.*, $Q_e = \sum_{x_{e'}} Q$ and similarly for $Q_{e'}$. The infimum in (12) is attained for some $Q^*_{e,e'} \in \mathcal{P}(\mathcal{X}^d)$ satisfying $I(Q^*_{e'}) = I(Q^*_e)$.

The proof of this theorem hinges on Sanov's theorem [6, Sec. 6.2] and the contraction principle [6, Thm 4.2.1]. We remark that the rate $J_{e,e'}$ is strictly positive since we assumed, a-priori, that e and e' satisfy $I(P_e) > I(P_{e'})$.

A. Example: The ‘Symmetric ‘Star” Graph

It is instructive to study a simple example – the "star" graph, shown in Fig. 1(a) with $\mathcal{E}_P = \{(1, i) : i = 2, \dots, d\}$. 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: (i) $P_{1,i} = Q_a$ for all $2 \le i \le d$. (ii) $P_{i,j} = Q_b$ for all $2 \le i, j \le d$. (iii) $P_{1,i,j,k} = Q_{a,b}$ for all $2 \le i, j, k \le d$. Thus, the joint distribution on the central node and any outer node is Q_a, while the joint distribution of any two outer nodes is Q_b. Note that Q_a and Q_b are the pairwise marginals of $Q_{a,b}$. Furthermore, we assume that $I(Q_a) > I(Q_b) > 0$.

Proposition 2: For the "star" graph with the distributions as described, K_P, the error exponent for structure learning, is

$$K_P = \inf_{R_{1,2,3,4}\in\mathcal{P}(\mathcal{X}^4)} \left\{D(R_{1,2,3,4}\,||\,Q_{a,b}) : I(R_{1,2}) = I(R_{3,4})\right\},$$

where $R_{1,2}$ and $R_{3,4}$ are the pairwise marginals of $R_{1,2,3,4}$.

In general, it is not easy to derive the error exponent K_P since crossover events for different node pairs affect the learned structure in a complex manner. We now provide an expression for K_P by identifying the dominant error tree.

IV. Error Exponent for Structure Learning

The analysis in the previous section characterized the rate for the crossover event for empirical mutual information $\mathcal{C}_{e,e'}$. In this section, we connect these events $\{\mathcal{C}_{e,e'}\}$ to the quantity of interest K_P in (3). Not all the events in the set $\{\mathcal{C}_{e,e'}\}$ contribute to the overall event error $\mathcal{A}_n$ in (2) because of the global spanning tree constraint for the learned structure $\mathcal{E}_{\mathrm{ML}}$.

A. Identifying the Dominant Error Tree

Definition Given a node pair $e' \notin \mathcal{E}_P$, its *dominant replacement edge* $r(e') \in \mathcal{E}_P$ is given by the edge along the unique path (See Fig. 1(b)) connecting the nodes in e' (denoted $\mathrm{Path}(e'; \mathcal{E}_P)$) and having minimum crossover rate, *i.e.*,

$$r(e') := \operatorname*{argmin}_{e\in\mathrm{Path}(e';\mathcal{E}_P)} J_{e,e'}. \tag{13}$$

Note that if we replace the true edge $r(e')$ by a non-edge e', the tree constraint is still satisfied. This is important since such replacements lead to an error in structure learning.

Theorem 3 (Error Exponent as a Single Crossover Event): A dominant error tree $T^*_P = (\mathcal{V}, \mathcal{E}^*_P)$, has edge set $\mathcal{E}^*_P = \mathcal{E}_P \cup \{e^*\} \setminus \{r(e^*)\}$, where

$$e^* := \operatorname*{argmin}_{e'\notin\mathcal{E}_P} J_{r(e'),e'}, \tag{14}$$

with $r(e')$, defined in (13), being the dominant replacement edge associated with $e' \notin \mathcal{E}_P$. Furthermore, the error exponent K_P, which is the rate at which the ML tree $\mathcal{E}_{\mathrm{ML}}$ differs from the true tree structure $\mathcal{E}_P$, is given by,

$$K_P = J_{r(e^*),e^*} = \min_{e'\notin\mathcal{E}_P} \min_{e\in\mathrm{Path}(e';\mathcal{E}_P)} J_{e,e'}, \tag{15}$$

where $e^* \notin \mathcal{E}_P$ is given in (14).

