Near-Optimal Learning of Tree-Structured Distributions by Chow-Liu

Arnab Bhattacharyya
National University of Singapore
arnabb@nus.edu.sg

Sutanu Gayen
National University of Singapore
sutanugayen@gmail.com

Eric Price
University of Texas at Austin
ecprice@cs.utexas.edu

N. V. Vinodchandran
University of Nebraska-Lincoln
vinod@cse.unl.edu

February 24, 2021

Abstract

We provide finite sample guarantees for the classical Chow-Liu algorithm (IEEE Trans. Inform. Theory, 1968) to learn a tree-structured graphical model of a distribution. For a distribution P on Σ^n and a tree T on n nodes, we say T is an ε-approximate tree for P if there is a T-structured distribution Q such that $D(P \parallel Q)$ is at most ε more than the best possible tree-structured distribution for P. We show that if P itself is tree-structured, then the Chow-Liu algorithm with the plug-in estimator for mutual information with $\widetilde{O}(|\Sigma|^3 n\varepsilon^{-1})$ i.i.d. samples outputs an ε-approximate tree for P with constant probability. In contrast, for a general P (which may not be tree-structured), $\Omega(n^2\varepsilon^{-2})$ samples are necessary to find an ε-approximate tree. Our upper bound is based on a new conditional independence tester that addresses an open problem posed by Canonne, Diakonikolas, Kane, and Stewart (STOC, 2018): we prove that for three random variables X, Y, Z each over Σ, testing if $I(X;Y \mid Z)$ is 0 or $\geq \varepsilon$ is possible with $\widetilde{O}(|\Sigma|^3/\varepsilon)$ samples. Finally, we show that for a specific tree T, with $\widetilde{O}(|\Sigma|^2 n\varepsilon^{-1})$ samples from a distribution P over Σ^n, one can efficiently learn the closest T-structured distribution in KL divergence by applying the add-1 estimator at each node.

1 Introduction

Probabilistic graphical models form a highly effective framework for encoding high-dimensional distributions. Graphical models yield human-interpretable representation of data as they explicitly describe the statistical dependencies among different features. From a computational standpoint, the graphical representation enables efficient algorithms for inference, e.g., message passing, loopy belief propagation, and other variational inference methods [KFL01]. Graphical models have found extensive applications in many domains, such as image processing, natural language processing and computational biology; see [Lau96, KF09, WJ08] and the references therein for examples.

A fundamental question in this area is to learn graphical models from independently drawn samples. In this paper, we focus on the basic problem of learning *tree-structured distributions*. Given a tree T on n nodes, fix an arbitrary root and orient it outwards. A distribution P over variables $X_1, \ldots, X_n$ is said to be *T-structured* iff for every non-root vertex i:

$$X_i = f_i(X_{\mathrm{pa}(i)}, U_i)$$

where $\mathrm{pa}(i)$ is the parent of i in the oriented tree, U_i is an independent random variable, and f_i is a (deterministic) function. A distribution is tree-structured if it is T-rooted for some tree T. Equivalently, a tree-structured distribution is a Markov random field where the underlying undirected graph is a tree.

In a seminal work [CL68], Chow and Liu observed that the tree-structured distribution maximizing the likelihood of the observed samples can be obtained by solving a maximum weight spanning tree problem. In particular, their algorithm assigns a weight equal to the empirical mutual information between each pair of variables and finds a maximum weight spanning tree in this weighted graph. The resulting tree can be oriented from an arbitrary root, so as to assign a parent $\mathrm{pa}(i)$ for all non-root vertices i. Finally, the conditional probability distributions $X_i \mid X_{\mathrm{pa}(i)}$ can be learned from the data.

Chow and Wagner [CW73] showed that the Chow-Liu algorithm consistently recovers structure, meaning that if the samples are generated by a T-structured distribution for a tree T, then it recovers T with probability approaching 1 as the number of samples tends to infinity. More recent works [TATW11, TTZ20] have used large-deviation theory to study the error exponent K_P of T-structured distributions P, where:

$$K_P = \lim_{N\to\infty} -\frac{1}{N} \log \mathbf{Pr}[\widehat{T} \neq T].$$

Here, $\widehat{T}$ is the tree output by the Chow-Liu algorithm from N samples. The bounds they obtain depend on the distribution structure, since it may be very hard to distinguish T from an alternative tree that is almost as good, and this dependence is only explicit in the special case of homogenous binary distributions.

In this work, we take a different viewpoint that is in the spirit of distribution learning and probabilistically approximately correct (PAC) analysis [Val84, KSS94, KS94]. Instead of trying to exactly recover the structure of a tree-structured distribution P, we consider the objective of *learning a tree-structured distribution Q that is close to P*. For many downstream tasks, most notably statistical inference, it is fine to not recover the exact structure as long as one can approximate probabilities of relevant events using the learned distribution. Also, this viewpoint allows us to analyze Chow-Liu for non-tree structured distributions P, by comparing how far P is from the output of Chow-Liu and how far from the closest tree-structured distribution.

