Scaling laws for learning high-dimensional Markov forest distributions

Scaling Laws for Learning High-Dimensional Markov Forest Distributions

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

* Stochastic Systems Group, LIDS, MIT, Cambridge, MA 02139, Email: {vtan,willsky}@mit.edu
† Center for Pervasive Communications and Computing, UC Irvine, Email: a.anandkumar@uci.edu

***Abstract*—The problem of learning forest-structured discrete graphical models from i.i.d. samples is considered. An algorithm based on pruning of the Chow-Liu tree through adaptive thresholding is proposed. It is shown that this algorithm is structurally consistent and the error probability of structure learning decays faster than any polynomial in the number of samples under fixed model size. For the high-dimensional scenario where the size of the model d and the number of edges k scale with the number of samples n, sufficient conditions on (n, d, k) are given for the algorithm to be structurally consistent. In addition, the extremal structures for learning are identified; we prove that the independent (resp. tree) model is the hardest (resp. easiest) to learn using the proposed algorithm in terms of error rates for structure learning.**

***Index Terms*—Graphical models, Forest distributions, Structural consistency, Method of types.**

I. Introduction

Graphical models (also known as Markov random fields) have a wide range of applications in diverse fields such as signal processing, coding theory and bioinformatics. See [1], [2] and references therein for examples. Inferring the structure and parameters of graphical models from samples is a starting point in all these applications. The structure of the model provides a quantitative interpretation of relationships amongst the given collection of random variables by specifying a set of conditional independence relationships. The parameters of the model quantify the strength of these interactions among the variables.

The challenge in learning graphical models is often compounded by the fact that typically only a small number of samples are available relative to the size of the model (dimension of data). This is referred to as the high-dimensional learning regime, which differs from classical statistics where a large number of samples of fixed dimensionality are available. As a concrete example, in order to analyze the effect of environmental and genetic factors on childhood asthma, clinician scientists in Manchester, UK have been conducting a longitudinal birth-cohort study since 1997 [3], [4]. The number of variables collected is of the order of $d \approx 10^6$ (dominated by the genetic data) but the number of children in the study is small ($n \approx 10^3$). The paucity of subjects in the study is due in part to the prohibitive cost of collecting high-quality clinical data from willing participants.

In order to learn high-dimensional graphical models, it is imperative to find the right balance between data fidelity and overfitting. To ameliorate the effect of overfitting, the samples are often fitted to a *sparse graphical model* [2], with a small number of edges. One popular and tractable class of sparse graphical models is the set of tree[1] models. When restricted to trees, the Chow-Liu algorithm [5], [6] provides an efficient implementation of the maximum-likelihood (ML) procedure to learn the structure from independent samples. However, in the high-dimensional regime, even a tree may overfit the data [7]. In this paper, we consider learning high-dimensional, forest-structured (discrete) graphical models from a given set of samples.

For learning the forest structure, the ML (Chow-Liu) algorithm does not produce a consistent estimate since ML favors richer model classes and hence, outputs a tree in general. We propose a consistent algorithm called CLThres, which has a thresholding mechanism to prune "weak" edges from the Chow-Liu tree. We provide tight bounds on the *overestimation* and *underestimation* errors, that is, the error probability that the output of the algorithm has more or fewer edges than the true model.

A. Main Contributions

We first prove that CLThres is structurally consistent, i.e., as the number of samples grows for a fixed model size, the probability of learning the incorrect structure (set of edges), decays to zero for a fixed model size. We show that the error rate is in fact, dominated by the rate of decay of the overestimation error probability.[2] We use an information-theoretic technique known as the *method of types* [8, Ch. 11] as well as a recently-developed technique known as Euclidean information theory [9]. We provide an upper bound on the error probability by using convex duality to find a surprising connection between the overestimation error rate and a semidefinite program [10] and show that the overestimation error in structure learning decays faster than any polynomial in n for a fixed data dimension d.

This work was supported by a AFOSR funded through Grant FA9559-08-1-1080, a MURI funded through ARO Grant W911NF-06-1-0076 and a MURI funded through AFOSR Grant FA9550-06-1-0324. V. Tan is also funded by A*STAR, Singapore. The full version of this abridged paper can be found at the arXiv link http://arxiv.org/abs/1005.0766.

[1] A *tree* is a *connected*, acyclic graph. We use the term *proper forest* to denote the set of *disconnected*, acyclic graphs.

[2] The overestimation error probability is the probability that the number of edges learned exceeds the true number of edges. The underestimation error is defined analogously.

We then consider the high-dimensional scenario and provide sufficient conditions on the growth of (n, d) (and also the true number of edges k) to ensure that CLThres is structurally consistent. We prove that even if d grows faster than any polynomial in n (in fact close to exponential in n), structure estimation remains consistent. We also show that independent models (resp. tree models) are the "hardest" (resp. "easiest") to learn in the sense that the asymptotic error rate is the highest (resp. lowest), over all models with the same scaling of (n, d). Thus, the empty graph and connected trees are the extremal forest structures for learning.