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 (2). We see that the dominant error tree T^*_P differs from the true tree T_P in exactly one edge. Note that the result in (15) is exponentially tight in the sense that $\mathbb{P}(\mathcal{A}_n) \doteq \exp(-nK_P)$. From (15), we see that if at least one of the crossover rates $J_{e,e'}$ is zero, then 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 $K_P > 0$, we have exponential decay in the desired probability of error.

Theorem 4 (Conditions for exponential decay): The following three statements are equivalent.

(a) The error probability decays exponentially, *i.e.*, $K_P > 0$.
(b) The mutual information quantities satisfy $I(P_{e'}) \neq I(P_e), \forall\, e \in \mathrm{Path}(e'; \mathcal{E}_P)$, $e' \notin \mathcal{E}_P$.
(c) T_P is *not* a proper forest[3] as assumed in Section II.

[2] If e and e' share a node, $P_{e,e'} \in \mathcal{P}(\mathcal{X}^3)$. This does not change the subsequent exposition significantly.

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

Condition (b) states that, for each non-edge e', we need $I(P_{e'})$ to be different from 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* non-degenerate tree distribution in which all the pairwise joint distributions are not product distributions, then we have exponential decay in the probability of error.

C. Computational Complexity to Compute Error Exponent

We now provide an upper bound on the complexity to compute K_P. The *diameter* of the tree $T_P = (\mathcal{V}, \mathcal{E}_P)$ is

$$\zeta(T_P) := \max_{u,v\in\mathcal{V}} L(u,v), \tag{16}$$

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 non-edge $e' = (u,v)$ in the subtree in Fig. 1(b).

Theorem 5 (Computational Complexity): The number of computations of $J_{e,e'}$ required to determine K_P is upper bounded by $(1/2)\zeta(T_P)(d-1)(d-2)$.

Thus, if the diameter of the tree is relatively low and independent of number of nodes d, the complexity is quadratic in d. For instance, for a "star" network, the diameter $\zeta(T_P) = 2$. For a balanced tree, $\zeta(T_P) = \mathcal{O}(\log d)$, hence the number of computations is $\mathcal{O}(d^2 \log d)$. The complexity is vastly reduced as compared to exhaustive search which requires $d^{d-2} - 1$ computations, since there are d^{d-2} trees on d nodes.

V. Euclidean Approximations

In order to gain more insight into the error exponent, we make use of *Euclidean approximations* [7] of information-theoretic quantities to obtain an approximate but closed-form solution to (12), which is non-convex and hard to solve exactly. To this end, we first approximate the crossover rate $J_{e,e'}$. This will allow us to understand which pairs of edges have a higher probability of crossing over and hence result in an error as defined in (2). It turns out that $J_{e,e'}$ intuitively depends on the "separation" of the mutual information values. It *also* depends on the uncertainty of the mutual information estimates.

Roughly speaking, our strategy is to "convexify" the objective and the constraints in (12). To do so, we recall that if P and Q are two distributions with the same support, the KL divergence can be approximated [7] by

$$D(Q\,||\,P) \approx \frac{1}{2}\|Q-P\|_P^2, \tag{17}$$

where $\|y\|_w^2$ denotes the weighted squared norm of y, *i.e.*, $\|y\|_w^2 = \sum_i y_i^2/w_i$. This bound is tight whenever $P \approx Q$. By $\approx$, we mean that P is close to Q entry-wise, *i.e.*, $\|P-Q\|_\infty < \epsilon$ for some small $\epsilon > 0$. In fact, if $P \approx Q$, $D(P\,||\,Q) \approx D(Q\,||\,P)$. We will also need following notion.