More formally, for a distribution P over Σ^n and a tree T on n vertices, let:

$$P_T := \operatorname*{arg\,min}_{\substack{T\text{-structured} \\ \text{distribution } Q}} D(P \parallel Q)$$

where $D(\cdot \parallel \cdot)$ denotes the KL-divergence. We say that a tree $\widehat{T}$ is an *ε-approximate tree* for P if:

$$D(P \parallel P_{\widehat{T}}) \leq \min_{\text{tree } T} D(P \parallel P_T) + \varepsilon.$$

The KL divergence, although not a metric, is a useful notion of distance to consider in this setting. Firstly, with infinite samples, Chow-Liu's output maximizes the likelihood of generating samples from P and hence, minimizes $D(P \parallel \cdot)$. Secondly, via Pinsker's inequality, bounding the KL divergence by ε implies a $\sqrt{2\varepsilon}$ bound on total variation distance which may be more directly useful.

1.1 Our Contributions

We study the number of samples required by Chow-Liu to output an ε-approximate tree with a fixed error probability. We first observe that for any distribution P, it can be guaranteed that the output of Chow-Liu is ε-approximate if each mutual information estimate is an additive $\pm\frac{\varepsilon}{2n}$ estimate. Known bounds for the plug-in entropy estimator imply the following sample complexity.

Lemma 1.1. *The* Chow-Liu *algorithm when run on* $\widetilde{O}\left(\frac{|\Sigma|^2 n}{\varepsilon} + \frac{n^2}{\varepsilon^2}\log\frac{1}{\delta}\right)$ *samples from a distribution* P *on* Σ^n *outputs an* ε*-approximate tree* T *with probability at least* $1-\delta$*. Moreover, the dependence of the sample complexity on* n *and* ε *are tight up to logarithmic factors.*

We show that the quadratic dependence on n and ε is unfortunately necessary for general distributions P. However, in the "realizable" setting where P is actually tree-structured, we show that the sample complexity can be improved to near-linear:

Theorem 1.2. *The* Chow-Liu *algorithm when run on* $\widetilde{O}(\frac{|\Sigma|^3 n}{\varepsilon}\log\frac{1}{\delta})$ *samples from a tree-structured distribution* P *on* Σ^n *outputs an* ε*-approximate tree* T *with probability at least* $1-\delta$*. Moreover, the dependence on* n *and* ε *are tight up to logarithmic factors.*

Hence, for example, for tree-structured Ising models (where $\Sigma = \{\pm 1\}$), there is a provable near-quadratic gap in the sample complexity for realizable versus non-realizable input distributions. Note that with $O(n/\varepsilon)$ samples, we do not get accurate estimates of the mutual information edge weights. However, as we explain in Section 2, the errors for the edge weights are not independent; in fact, the errors are correlated in such a way that Chow-Liu still recovers an approximate tree! We note that our $\Omega(n/\varepsilon)$-sample complexity lower bound is specifically for *recovering the structure* of the unknown tree. Daskalakis, Dikkala, and Kamath [DDK19] have shown the same lower bound for learning the distribution, but learning the tree might have been easier.

Our main tool for proving Theorem 1.2 is a result on testing conditional independence using the plug-in conditional mutual information estimator. We show that $\widetilde{O}(|\Sigma|^3/\varepsilon)$ samples suffice to distinguish $I(X;Y \mid Z) = 0$ from $I(X;Y \mid Z) \geq \varepsilon$ with constant probability. In more detail:

Theorem 1.3 (Conditional Mutual Information Tester)**.** *Let* (X,Y,Z) *be three random variables over* Σ*, and* $(\widehat{X},\widehat{Y},\widehat{Z})$ *be the empirical distribution over a size* N *sample of* (X,Y,Z)*. There exists a universal constant* $0 < C < 1$ *so that for any*

$$N \geq \Theta\left(\frac{|\Sigma|^3}{\varepsilon}\log\frac{|\Sigma|}{\delta}\log\frac{|\Sigma|\log(1/\delta)}{\varepsilon}\right),$$

the following results hold with probability $1-\delta$*:*

1. *If* $I(X;Y \mid Z) = 0$*, then* $I(\widehat{X};\widehat{Y} \mid \widehat{Z}) < \varepsilon$*.*
2. *If* $I(X;Y \mid Z) \geq \varepsilon$*, then* $I(\widehat{X};\widehat{Y} \mid \widehat{Z}) > C \cdot I(X;Y \mid Z)$*.*

We also get a similar result for unconditional mutual information testing (testing if $I(X;Y) = 0$ or $I(X;Y) \geq \varepsilon$) with a $|\Sigma|$ factor smaller N. Conditional independence testing has previously been studied in [CDKS18], which gave optimal bounds for testing whether (X,Y,Z) is conditionally independent or ε-far from independent in *total variation distance*. Developing a (conditional) independence tester with respect to *mutual information* with $o(\frac{1}{\varepsilon^2})$ sample complexity was posed as an open problem in [CDKS18]; Theorem 1.3 resolves this with optimal ε dependence. Moreover, the test statistic used by Theorem 1.3 is simply the empirical mutual information, which is key for our application to Chow-Liu.