B. Related Work

There are many papers that discuss learning graphical models from data. See [11]–[15], and references therein. Most of these methods pose the learning problem as a parameterized optimization problem, typically with a regularization term to enforce sparsity in the resulting graph. Consistency guarantees in terms of n and d (and possibly the maximum degree) are provided. Information-theoretic limits for learning graphical models have also be derived in [16]. In Zuk et al. [17], bounds on the error rate for learning the structure of Bayesian networks using the Bayesian Information Criterion (BIC) were provided. Bach and Jordan in [18] learned tree-structured models for solving the independent component analysis (ICA) problem. A PAC analysis for learning thin junction trees was given in [19].

By using the theory of large-deviations [20], we derived and analyzed the error exponent for learning trees for discrete [21] and Gaussian [22] graphical models. The error exponent is a quantitative measure of performance of the learning algorithm since a larger exponent implies a faster decay of the error probability. However, the analysis does not readily extend to learning forest models. In addition, we posed the structure learning problem for trees as a composite hypothesis testing problem [23] and derived a closed-form expression for the Chernoff-Stein exponent in terms of the mutual information on the bottleneck edge. In two recent works which are most closely related to ours, Liu et al. [7] and Gupta et al. [24] derived consistency (and sparsistency) guarantees for learning tree and forest models. The pairwise joint distributions are modeled using kernel density estimates, where the kernels are Hölder continuous. This differs from our approach since we assume that each variable can only take finitely many values, leading to stronger results on error rates for structure learning via the method of types, a powerful proof technique in information theory. Furthermore, the algorithm suggested in both papers uses a subset (usually half) of the dataset to learn the full tree model and then uses the remaining subset to prune the model based on the log-likelihood on the held-out set. We suggest a more direct and consistent method based on thresholding, which uses the *entire* dataset to learn and prune the model without recourse to validation on a held-out dataset. It is well known that validation is both computationally expensive [25, pp. 33] and a potential waste of valuable data which may otherwise be employed to learn a better model. In [24], the problem of estimating forests with restricted component sizes was considered and was proven to be NP-hard. We do not restrict the component size in this paper but instead attempt to learn the model with the minimum number of edges which best fits the data.

Our work is also related to and inspired by the body of literature in information theory on Markov order estimation. In these works, the authors use various regularization and model selection schemes to find the optimal order of a Markov chain [26]–[28], hidden Markov model [29] or exponential family [30]. We build on some of these ideas and proof techniques to identify the correct set of edges (and in particular the number of edges) in the forest model and also to provide strong theoretical guarantees of the rate of convergence of the estimated forest-structured distribution to the true one.

Because of space constraints, all the proofs of the results in this short paper can be found in the journal version at following arXiv link: http://arxiv.org/abs/1005.0766.

II. Preliminaries and Problem Formulation

Let $G = (V, E)$ be an undirected graph with vertex (or node) set $V := \{1, \ldots, d\}$ and edge set $E \subset \binom{V}{2}$ and let $\mathrm{nbd}(i) := \{j \in V : (i, j) \in E\}$ be the set of neighbors of vertex i. Let the set of labeled *trees* (connected, acyclic graphs) with d nodes be $\mathcal{T}^d$ and let the set of *forests* (acyclic graphs) with k edges and d nodes be $\mathcal{T}_k^d$ for $0 \leq k \leq d-1$. The set of forests includes all the trees. We reserve the term *proper forests* for the set of disconnected acylic graphs $\cup_{k=0}^{d-2} \mathcal{T}_k^d$. We also use the notation $\mathcal{F}^d := \cup_{k=0}^{d-1} \mathcal{T}_k^d$ to denote the set of labeled forests with d nodes.

A *graphical model* [1] is a family of multivariate probability distributions (probability mass functions) in which each distribution factorizes according to a given undirected graph and where each variable is associated to a node in the graph. Let $\mathcal{X} = \{1, \ldots, r\}$ (where $2 \leq r < \infty$) be a finite set and $\mathcal{X}^d$ the d-fold Cartesian product of the set $\mathcal{X}$. As usual, let $\mathcal{P}(\mathcal{X}^d)$ denote the probability simplex over the alphabet $\mathcal{X}^d$. We say that the random vector $\mathbf{X} = (X_1, \ldots, X_d)$ with distribution $Q \in \mathcal{P}(\mathcal{X}^d)$ is *Markov on* the graph $G = (V, E)$ if

$$Q(x_i | x_{\mathrm{nbd}(i)}) = Q(x_i | x_{V \setminus i}), \qquad \forall\, i \in V, \tag{1}$$

where $x_{V \setminus i}$ is the collection of variables excluding variable i. Eq. (1) is known as the *local Markov property* [1]. In this paper, we always assume that graphs are *minimal representations* for the corresponding graphical model, i.e., if Q is Markov on G, then G has the smallest number of edges for the conditional independence relations in (1) to hold. We say the distribution Q is a *forest-structured distribution* if it is Markov on a forest. We also use the notation $\mathcal{D}(\mathcal{T}_k^d) \subset \mathcal{P}(\mathcal{X}^d)$ to denote the set of d-variate distributions Markov on a forest with k edges. Similarly, $\mathcal{D}(\mathcal{F}^d)$ is the set of forest-structured distributions.