Definition Given a joint distribution $P_e = P_{i,j}$ on $\mathcal{X}^2$ with marginals P_i and P_j, the *information density* [10] function, denoted by $s_{i,j} : \mathcal{X}^2 \to \mathbb{R}$, is defined as

$$s_{i,j}(x_i,x_j) := \log\frac{P_{i,j}(x_i,x_j)}{P_i(x_i)P_j(x_j)}, \quad \forall\,(x_i,x_j)\in\mathcal{X}^2. \tag{18}$$

If $e = (i,j)$, we will also use the notation $s_e(x_i,x_j) = s_{i,j}(x_i,x_j)$. The mutual information is the expectation of the information density, *i.e.*, $I(P_e) = \mathbb{E}[s_e]$.

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.

Definition 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

$$\|P_e - P_{e'}\|_\infty := \max_{(x_i,x_j)\in\mathcal{X}^2} |P_e(x_i,x_j) - P_{e'}(x_i,x_j)| < \epsilon. \tag{19}$$

This condition is needed because if (19) holds, then by continuity of the mutual information, there exists a $\delta > 0$ such that $|I(P_e) - I(P_{e'})| < \delta$, which means that the mutual information quantities are difficult to distinguish and the approximation in (17) is accurate. Note that proximity of the mutual informations is not sufficient for the approximation to hold since we have seen from Theorem 1 that $J_{e,e'}$ depends not only on the mutual informations but on $P_{e,e'}$. We now define the *approximate crossover rate* on e and e' as

$$\widetilde{J}_{e,e'} := \inf_{Q\in\mathcal{P}(\mathcal{X}^4)} \left\{\frac{1}{2}\|Q - P_{e,e'}\|^2_{P_{e,e'}} : Q \in \mathcal{Q}(P_{e,e'})\right\}, \tag{20}$$

where the (linearized) constraint set is

$$\begin{aligned}\mathcal{Q}(P_{e,e'}) := \Big\{Q \in \mathcal{P}(\mathcal{X}^4) : I(P_{e'}) + \langle \nabla_{P_{e'}} I(P_{e'}), Q - P_{e,e'}\rangle \\ = I(P_e) + \langle \nabla_{P_e} I(P_e), Q - P_{e,e'}\rangle\Big\},\end{aligned} \tag{21}$$

where $\nabla_{P_e} I(P_e)$ is the gradient vector of the mutual information with respect to the joint distribution P_e. We also define the *approximate error exponent* as

$$\widetilde{K}_P := \min_{e'\notin\mathcal{E}_P} \min_{e\in\mathrm{Path}(e';\mathcal{E}_P)} \widetilde{J}_{e,e'}. \tag{22}$$

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.[4]

Theorem 6 (Euclidean approximation of $J_{e,e'}$): The approximate crossover rate for the empirical mutual information quantities, defined in (20), is given by

$$\widetilde{J}_{e,e'} = \frac{(\mathbb{E}[s_{e'} - s_e])^2}{2\,\mathrm{Var}(s_{e'} - s_e)} = \frac{(I(P_{e'}) - I(P_e))^2}{2\,\mathrm{Var}(s_{e'} - s_e)}, \tag{23}$$

where s_e is the information density defined in (18) and the expectations are with respect to $P_{e,e'}$. The approximation $\widetilde{J}_{e,e'}$ is asymptotically accurate, *i.e.*, as $\epsilon \to 0$, $\widetilde{J}_{e,e'} \to J_{e,e'}$.

Corollary 7 (Euclidean approximation of K_P): The approximation $\widetilde{K}_P$ is asymptotically accurate if either of the following 2 conditions.

(a) The joint distribution $P_{r(e'),e'}$ satisfies the ϵ-very noisy condition for every $e' \notin \mathcal{E}_P$.

[4] We say that a family of approximations $\{\widetilde{\theta}(\epsilon) : \epsilon > 0\}$ of a true parameter θ is asymptotically accurate if the approximations converge to θ as $\epsilon \to 0$.

(b) The joint distribution $P_{r(e^*),e^*}$ satisfies the ϵ-very noisy condition but all the other joint distributions on the non-edges $e' \notin \mathcal{E}_P \cup \{e^*\}$ and their dominant replacement edges $r(e')$ *do not* 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 (23) 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 $\widetilde{J}_{e,e'}$. The expression in (23) 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. 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 in (22).