Theorem 1.2 describes how Chow-Liu finds a good tree T. Our final result shows how to estimate the nearest T-structured distribution for fixed T. As above, the spirit of our approach is to make the algorithms as simple as possible (moving possible complications to the analysis). For the fixed-structure learning problem, the most natural approach is to empirically estimate $X_i \mid X_{\mathrm{pa}(i)} = x$ for each non-root node i and for each setting x of the parent of i. However, for KL divergence, the empirical estimator is known to not work; so, we move to the next most natural estimator: Laplace's add-1 estimator [Lap95].

Theorem 1.4. *Let P be a discrete distribution over Σ^n. Let T be a tree on n vertices, and Q be a T-structured distribution with conditional probabilities at each node estimated using the empirical add-1 estimator on*

$$N = \Theta\left(\frac{n|\Sigma|^2}{\varepsilon}\log\frac{n|\Sigma|}{\delta}\log\left(\frac{n|\Sigma|}{\varepsilon}\log\frac{1}{\delta}\right)\right)$$

samples from P. Then

$$D(P \parallel Q) - D(P \parallel P_T) \leq \varepsilon$$

with probability $1-\delta$.

The result actually holds for arbitrary Bayes net models, not just trees. The sample complexity becomes $\widetilde{O}(n|\Sigma|^{d+1}/\varepsilon)$ for Bayes nets with in-degree at most d. To the best of our knowledge, this is the first efficient algorithm with this guarantee. Combining Theorem 1.4 with Theorem 1.2 shows that, for any tree-structured distribution P, after $\widetilde{O}(|\Sigma|^3 n/\varepsilon)$ samples, we can properly learn a tree-structured distribution Q satisfying $D(P \parallel Q) \leq \varepsilon$ (and hence $\|P-Q\|_{TV} \leq \sqrt{2\varepsilon}$).

1.2 Related Work

Learning a multivariate distribution from samples is an important problem in machine learning and statistics with many applications. The problem is provably intractable for general high-dimensional multivariate distributions (e.g., [KOPS15]). Thus, structural assumptions need to be made for designing efficient and practical learning algorithms for high-dimensional distributions. Graphical models including Bayesian networks and Markov Random Fields (MRFs) are widely popular natural classes of structured distributions. In this setup, the learning problem naturally decomposes into two subproblems: *structure learning* and *parameter learning*.

For structure learning, the goal is to output the best structure (eg: a Bayes net that maximizes the likelihood of the data), given independent samples. Unfortunately, in general for both Bayes nets and MRFs, finding the best structure is known to be NP-hard [Chi95, DL97, KS01, Mee01]. In this context, Chow-Liu algorithm remains one of the few efficient structure learning algorithms that does not require any additional assumptions. Since its publication, researchers have continued to look into analyzing properties of this algorithm [CW73, Mei99, TATW11, TTZ20] and generalizing it to other classes of graphs, e.g., polytrees [Das13], bounded treewidth graphs [Sre03, NB04], and mixtures of trees [MJ00, AHHK12]. Most of these works focus on establishing conditions guaranteeing that the algorithm recovers the *exact* tree structure in the limit that the number of samples tends to infinity. Also, for general graph-structured Ising and Markov random fields, several algorithms [BMS13, WSN13, Bre15, KM17, WSD19, Goe20] have been proposed that recover the graphical structure under various distributional assumptions.

As mentioned in the introduction, a common motivation for structure learning is to subsequently use the structure for inference algorithms. For such applications, instead of recovering the exact structure, it is more relevant to recover a model that approximates the original distribution statistically and on which inference can be performed efficiently. For example, Wainwright [Wai06] discusses situations in which it is computationally beneficial to use inconsistent learning algorithms (even in the infinite sample limit) to feed into approximate inference algorithms. Trees play an important role for inference algorithms, since the commonly used sum-product algorithm assumes tree structure, and other more general inference algorithms (like the junction tree algorithm and various approximate inference techniques) rely on tree-like structure. This is what motivates the notion of learning ε-approximate trees considered in this paper.

The problem of learning ε-approximate graphical models has a long history. Höffgen [Höf93] proved* Lemma 1.1 for distributions on $\{0,1\}^n$. There have been several other works which provide PAC-learning guarantees for generalizations of trees: bounded tree-width junction trees [NB04, CG08], factor graphs [AKN06], and forest-structured MRF's [LXG+11]. While we consider the KL divergence between the true distribution and the output of Chow-Liu, Bresler and Karzand [BK20] recently studied the same question with respect to maximum total variation distance between pairwise marginals. Their work is restricted to Ising models, and their sample complexity depends on bounds on distributional parameters (edge weights) while ours do not. In another recent work, Brustle, Cai and Daskalakis [BCD20] (generalizing the results in

*Höffgen's result is slightly different in that he doesn't use the plug-in estimator for mutual information.

[DMR20]) get bounds on the sample complexity of learning ε-approximate[†] MRF's with bounded hyper-edges and Bayesian networks with bounded in-degree, but they do not get efficient algorithms for these problems.