Let $P \in \mathcal{D}(\mathcal{T}_k^d)$ be a discrete forest-structured distribution Markov on $T_P = (V, E_P) \in \mathcal{T}_k^d$ (for some $k = 0, \ldots, d-1$). It is known that the joint distribution P factorizes as follows [1]:

$$P(\mathbf{x}) = \prod_{i \in V} P_i(x_i) \prod_{(i,j) \in E_P} \frac{P_{i,j}(x_i, x_j)}{P_i(x_i) P_j(x_j)}, \tag{2}$$

where $\{P_i\}_{i\in V}$ and $\{P_{i,j}\}_{(i,j)\in E_P}$ are the node and pairwise marginals which are assumed to be positive everywhere.

The mutual information (MI) of two random variables X_i and X_j with joint distribution $P_{i,j}$ is the function $I(\cdot) : \mathcal{P}(\mathcal{X}^2) \to [0,\infty)$ defined as

$$I(P_{i,j}) := \sum_{x_i,x_j\in\mathcal{X}} P_{i,j}(x_i,x_j)\log\frac{P_{i,j}(x_i,x_j)}{P_i(x_i)P_j(x_j)}. \quad (3)$$

This notation for mutual information differs from the usual $I(X_i;X_j)$ used in [8]; we emphasize the dependence of I on the joint distribution $P_{i,j}$. The *minimum mutual information* in the forest, denoted as $I_{\min} := \min_{(i,j)\in E_P} I(P_{i,j})$ will turn out to be a fundamental quantity in the subsequent analysis. Note from our minimality assumption that $I_{\min} > 0$ since all edges in the forest have positive mutual information (none of the edges are degenerate). When we consider the scenario where d grows with n in Section V, we assume that $I_{\min}$ is *uniformly* bounded away from zero.

A. Problem Statement

We now state the basic problem formally. We are given a set of i.i.d. samples, denoted as $\mathbf{x}^n := \{\mathbf{x}_1,\dots,\mathbf{x}_n\}$. Each sample $\mathbf{x}_l = (x_{l,1},\dots,x_{l,d}) \in \mathcal{X}^d$ is drawn independently from $P \in \mathcal{D}(\mathcal{T}_k^d)$ a forest-structured distribution. From these samples, and the prior knowledge that the undirected graph is acyclic (but not necessarily connected), estimate the true set of edges E_P as well as the true distribution P consistently.

III. The Forest Learning Algorithm: CLThres

We now describe our algorithm for estimating the edge set E_P and the distribution P. This algorithm is a modification of the celebrated Chow-Liu algorithm for maximum-likelihood (ML) learning of tree-structured distributions [5]. We call our algorithm CLThres which stands for *Chow-Liu with Thresholding*.

The inputs to the algorithm are the set of samples $\mathbf{x}^n$ and a *regularization* sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$ (to be specified precisely later) that typically decays to zero, i.e., $\lim_{n\to\infty}\varepsilon_n = 0$. The outputs are the estimated edge set, denoted $\widehat{E}_{\widehat{k}_n}$, and the estimated distribution, denoted P^*.

1) Given $\mathbf{x}^n$, calculate the set of *pairwise empirical distributions*[3] (or *pairwise types*) $\{\widehat{P}_{i,j}\}_{i,j\in V}$. This is just a normalized version of the counts of each observed symbol in $\mathcal{X}^2$ and serves as a set of sufficient statistics for the estimation problem. The dependence of $\widehat{P}_{i,j}$ on the samples $\mathbf{x}^n$ is suppressed.
2) Form the set of *empirical mutual information* quantities:
$$I(\widehat{P}_{i,j}) := \sum_{(x_i,x_j)\in\mathcal{X}^2} \widehat{P}_{i,j}(x_i,x_j)\log\frac{\widehat{P}_{i,j}(x_i,x_j)}{\widehat{P}_i(x_i)\widehat{P}_j(x_j)},$$
for $1 \le i,j \le d$. This is a consistent estimator of the true mutual information in (3).
3) Run a max-weight spanning tree (MWST) algorithm [31], [32] to obtain an estimate of the edge set:
$$\widehat{E}_{d-1} := \operatorname*{argmax}_{E:T=(V,E)\in\mathcal{T}^d} \sum_{(i,j)\in E} I(\widehat{P}_{i,j}).$$
Let the estimated edge set be $\widehat{E}_{d-1} := \{\widehat{e}_1,\dots,\widehat{e}_{d-1}\}$ where the edges $\widehat{e}_i$ are sorted according to decreasing empirical mutual information values. We index the edge set by $d-1$ to emphasize that it has $d-1$ edges and hence is connected. We denote the sorted empirical mutual information quantities as $I(\widehat{P}_{\widehat{e}_1}) \ge \dots \ge I(\widehat{P}_{\widehat{e}_{d-1}})$. These first three steps constitute the Chow-Liu algorithm [5].
4) Estimate the true number of edges using the *thresholding estimator*:
$$\widehat{k}_n := \operatorname*{argmin}_{1\le j\le d-1}\Big\{I(\widehat{P}_{\widehat{e}_j}) : I(\widehat{P}_{\widehat{e}_j}) \ge \varepsilon_n, I(\widehat{P}_{\widehat{e}_{j+1}}) \le \varepsilon_n\Big\}. \quad (4)$$
If there exists an empirical mutual information $I(\widehat{P}_{\widehat{e}_j})$ such that $I(\widehat{P}_{\widehat{e}_j}) = \varepsilon_n$, break the tie arbitrarily.[4]
5) Prune the tree by retaining only the top $\widehat{k}_n$ edges, i.e., define the *estimated edge set* of the forest to be
$$\widehat{E}_{\widehat{k}_n} := \{\widehat{e}_1,\dots,\widehat{e}_{\widehat{k}_n}\},$$
where $\{\widehat{e}_i : 1 \le i \le d-1\}$ is the ordered edge set defined in Step 3. Define the estimated tree to be $\widehat{T}_{\widehat{k}_n} := (V, \widehat{E}_{\widehat{k}_n})$.
6) Finally, define the estimated distribution P^* to be the *reverse I-projection* [33] of the joint type $\widehat{P}$ onto $\widehat{T}_{\widehat{k}_n}$, i.e.,
$$P^*(\mathbf{x}) := \operatorname*{argmin}_{Q\in\mathcal{D}(\widehat{T}_{\widehat{k}_n})} D(\widehat{P}\,||\,Q).$$
It can easily be shown that the projection can be expressed in terms of the marginal and pairwise joint types:
$$P^*(\mathbf{x}) = \prod_{i\in V}\widehat{P}_i(x_i)\prod_{(i,j)\in\widehat{E}_{\widehat{k}_n}}\frac{\widehat{P}_{i,j}(x_i,x_j)}{\widehat{P}_i(x_i)\widehat{P}_j(x_j)}.$$