We now comment on our assumption of $P_{e,e'}$ satisfying the ϵ-very noisy condition, under which the approximation is tight. When $P_{e,e'}$ is ϵ-very noisy, then we have $|I(P_e)-I(P_{e'})| < \delta$. Thus it is very hard to distinguish the relative magnitudes of $I(\widehat{P}_e)$ and $I(\widehat{P}_{e'})$. The particular problem of learning the distribution $P_{e,e'}$ from samples is *very noisy*. Under these conditions, the approximation in (23) is accurate. In fact, ratio of the approximate crossover rate and the true crossover rate approaches unity as $\epsilon \to 0$.

VI. Numerical Experiments

In this section, we perform numerical experiments to study the accuracy of the Euclidean approximations. We do this by analyzing under which regimes $\widetilde{J}_{e,e'}$ in (23) is close to the true crossover rate $J_{e,e'}$ in (12). We parameterize a symmetric "star" graph distribution with $d = 4$ variables and with a single parameter $\gamma > 0$. We let $\mathcal{X} = \{0, 1\}$, *i.e.*, all the variables are binary. The central node is x_1 and the outer nodes are x_2, x_3, x_4. For the parameters, we set $P_1(x_1 = 0) = 1/3$ and

$$P_{i|1}(x_i = 0|x_1 = a) = \frac{1}{2} + (-1)^a\gamma, \quad a = 0, 1, \qquad (24)$$

for $i = 2, 3, 4$. 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 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 edge by varying γ. By symmetry, there is only one crossover rate and hence it is also the error exponent for the error event $\mathcal{A}_n$ in (2).

We vary γ from 0 to 0.2 and plot both the true and approximate rates against $I(P_e) - I(P_{e'})$ in Fig. 2, where e denotes any edge and e' denotes any non-edge. We note from Fig. 2 that both rates increase as the difference $I(P_e)-I(P_{e'})$ increases. This is in line with our intuition because if $I(P_e) - I(P_{e'})$ is large, the crossover rate is also large. We also observe that if the difference is small, the true and approximate rates are close. This is in line with the assumptions of Theorem 6; that if $P_{e,e'}$ satisfies the ϵ-very noisy condition, then the mutual informations $I(P_e)$ and $I(P_{e'})$ are close and 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.

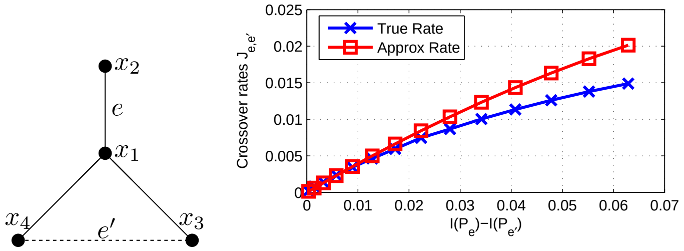

Fig. 2. Left: The true model is a symmetric star where the mutual informations satisfy $I(P_{1,2}) = I(P_{1,3}) = I(P_{1,4})$ and $I(P_{e'}) < I(P_{1,2})$ for any non-edge e'. Right: Comparison of True and Approximate Rates.

VII. Conclusion

In this paper, we presented a solution to the problem of finding the error exponent for ML tree structure learning by employing 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 3 and 6 provide the intuition as to how errors occur for learning discrete tree distributions.

References

[1] S. Lauritzen, *Graphical Models*. Oxford University Press, USA, 1996.

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

[3] M. Dudik, S. Phillips, and R. Schapire, "Performance guarantees for regularized maximum entropy density estimation," in *COLT*, 2004.

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

[5] M. J. Wainwright, P. Ravikumar, and J. D. Lafferty, "High-Dimensional Graphical Model Selection Using ℓ_1-Regularized Logistic Regression," in *NIPS*, 2006, pp. 1465–1472.

[6] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. Springer, 1998.

[7] S. Borade and L. Zheng, "Euclidean Information Theory," in *Allerton Conference*, 2007.

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

[9] T. Cormen, C. Leiserson, R. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd ed. McGraw-Hill Science/Engineering/Math, 2003.

[10] J. N. Laneman, "On the Distribution of Mutual Information," in *Information Theory and Applications Workshop*, 2006.