The main technical component of our analysis of the Chow-Liu algorithm is a new *conditional independence tester* which falls in the framework of *distribution property testing* [GR11, BFF+01]. We refer the reader to the surveys [Can15, Rub12] and the textbook [Gol17] and references therein for more details and results in this rapidly progressing field. Testing independence of two or more random variables has received some attention in distribution testing [BFF+01, ADK15, CDKS18]. The simplest formulation of the problem is the following: Given samples from an unknown joint distribution on variables (X, Y): decide with probability $\geq 2/3$ whether X and Y are independent or they are ε far (under some distance measure) from the product distribution $X \times Y$. Recently, [CDKS18] considered the problem of conditional independence testing. In particular, given samples from an unknown discrete random variable (X, Y, Z), the [CDKS18] tester can distinguish the case that X and Y are conditionally independent given Z from the case that (X, Y, Z) is ε-far in TV distance from every distribution that has this property. The key difference in our setting is that our alternative case is that $I(X; Y \mid Z) \geq \varepsilon$; Theorem A.1 of [CDKS18] gives an $\widetilde{O}(1/\varepsilon^2)$ bound for our setting, but notes that $O(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon})$ is plausible. This is precisely what we obtain (though our dependence on the domain size is worse than [CDKS18] and presumably suboptimal).

The parameter learning problem (i.e., learning the distribution with given structure) is also well-studied. Dasgupta [Das97] showed an $\widetilde{O}(n^2 2^d \log(1/\delta)/\varepsilon^2)$ for learning an ε-approximate Bayes net on n boolean variables with in-degree at most d. We improve the dependence on n and ε^{-1} to linear. For the realizable setting[‡], [BGMV20] also obtained the same improvement but for constant error probability δ. The key to obtaining our $\log(1/\delta)$ dependence is a PAC analysis of the add-1 estimator, which is new to the best of our knowledge. [KOPS15] analyze the expected risk, which does not directly imply a high-probability bound.

Finally, we note that while our work focuses on learning tree structured distributions, recent works [DDK19, CDKS20, DP17, BGMV20] have investigated testing problems for more general classes of high-dimensional distributions.

Concurrent Work. During preparation of this paper, a concurrent work by Daskalakis and Pan [DP20] was posted online. The headline result—that Chow-Liu learns tree-structured distributions with near-linear number of samples—is the same. The techniques employed there are quite different and more involved, with [DP20] working in squared Hellinger distance rather than KL and not involving the connection to conditional independence testing (Theorem 1.3). The details of the theorem are also somewhat different, most notably in that our result uses a $\log n$ factor more samples while [DP20] only works for a binary alphabet Σ.

2 Proof Overview

For the purposes of this proof overview, we consider a constant size alphabet Σ.

2.1 Finding an Approximate Tree

For any distribution P and a tree T, it is known that P_T is simply the distribution that matches the marginals of P on each edge of T. The Chow-Liu algorithm [CL68] is based on the following observation:

$$D(P \parallel P_T) = J_P - \mathrm{wt}_P(T) \tag{2.1}$$

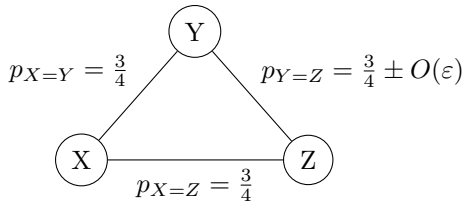
where $J_P = \sum_v H(P_v) - H(P)$ is independent of T (P_v is the marginal on variable v), and

$$\mathrm{wt}_P(T) := \sum_{(X,Y) \in T} I(X; Y)$$

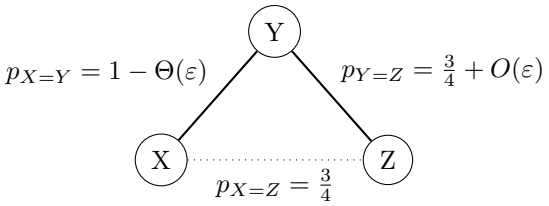
is the weight of T in the complete graph weighted by pairwise mutual information. Therefore $D(P \parallel P_T)$ is minimized when T is the maximum weight spanning tree T^* of this weighted complete graph.

[†] In total variation distance rather than KL

[‡] [CDKS20] also claimed this result (for TV distance) in an appendix, but the analysis there appears to incomplete [Can20].



(a) Hard instance for non-realizable setting. X, Y, and Z are individually uniform on $\{0,1\}$, and pairwise match with probability $\approx \frac{3}{4}$. Any ε-optimal tree will include edge YZ if $p_{Y=Z} = \frac{3}{4} + O(\varepsilon)$ and not if $p_{Y=Z} = \frac{3}{4} - O(\varepsilon)$; determining which takes $\Omega(1/\varepsilon^2)$ samples.



(b) A similar example in the realizable setting fails: if P is actually X-Y-Z-structured, and $p_{Y=Z}, p_{X=Z}$ are as on the left, then $p_{X=Y}$ must be $1 - O(\varepsilon)$. This means that $I(\widehat{Y};\widehat{Z})$ is highly correlated with $I(\widehat{X};\widehat{Z})$, so $I(\widehat{Y};\widehat{Z}) - I(\widehat{X};\widehat{Z})$ is ε-accurate with only $O(1/\varepsilon)$ samples.

Figure 1: The $\Omega(1/\varepsilon^2)$ bound in the non-realizable setting, and its inapplicability to the realizable setting.