Intuitively, CLThres first constructs a connected tree $(V,\widehat{E}_{d-1})$ via Chow-Liu (in Steps 1 – 3) before pruning the weak edges (with small mutual information) to obtain the final structure $\widehat{E}_{\widehat{k}_n}$. The estimated distribution P^* is simply the ML estimate of the parameters subject to the constraint that P^* is Markov on the learned tree $\widehat{T}_{\widehat{k}_n}$.

Note that if Step 4 is omitted and $\widehat{k}_n$ is defined to be $d-1$, then CLThres simply reduces to the Chow-Liu ML algorithm. Of course Chow-Liu, which outputs a tree, is guaranteed to fail (not be structurally consistent) if the number of edges in the true model $k < d-1$, which is the problem of interest in this paper. Thus, Step 4, a model selection step, is essential in estimating the true number of edges k.

[3] In this paper, the terms *empirical distribution* and *type* are used interchangeably.

[4] Here were allow a bit of imprecision by noting that the non-strict inequalities in (4) simplify the subsequent analyses because the constraint sets that appear in optimization problems will be closed, hence compact, insuring the existence of optimizers.

This step is a generalization of the test for independence of discrete memoryless sources discussed in [30]. In our work, we exploit the fact that the empirical mutual information $I(\widehat{P}_{\widehat{e}_j})$ corresponding to a pair of independent variables $\widehat{e}_j$ will be very small when n is large, thus a thresholding procedure using the (appropriately chosen) regularization sequence $\{\varepsilon_n\}$ will remove these edges. In fact, the subsequent analysis allows us to conclude that Step 4, in a formal sense, *dominates* the error probability in structure learning. CLThres is also efficient as shown by the following result.

Proposition 1 (Complexity of CLThres*):* CLThres runs in time $O((n+\log d)d^2)$.

IV. Structural Consistency For Fixed Model Size

In this section, we keep d and k fixed and consider a probability model P, which is assumed to be Markov on a forest in $\mathcal{T}_k^d$. This is to gain better insight into the problem before we analyze the high-dimensional scenario in Section V where d and k scale[5] with the sample size n. More precisely, we are interested in quantifying the rate at which the probability of the error event of structure learning

$$\mathcal{A}_n := \Big\{\mathbf{x}^n \in (\mathcal{X}^d)^n : \widehat{E}_{\widehat{k}_n} \neq E_P\Big\} \tag{5}$$

decays to zero as n tends to infinity. Recall that $\widehat{E}_{\widehat{k}_n}$, with cardinality $\widehat{k}_n$, is the learned edge set by using CLThres. As usual, P^n is the n-fold product probability measure corresponding to the forest-structured distribution P.

Before stating the main result of this section in Theorem 3, we first state an auxiliary result that essentially says that if one is provided with oracle knowledge of $I_{\min}$, the minimum mutual information in the forest, then the problem is greatly simplified.

Proposition 2 (Error Rate with knowledge of $I_{\min}$): Assume that $I_{\min}$ is known in CLThres. Then by letting the regularization sequence be $\varepsilon_n = I_{\min}/2$ for all n, we have

$$\lim_{n\to\infty} \frac{1}{n}\log P^n(\mathcal{A}_n) < 0, \tag{6}$$

i.e., the error probability decays exponentially fast.

Thus, the primary difficulty lies in estimating $I_{\min}$ or equivalently, the number of edges k. Note that if k is known, a simple modification to the Chow-Liu procedure by imposing the constraint that the final structure contains k edges will also yield exponential decay as in (6). However, in the realistic case where both $I_{\min}$ and k are unknown, we show in the rest of this section that we can design the regularization sequence ε_n in such a way that the rate of decay of $P^n(\mathcal{A}_n)$ decays almost exponentially fast.