The main question is how many samples are necessary for the maximum-weight spanning tree of the *empirical* distribution $\widehat{P}$ to have nearly-optimal weight under the true distribution P. That is, for Chow-Liu to recover a $\widehat{T}$ with $D(P \parallel P_{\widehat{T}}) \le D(P \parallel P_{T^*}) + \varepsilon$, it is necessary and sufficient that

$$\widehat{T} = \arg\max \mathrm{wt}_{\widehat{P}}(T)$$

satisfies

$$\mathrm{wt}_P(\widehat{T}) \ge \mathrm{wt}_P(T^*) - \varepsilon. \tag{2.2}$$

The non-realizable setting. The simplest approach to achieving (2.2) would be to ensure that $|I(\widehat{X};\widehat{Y}) - I(X;Y)| \le \frac{\varepsilon}{2n}$ for all vertex pairs (X,Y). This guarantees for every T that $|\mathrm{wt}_{\widehat{P}}(T) - \mathrm{wt}_P(T)| \le |T|\frac{\varepsilon}{2n} < \varepsilon/2$, which gives (2.2). Estimating mutual information to within $\frac{\varepsilon}{2n}$ is possible with $\widetilde{\Theta}(n^2/\varepsilon^2)$ samples, with high probability. A union bound over all vertex pairs then gives the Lemma 1.1 upper bound. We also show that this bound is tight. Estimating $I(X;Y)$ to $\pm\varepsilon$ really does require $1/\varepsilon^2$ samples (for example, if X is uniform on $\{0,1\}$ and $\mathbf{Pr}[Y = X] = p \approx \frac{3}{4}$, then estimating $I(X;Y) = 1 - h(p)$ requires estimating p to within $\pm\Theta(\varepsilon)$). We can translate this hardness into a $\Omega(1/\varepsilon^2)$ lower bound for a (non–tree-structured) three-variable P [see Figure 1a]; and by concatenating $\Omega(n)$ of these instances together, we get an $\Omega(n^2/\varepsilon^2)$ lower bound.

The realizable setting. Fortunately, we can do much better in the realizable setting, where P is actually T^*-structured for some tree T^*. We show that the errors in estimating mutual information are correlated, as illustrated in Figure 1b, so that the *difference* between mutual informations will be estimated more accurately than the mutual information itself.

As an example, consider the three variable case, where the true T^* is X-Y-Z and we want to ensure the algorithm does not pick edge XZ over YZ. We use the identity:

$$I(Y;Z) - I(X;Z) = I(Y;Z \mid X) - I(X;Z \mid Y).$$

In order for picking XZ over YZ to be ε-bad, the left hand side must be at least ε. On the other hand, because P is X-Y-Z-structured, $I(X;Z \mid Y) = 0$, and hence $I(Y;Z \mid X) \ge \varepsilon$.

Chow-Liu looks at the empirical mutual information, which has the same identity:

$$I(\widehat{Y};\widehat{Z}) - I(\widehat{X};\widehat{Z}) = I(\widehat{Y};\widehat{Z} \mid \widehat{X}) - I(\widehat{X};\widehat{Z} \mid \widehat{Y}). \tag{2.3}$$

In order for Chow-Liu to return the wrong tree by picking XZ over YZ, this must be negative. For this to happen, either $I(\widehat{X};\widehat{Z} \mid \widehat{Y}) > \varepsilon/10$ or $I(\widehat{Y};\widehat{Z} \mid \widehat{X}) \le \varepsilon/10$. This is, effectively, a question about conditional independence testing—after how many samples can we distinguish the conditionally independent distribution $(X, Z \mid Y)$ from the ε-far from conditionally independent distribution $(Y, Z \mid X)$? Our Theorem 1.3 (discussed in the next section) shows that $\widetilde{O}(1/\varepsilon)$ samples suffice for the empirical conditional mutual information to distinguish these cases, so that (2.3) will be positive.

For the general n-variable case, consider the tree $\widehat{T}$ returned by Chow-Liu. We can pair up the edges in $\widehat{T} \setminus T^*$ with those from $T^* \setminus \widehat{T}$, such that each edge WZ in $\widehat{T} \setminus T^*$ is matched to an edge $XY \in T^* \setminus \widehat{T}$ along the $W \rightsquigarrow Z$ path in T^*. We then use the more complicated identity

$$\begin{aligned} I(X;Y) - I(W;Z) &= I(X;Y) - I(X;Z) + I(X;Z) - I(W;Z) \\ &= I(X;Y \mid Z) - I(X;Z \mid Y) + I(Z;X \mid W) - I(Z;W \mid X). \end{aligned}$$

As in the three-variable case, the negative terms in this RHS are zero, and (if picking WZ over XY is ε/n-bad) at least one of the positive terms is at least $\varepsilon/(2n)$. If this is the case, then again Theorem 1.3 means that the empirical estimates of these terms, after $\widetilde{O}(n/\varepsilon)$ samples, will be sufficiently accurate that $I(\widehat{X};\widehat{Y}) - I(\widehat{W};\widehat{Z}) > 0$, and hence Chow-Liu will choose XY over WZ. As a result, with $\widetilde{O}(n/\varepsilon)$ samples, the tree $\widehat{T}$ recovered by Chow-Liu will satisfy (2.2), giving Theorem 1.2.