A. Error Rate for Forest Structure Learning

We now state one of the main results in this paper. We emphasize that the following result is stated for a fixed forest-structured distribution $P \in \mathcal{D}(\mathcal{T}_k^d)$ so d and k are also fixed natural numbers.

[5]In that case P must also scale, i.e., we learn a *family* of models as d and k scale.

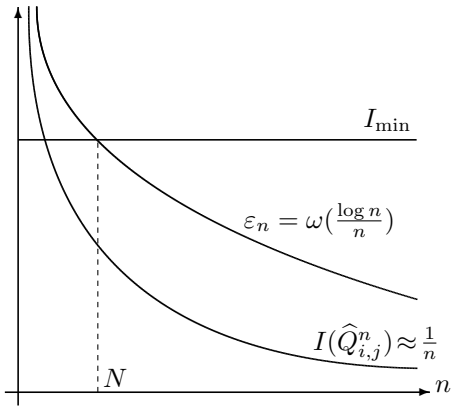

Fig. 1. Graphical interpretation of the condition on ε_n. As $n \to \infty$, the regularization sequence ε_n will be smaller than $I_{\min}$ and larger than $I(\widehat{Q}^n_{i,j})$ with high probability.

Theorem 3 (Error Rate for Structure Learning): Assume that the regularization sequence $\{\varepsilon_n\}_{n\in\mathbb{N}}$ satisfies the following two conditions:

$$\lim_{n\to\infty} \varepsilon_n = 0, \qquad \lim_{n\to\infty} \frac{n\varepsilon_n}{\log n} = \infty. \tag{7}$$

Then, if the true model $T_P = (V, E_P)$ is a proper forest ($k < d-1$), there exists a constant $C_P \in (1,\infty)$ such that

$$-C_P \leq \liminf_{n\to\infty} \frac{1}{n\varepsilon_n}\log P^n(\mathcal{A}_n) \tag{8}$$

$$\leq \limsup_{n\to\infty} \frac{1}{n\varepsilon_n}\log P^n(\mathcal{A}_n) \leq -1. \tag{9}$$

Finally, if the true model $T_P = (V, E_P)$ is a tree ($k = d-1$), then

$$\lim_{n\to\infty} \frac{1}{n}\log P^n(\mathcal{A}_n) < 0, \tag{10}$$

i.e., the error probability decays exponentially fast.

B. Interpretation of Result

From (9), the rate of decay of the error probability for proper forests is subexponential but nonetheless can be made faster than any polynomial for an appropriate choice of ε_n. The reason for the subexponential rate is because of our lack of knowledge of $I_{\min}$, the minimum mutual information in the true forest T_P. For trees, the rate[6] is exponential ($\doteq \exp(-nF)$ for some positive constant F). Learning proper forests is thus, strictly "harder" than learning trees. The condition on ε_n in (7) is needed for the following intuitive reasons:

1) Firstly, (7) ensures that for all sufficiently large n, we have $\varepsilon_n < I_{\min}$. Thus, the true edges will be correctly identified by CLThres implying that with high probability, there will not be underestimation as $n \to \infty$.

[6]We use the asymptotic notation from information theory $\doteq$ to denote equality to first order in the exponent. More precisely, for two positive sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ we say that $a_n \doteq b_n$ iff $\lim_{n\to\infty} n^{-1}\log(a_n/b_n) = 0$.

2) Secondly, for two independent random variables X_i and X_j with distribution $Q_{i,j} = Q_i Q_j$, the sequence[7] $\sigma(I(\widehat{Q}^n_{i,j})) = \Theta(1/n)$, where $\widehat{Q}^n_{i,j}$ is the joint empirical distribution of n i.i.d. samples drawn from $Q_{i,j}$. Since the regularization sequence $\varepsilon_n = \omega(\log n/n)$ has a slower rate of decay than $\sigma(I(\widehat{Q}^n_{i,j}))$, $\varepsilon_n > I(\widehat{Q}^n_{i,j})$ with high probability as $n \to \infty$. Thus, with high probability there will not be overestimation as $n \to \infty$.

See Figure 1 for an illustration of this intuition. The formal proof follows from a method of types argument and we provide an outline in Section IV-C. A convenient choice of ε_n that satisfies (7) is

$$\varepsilon_n := n^{-\beta}, \qquad \forall\, \beta \in (0,1). \tag{11}$$

Note further that the upper bound in (9) is also independent of P since it is equal to -1 for all P. Thus, (9) is a *universal* result for all forest distributions $P \in \mathcal{D}(\mathcal{F}^d)$. The intuition for this universality is because in the large-n regime, the typical way an error occurs is due to overestimation. The overestimation error results from testing whether pairs of random variables are independent and our asymptotic bound for the error probability of this test does not depend on the true distribution P.

The lower bound C_P in (8) means that we cannot hope to do much better using CLThres if the original structure (edge set) is a proper forest. Together, (8) and (9) imply that the rate of decay of the error probability for structure learning is tight to within a constant factor in the exponent. We believe that the error rates given in Theorem 3 cannot, in general, be improved without knowledge of $I_{\min}$.