2.2 Conditional Independence Testing

Independence Testing. To build up to conditional independence testing with respect to mutual information, consider *unconditional* independence testing: given samples $(X,Y) \sim P_{XY}$, determine whether $I(X;Y)$ is 0 or $\geq \varepsilon$. We would like to show that, with $O(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon})$ samples, the empirical mutual information $I(\widehat{X};\widehat{Y})$ will distinguish between these two cases. [Note that $\Omega(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon})$ samples are necessary even in the binary setting: if $X = Y$ always, but $\mathbf{Pr}[X=1]$ is either 0 or $\varepsilon/\log(1/\varepsilon)$, the mutual information is either 0 or $\Theta(\varepsilon)$, but the first $\Omega(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon})$ samples will probably all be zero in either case.]

For intuition, consider the binary setting. Let $p_y = \mathbf{Pr}[X = 1 \mid Y = y]$, and $p = \mathbf{Pr}[X=1] = \mathbb{E}_y[p_y]$, so

$$I(X;Y) = H(X) - H(X \mid Y) = h(p) - \mathop{\mathbb{E}}_{y\sim Y}[h(p_y)]$$

for the binary entropy function h. Now, estimating either $h(p)$ or the $h(p_y)$ to $\pm\varepsilon$ would require $1/\varepsilon^2$ samples: if $p \approx \frac{1}{4}$, we would need $|\widehat{p} - p| \lesssim \varepsilon$ to estimate the individual entropies accurately. But if we expand $h(p_y)$ in a Taylor expansion around $p = \mathbb{E}_y[p_y]$, the *constant and linear terms cancel* leaving $I(X;Y) \approx \frac{1}{2}h''(p)E_y[(p_y - p)^2]$. So distinguishing $I(X;Y) = 0$ from $I(X;Y) \geq \varepsilon$ involves distinguishing between $\mathbb{E}[(p_Y - p)^2] = 0$ and $\mathbb{E}[(p_Y - p)^2] \gtrsim \varepsilon/h''(p)$. Up to a log factor coming from $h''(p)$, at least if the distribution of y is fairly balanced, this means it suffices to estimate each p_y to within $\pm\sqrt{\varepsilon}/10$, which takes $O(1/\varepsilon)$ samples.

More formally and more generally, by expressing mutual information as KL and removing each entry in the sum's linear dependence on $\Delta_{xy} := P_{xy} - P_xP_y$, we can write

$$I(X;Y) = D(P_{XY} \parallel P_XP_Y) = \sum_{x,y\in\Sigma^2} f(\Delta_{xy}, P_xP_y) \tag{2.4}$$

for some function f satisfying

$$f(a,b) = \Theta\left(\min\left(\frac{a^2}{b}, |a|\log\left(2 + \frac{|a|}{b}\right)\right)\right).$$

We then apply Chernoff bounds to show that every individual entry of the sum (2.4) concentrates: in the completeness case (Lemma 4.5), for any x, y,

$$f(\widehat{\Delta}_{xy}, \widehat{P}_x\widehat{P}_y) \gtrsim f(\Delta_{xy}, P_xP_y) - \frac{\log N \log(1/\delta)}{N}$$

with probability $1-\delta$, and in the soundness case (Lemma 4.7)

$$f(\widehat{\Delta}_{xy}, \widehat{P}_x\widehat{P}_y) \lesssim \frac{\log N \log(1/\delta)}{N}.$$

Taking a union bound over X and Y, and plugging this into the sum (2.4), gives the desired tester: as long as $\frac{N}{\log N} \gtrsim \Sigma^2\frac{\log(\Sigma/\delta)}{\varepsilon}$, the empirical mutual information will distinguish between $I(X;Y) > \varepsilon$ and $I(X;Y) = 0$.

The proofs of Lemma 4.5 and Lemma 4.7 are somewhat technical, but straightforward. We give intuition for the soundness case and constant probability. We use the two branches of f depending on whether $P_{xy} = P_x P_y$ is large or small. If $P_{xy} \lesssim 1/N$, then we will typically have $\widehat{\Delta}_{xy} \leq \widehat{P}_{xy} \lesssim 1/N$, so

$$f(\widehat{\Delta}_{xy}, \widehat{P}_x \widehat{P}_y) \lesssim \widehat{\Delta}_{xy} \log N \lesssim \frac{\log N}{N}.$$

On the other hand, if $P_{xy} \gg 1/N$, then typically $\widehat{P}_x = \Theta(P_x)$, $\widehat{P}_y = \Theta(P_y)$, and (we show) $|\widehat{\Delta}_{xy}| \lesssim \sqrt{P_x P_y / N}$. Hence

$$f(\widehat{\Delta}_{xy}, \widehat{P}_x \widehat{P}_y) \lesssim \frac{\widehat{\Delta}_{xy}^2}{\widehat{P}_x \widehat{P}_y} \lesssim \frac{1}{N}.$$