C. Proof Idea

The method of proof for Theorem 3 involves using the Gallager-Fano bounding technique [35, pp. 24] and the union bound to decompose the overall error probability $P^n(\mathcal{A}_n)$ into three distinct terms: (i) the rate of decay of the error probability for learning the top k edges (in terms of the mutual information quantities) correctly – known as the *Chow-Liu error*, (ii) the rate of decay of the *overestimation error* $\{\widehat{k}_n > k\}$ and (iii) the rate of decay of the *underestimation error* $\{\widehat{k}_n < k\}$. Each of these terms is upper bounded using a method of types [8, Ch. 11] argument. It turns out, as is the case with the literature on Markov order estimation (e.g., [27]), that bounding the overestimation error poses the greatest challenge. Indeed, we show that the underestimation and Chow-Liu errors have exponential decay in n. However, the overestimation error has subexponential decay ($\approx \exp(-n\varepsilon_n)$).

The main technique used to analyze the overestimation error relies on *Euclidean information theory* [9] which states that if two distributions ν_0 and ν_1 (both supported on a common finite alphabet $\mathcal{Y}$) are close entry-wise, then various information-theoretic measures can be approximated locally by quantities related to Euclidean norms. For example, the KL-divergence $D(\nu_0 \,||\, \nu_1)$ can be approximated by the square of a weighted Euclidean norm:

$$D(\nu_0 \,||\, \nu_1) = \frac{1}{2} \sum_{a \in \mathcal{Y}} \frac{(\nu_0(a) - \nu_1(a))^2}{\nu_0(a)} + o(\|\nu_0 - \nu_1\|^2_\infty). \tag{12}$$

Using this approximation and Lagrangian duality [36], we reduce a non-convex I-projection [33] problem involving information-theoretic quantities (such as divergence) to a relatively simple *semidefinite program* [10] which admits a closed-form solution. Furthermore, the approximation in (12) becomes *exact* as $n \to \infty$ (i.e., $\varepsilon_n \to 0$), which is the asymptotic regime of interest.

D. Error Rate for Learning the Forest Projection

In our discussion thus far, P has been assumed to be Markov on a forest. In this subsection, we consider the situation when the underlying unknown distribution P is not forest-structured but we wish to learn its best forest approximation. To this end, we define the projection of P onto the set of forests (or *forest projection*) to be

$$\widetilde{P} := \operatorname*{argmin}_{Q \in \mathcal{D}(\mathcal{F}^d)} D(P \,||\, Q). \tag{13}$$

If there are multiple optimizing distribution, choose a projection $\widetilde{P}$ that is minimal, i.e., its graph $T_{\widetilde{P}} = (V, E_{\widetilde{P}})$ has the *fewest number of edges* such that (13) holds. If we redefine the event $\mathcal{A}_n$ in (5) to be $\widetilde{\mathcal{A}}_n := \{\widehat{E}_{\widehat{k}_n} \neq E_{\widetilde{P}}\}$, we have the following analogue of Theorem 3.

Corollary 4 (Error Rate for Learning Forest Projection): Let P be an arbitrary distribution and the event $\widetilde{\mathcal{A}}_n$ be defined as above. Then the conclusions in (8) – (10) in Theorem 3 hold if the regularization sequence $\{\varepsilon_n\}_{n \in \mathbb{N}}$ satisfies (7).

V. High-Dimensional Structural Consistency

In the previous section, we considered learning a fixed forest-structured distribution P (and hence fixed d and k) and derived bounds on the error rate for structure learning. However, for most problems of practical interest, the number of data samples is small compared to the data dimension d (see the asthma example in the introduction). In this section, we prove sufficient conditions on the scaling of (n, d, k) for structure learning to remain consistent. We will see that even if d and k are much larger than n, under some reasonable regularity conditions, structure learning remains consistent.

A. Structure Scaling Law

To pose the learning problem formally, we consider a *sequence* of structure learning problems indexed by the number of data points n. For the particular problem indexed by n, we have a dataset $\mathbf{x}^n = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of size n where each sample $\mathbf{x}_l \in \mathcal{X}^d$ is drawn independently from an unknown d-variate forest-structured distribution $P^{(d)} \in \mathcal{D}(\mathcal{T}^d_k)$, which has d nodes and k edges and where d and k depend on n. This *high-dimensional* setup allows us to model and subsequently

[7] The notation $\sigma(Z)$ denotes the standard deviation of the random variable Z. The fact that the standard deviation of the empirical MI $\sigma(I(\widehat{Q}^n_{i,j}))$ decays as $1/n$ can be verified by Taylor expanding $I(\widehat{Q}^n_{i,j})$ around $Q_{i,j} = Q_i Q_j$ and using the fact that the ML estimate converges at a rate of $n^{-1/2}$ [34].

analyze how d and k can scale with n while maintaining consistency. We will sometimes make the dependence of d and k on n explicit, i.e., $d = d_n$ and $k = k_n$.