Conditional independence testing. By definition,

$$I(X; Y \mid Z) = \sum_z \mathbf{Pr}[Z = z] I(X; Y \mid Z = z).$$

Given N samples of (X, Y, Z), we expect about $N \mathbf{Pr}[Z = z]$ samples from $(X; Y \mid Z = z)$. This means our unconditional mutual independence tester will distinguish $I(X; Y \mid Z = z) = 0$ from $I(X; Y \mid Z = z) > \frac{|\Sigma|^2}{\widetilde{\Omega}(N \mathbf{Pr}[Z=z])}$. If the distribution passes all these independence checks, then

$$I(X; Y \mid Z) \leq \sum_z \mathbf{Pr}[Z = z] \frac{|\Sigma|^2}{\widetilde{\Omega}(N \mathbf{Pr}[Z = z])} = \frac{|\Sigma|^3}{\widetilde{\Omega}(N)}.$$

Thus $N = \widetilde{O}(|\Sigma|^3/\varepsilon)$ samples suffice to test conditional independence. A bit more care shows that the empirical conditional mutual information works as a test statistic, achieving Theorem 1.3.

2.3 Distribution Learning with Known Structure

For Theorem 1.4, it is implicit in [CL68, Das97] that it suffices to learn the conditional distributions in KL. While the empirical add-1 estimator of a discrete distribution was known to have small *expected* KL error with $\widetilde{O}(|\Sigma|/\varepsilon)$ samples [KOPS15], to our knowledge a high-probability bound was not known. We use a similar analysis to our independence tester—including the same decomposition (2.4) of KL—to show that the empirical add-1 estimator is accurate with high probability (Theorem 6.1). We then show that our samples from P give enough samples from each individual conditional distribution to estimate P well.

2.4 Questions

A natural question is whether the $|\Sigma|^3$ dependence in our bounds can be improved. The $|\Sigma|^3$ term is necessary to achieve Theorem 1.3 as stated; with fewer measurements, in the soundness case of a perfectly uniform distribution, the empirical conditional mutual information will exceed ε. However, it is quite possible that the empirical conditional mutual information—though $\gg \varepsilon$—is still smaller than in the completeness case. Just such behavior occurs when using the empirical total variation statistic in testing total variation from uniformity [DGPP18].

Another natural question is whether one can reduce tree structure learning to conditional independence testing as a black box. The Chow-Liu algorithm only considers pairwise mutual information, and never looks at conditional mutual information at all. Our analysis introduces conditional mutual information through (2.3), which relies on the test statistic being the empirical mutual information. If future work develops better conditional independence testers based on different test statistics, does that imply more sample-efficient (but possibly slower) algorithms for tree structure learning?

Recovering the structure of an unknown bounded degree Bayesian network remains an outstanding open question. Recently, Brustle, Cai, and Daskalakis [BCD20] have settled the sample complexity of this problem. But finding a polynomial time algorithm remains a challenge, even if we assume a correct topological ordering of the variables.

2.5 Organization

The rest of the paper is organized as follows. Section 3 describes the background and fixes notation. Section 4 analyzes conditional independence testing via empirical mutual information (Theorem 1.3). Section 5 uses this to show that Chow-Liu recovers an ε-approximate tree (Theorem 1.2). Section 6 shows how to recover the distribution given the tree (Theorem 1.4). Finally, Section 7 gives lower bounds for finding an ε-approximate tree T, showing that Lemma 1.1 and Theorem 1.2 are nearly optimal.

3 Notation and Preliminaries

For an undirected tree T, a *rooted orientation* of T fixes a root vertex and orients the edges outwards from it. For a rooted orientation of T, if i is a vertex in T, $\mathrm{pa}(i)$ denotes its parent node if any, and $\mathrm{nd}(i)$ denotes the subset of vertices not reachable from i.

Definition 3.1 (Tree-structured distributions)**.** *Let T be a tree. Fix any rooted orientation of it. Label the nodes of T in topological order (so, node 1 is the root). A probability distribution P over $X = (X_1, \ldots, X_n) \in \Sigma^n$ is said to be* T-structured *if: every variable X_i is conditionally independent of $\{X_j : j \in \mathrm{nd}(i)\}$ given $X_{\mathrm{pa}(i)}$.Equivalently, P admits the following factorization:*

$$\mathbf{Pr}[X = x] \coloneqq \mathbf{Pr}[X_1 = x_1] \cdot \prod_{i=2}^{n} \mathbf{Pr}[X_i = x_i \mid X_{\mathrm{pa}(i)} = x_{\mathrm{pa}(i)}]$$

A tree-structured distribution *is T-structured for some tree T.*

The following classical result justifies why the rooted orientation does not matter in Definition 3.1.

Theorem 3.2 ([VP90])**.** *Let T be a tree on n variables, and suppose P is a T-structured distribution on $(X_1, \ldots, X_n)$. For any 3 nodes $i, j, k \in [n]$, if the unique path between i and k in T passes through j, then X_i and X_k are independent conditioned on X_j.*

To compare distributions, we use the well-known notion of KL-divergence. Given two discrete probability distributions P and Q over Σ, their KL-divergence is defined as[§]

$$D(P \parallel Q) \coloneqq \sum_{x \in \Sigma} P(x) \log \frac{P(x)}{Q(x)}.$$

Recall that we say a tree T is *ε-approximate for a distribution P* if there exists a T-structured distribution Q such that:

$$D(P \parallel Q) \le \varepsilon + \min_{\text{tree } T'} \min_{\substack{T'\text{-structured} \\ \text{distribution } Q'}} D(P \parallel Q').$$

The following lemma is implicit in [CL68, Das97]. We defer the proof to Appendix A.