In order to be able to learn the structure of the models we assume that

$$\text{(A1)} \quad I_{\inf} := \inf_{d\in\mathbb{N}} \min_{(i,j)\in E_{P^{(d)}}} I(P^{(d)}_{i,j}) > 0, \tag{14}$$

$$\text{(A2)} \quad \kappa := \inf_{d\in\mathbb{N}} \min_{x_i,x_j\in\mathcal{X}} P^{(d)}_{i,j}(x_i, x_j) > 0. \tag{15}$$

That is, assumptions (A1) and (A2) insure that there exists *uniform* lower bounds on the minimum mutual information and the minimum entry in the pairwise probabilities in the forest models as the size of the graph grows. These are typical regularity assumptions for the high-dimensional setting. See [13] and [14] for examples. We again emphasize that the proposed learning algorithm CLThres has knowledge of neither $I_{\inf}$ nor κ. Equipped with (A1) and (A2) and assuming the asymptotic behavior of ε_n in (7), we claim the following theorem for CLThres.

Theorem 5 (Structure Scaling Law): There exists two finite, positive constants $C_1 = C_1(I_{\inf}, \kappa)$ and $C_2 = C_2(I_{\inf}, \kappa)$ such that if

$$n > \max\left\{(2\log(d-k))^{1+\zeta},\, C_1 \log d,\, C_2 \log k\right\}, \tag{16}$$

for any $\zeta > 0$, then the error probability of incorrectly learning the sequence of edge sets $\{E_{P^{(d)}}\}_{d\in\mathbb{N}}$ tends to zero as $(n, d, k) \to \infty$. When the sequence of forests are trees, $n > \max\{C_1, C_2\} \log d$ suffices for high-dimensional structure recovery.

Thus, if the model parameters (n, d, k) all grow with n but $d = o(\exp(n/C_1))$, $k = o(\exp(n/C_2))$ and $d - k = o(\exp(n^{1-\beta}/2))$ (for all $\beta > 0$), consistent structure recovery is possible in high dimensions. In other words, the number of nodes d can grow faster than any polynomial in the sample size n. In [7], the bivariate densities are modeled by functions from a Hölder class with exponent α and it was mentioned (in Theorem 4.3) that the number of variables can grow like $o(\exp(n^{\alpha/(1+\alpha)}))$ for structural consistency. Our result is somewhat stronger but we model the pairwise joint distributions as (simpler) probability mass functions (the alphabet $\mathcal{X}$ is a finite set).

B. Extremal Forest Structures

In this subsection, we study the extremal structures for learning, that is, the structures that, roughly speaking, lead to the largest and smallest error probabilities for structure learning. Define the sequence

$$h_n(P) := \frac{1}{n\varepsilon_n} \log P^n(\mathcal{A}_n), \quad \forall\, n \in \mathbb{N}. \tag{17}$$

Note that h_n is a function of both the number of variables $d = d_n$ and the number of edges $k = k_n$ in the models $P^{(d)}$ since it is a sequence indexed by n. In the next result, we assume (n, d, k) satisfies the scaling law in (16) and answer the following question: How does h_n in (17) depend on the number of edges k_n for a given d_n? Let $P_1^{(d)}$ and $P_2^{(d)}$ be two sequences of forest-structured distributions with a common number of nodes d_n and number of edges $k_n(P_1^{(d)})$ and $k_n(P_2^{(d)})$ respectively.

Corollary 6 (Extremal Forests): As $n \to \infty$, $h_n(P_1^{(d)}) \leq h_n(P_2^{(d)})$ whenever $k_n(P_1^{(d)}) \geq k_n(P_2^{(d)})$ implying that h_n is maximized when $P^{(d)}$ are product distributions (i.e., $k_n = 0$) and minimized when $P^{(d)}$ are tree-structured distributions (i.e., $k_n = d_n - 1$). Furthermore, if $k_n(P_1^{(d)}) = k_n(P_2^{(d)})$, then $h_n(P_1^{(d)}) = h_n(P_2^{(d)})$.

The intuition for this result is the following: We recall from the discussion after Theorem 3 that the overestimation error dominates the probability of error for structure learning. Thus, the performance of CLThres degrades with the number of missing edges. If there are very few edges (i.e., k_n is very small relative to d_n), the CLThres estimator is more likely to overestimate the number of edges as compared to if there are many edges (i.e., k_n/d_n is close to 1). We conclude that a distribution which is Markov on an *empty graph* (all variables are independent) is the *hardest* to learn (in the sense of Corollary 6 above). Conversely, *trees* are the *easiest* structures to learn using CLThres.

VI. Conclusion

In this paper, we proposed an efficient algorithm CLThres for learning the parameters and the structure of forest-structured graphical models. We provided error rates for structure learning and scaling laws on the number of samples, the number of variables and the number of edges so that structure learning remains consistent in high-dimensions. In the full version of this paper, we also develop results for risk consistency, i.e., the rate at which the estimated parameters converge to the true ones. There are many open problems that could possibly leverage on the proof techniques employed here. For example, we are currently interested to analyze the learning of general graphical models using similar thresholding-like techniques on the empirical correlation coefficients. The analyses could potentially leverage on the use of the method of types. We are currently exploring this promising line of research.

References

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

[2] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference." Technical Report, University of California, Berkeley, Tech. Rep., 2003.

[3] A. Custovic, B. M. Simpson, C. S. Murray, L. Lowe, and A. Woodcock, "The National Asthma Campaign Manchester Asthma and Allergy Study," *Pediatr Allergy Immunol*, vol. 13, pp. 32–37, 2002.

[4] A. Simpson, V. Y. F. Tan, J. M. Winn, M. Svensén, C. M. Bishop, D. E. Heckerman, I. Buchan, and A. Custovic, "Beyond Atopy: Multiple Patterns of Sensitization in Relation to Asthma in a Birth Cohort Study," *Am J Respir Crit Care Med*, 2010.

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

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

[7] H. Liu, J. Lafferty, and L. Wasserman, "Tree density estimation," *arXiv:1001.1557 [stat.ML]*, Jan 2010.

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, 2006.

[9] S. Borade and L. Zheng, "Euclidean Information Theory," in *IEEE International Zurich Seminar on Communications*, 2008, pp. 14–17.

[10] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Review*, vol. 38, pp. 49–95, Mar 1996.

[11] M. Dudik, S. J. Phillips, and R. E. Schapire, "Performance guarantees for regularized maximum entropy density estimation," in *Conference on Learning Theory (COLT)*, 2004.

[12] P. Abbeel, D. Koller, and A. Y. Ng, "Learning factor graphs in polynomial time and sample complexity," *Journal of Machine Learning Research*, Dec 2006.

[13] M. J. Wainwright, P. Ravikumar, and J. D. Lafferty, "High-Dimensional Graphical Model Selection Using ℓ_1-Regularized Logistic Regression," in *Advances of Neural Information Processing Systems (NIPS)*, 2006, pp. 1465–1472.

[14] 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.

[15] J. Johnson, V. Chandrasekaran, and A. S. Willsky, "Learning Markov Structure by Maximum Entropy Relaxation," in *Artificial Intelligence and Statistics (AISTATS)*, 2007.

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

[17] O. Zuk, S. Margel, and E. Domany, "On the number of samples needed to learn the correct structure of a Bayesian network," in *Proc of Uncertainty in Artificial Intelligence (UAI)*, 2006.

[18] F. Bach and M. I. Jordan, "Beyond independent components: trees and clusters," *Journal of Machine Learning Research*, vol. 4, pp. 1205–1233, 2003.

[19] A. Chechetka and C. Guestrin, "Efficient Principled Learning of Thin Junction Trees," in *Advances of Neural Information Processing Systems (NIPS)*, 2007.

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

[21] V. Y. F. Tan, A. Anandkumar, L. Tong, and A. S. Willsky, "A Large-Deviation Analysis for the Maximum-Likelihood Learning of Markov Tree Structures," *submitted to IEEE Transactions on Infomation Theory, arXiv:0905.0940 [cs.IT]*, May 2009.

[22] V. Y. F. Tan, A. Anandkumar, and A. S. Willsky, "Learning Gaussian Tree Models: Analysis of Error Exponents and Extremal Structures," *IEEE Transactions on Signal Processing*, vol. 58, no. 5, pp. 2701 – 2714, May 2010.

[23] ——, "Error Exponents for Composite Hypothesis Testing of Markov Forest Distributions," in *Proc. of Intl. Symp. on Info. Th.*, June 2010.

[24] A. Gupta, J. Lafferty, H. Liu, L. Wasserman, and M. Xu, "Forest density estimation," in *Info. Th. and Applications (ITA) Workshop*, San Diego, CA, 2010.

[25] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2008.

[26] N. Merhav, M. Gutman., and J. Ziv, "On the estimation of the order of a Markov chain and universal data compression," *IEEE Transactions on Infomation Theory*, vol. 35, pp. 1014–1019, 1989.

[27] L. Finesso, C. C. Liu, and P. Narayan, "The Optimal Error Exponent for Markov Order Estimation," *IEEE Trans. on Info Th.*, vol. 42, no. 5, pp. 1488–1497, 1996.

[28] I. Csiszár and P. Shields, "The consistency of the BIC Markov order estimator," *Ann. Statist.*, vol. 28, no. 6, pp. 1601–1619, 2000.

[29] E. Gassiat and S. Boucheron, "Optimal Error Exponents in Hidden Markov Models Order Estimation," *IEEE Transactions on Infomation Theory*, vol. 49, no. 4, pp. 964–980, Apr 2003.

[30] N. Merhav, "The Estimation of the Model Order in Exponential Families," *IEEE Transactions on Infomation Theory*, vol. 35, no. 5, pp. 1109–1115, 1989.

[31] R. C. Prim, "Shortest connection networks and some generalizations," *Bell System Technical Journal*, vol. 36, 1957.

[32] J. B. Kruskal, "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem," *Proceedings of the American Mathematical Society*, vol. 7, no. 1, Feb 1956.

[33] I. Csiszár and F. Matúš, "Information projections revisited," *IEEE Transactions on Infomation Theory*, vol. 49, no. 6, pp. 1474–1490, 2003.

[34] R. J. Serfling, *Approximation Theorems of Mathematical Statistics*. Wiley-Interscience, Nov 1980.

[35] R. M. Fano, *Transmission of Information*. New York: Wiley, 1961.

[36] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.