Lemma 3.3. *For a fixed tree T, let $pa(v)$ denote the parent of v in T (or $\perp$ if v is the root). Let $X \sim P$ and $X' \sim Q$ for some T-structured Q. Then, if $D(P \parallel Q)$ is bounded:*

$$\begin{aligned} D(P \parallel Q) = &\left(-H(X) + \sum_{v \in V} H(X_v) \right) - \sum_{v \in V} I(X_v; X_{pa(v)}) \\ &+ \sum_{v \in V} \sum_{x \in \Sigma} \mathbf{Pr}[X_{pa(v)} = x] D(X_v \mid X_{pa(v)} = x \parallel X'_v \mid X'_{pa(v)} = x) \end{aligned}$$

Lemma 3.3 is a generalization of (2.1): since KL divergence is nonnegative, the $Q = P_T$ minimizing $D(P \parallel Q)$ has every entry in the final sum is zero, which happens if Q matches the marginals of P on each edge of T. In that case, $D(P \parallel P_T)$ drops the final sum and gives (2.1), which we write formally:

[§]All logarithms in this paper are natural, so we measure information-theoretic quantities in nats not bits.

Algorithm 1: Learning the Skeleton of Tree-structured Distributions from Samples

input : Sample access to P over $X_1, \ldots, X_n \in \Sigma$
output: A tree T

$\hat{P} \leftarrow$ the empirical distribution of m i.i.d. samples from P;
for *every* $1 \le i < j \le n$ **do**
  $I(\widehat{X}_i, \widehat{X}_j) \leftarrow$ the mutual information between the variables X_i and X_j with respect to $\hat{P}$;
end
$G \leftarrow$ the weighted complete undirected graph on $[n]$ whose edge-weight (i, j) is $I(\widehat{X}_i, \widehat{X}_j)$;
$T \leftarrow$ a maximum weight spanning tree of G;
return T

Figure 2: The Chow-Liu Algorithm

Corollary 3.4 ([CL68]). *Let P be a distribution over Σ^n and T be an undirected tree over the vertex set $[n]$. Let P_T be the most likely distribution of P for the tree T. Then,*

$$D(P \parallel P_T) = J_P - \mathrm{wt}_P(T) \tag{3.1}$$

where $J_P = \sum_v H(P_v) - H(P)$ is independent of T (P_v is the marginal on variable v), and $\mathrm{wt}_P(T) := \sum_{(X,Y)\in T} I(X;Y)$.

This suggests the Chow-Liu algorithm (see Figure 2) that we analyze.

4 Testing Independence and Conditional Independence

Setup. We assume all random variables are over a discrete domain Σ. Let X and Y be random variables over Σ distributed jointly according to P. For any pair $x, y \in \Sigma$, let P_x, P_y, P_{xy} denote $\mathbf{Pr}[X = x], \mathbf{Pr}[Y = y], \mathbf{Pr}[(X, Y) = (x, y)]$, respectively. Let $\Delta_{xy} := P_{xy} - P_x P_y$. Hence $\sum_{xy} \Delta_{xy} = 0$. Let $(\widehat{X}, \widehat{Y})$ be the random variable distributed according to the *empirical* distribution $\widehat{P}$ over (X, Y) over a finite set of independent samples. Let $\widehat{P}_x, \widehat{P}_y, \widehat{P}_{xy}, \widehat{\Delta}_{xy}$ denote the same values for $(\widehat{X}, \widehat{Y})$.

Define

$$f(a, b) := (a + b)\log(1 + a/b) - a$$

for all $b \in [0, 1], a \in [-b, 1 - b]$ [with $f(-b, b) = b$, being the limiting value].

Claim 4.1. *For two random variables X and Y over Σ, $I(X;Y) = \sum_{x,y} f(\Delta_{xy}, P_x P_y)$.*

Proof.

$$\begin{aligned} I(X;Y) &= D(P_{XY} || P_X P_Y) \\ &= \sum_{x,y} (P_x P_y + \Delta_{xy}) \log(1 + \Delta_{xy}/(P_x P_y)) \\ &= \sum_{x,y} [(P_x P_y + \Delta_{xy}) \log(1 + \Delta_{xy}/(P_x P_y)) - \Delta_{xy}] \\ &= \sum_{x,y} f(\Delta_{xy}, P_x P_y). \end{aligned} \tag{4.1}$$

□

4.1 Analysis of f

Lemma 4.2. *For any $a \ge -b$ and $b \ge 0$,*

$$f(a, b) = C_{a,b} \min\left(\frac{a^2}{b}, |a| \log\left(2 + \frac{|a|}{b}\right)\right)$$

where the coefficient $1/3 \le C_{a,b} \le 1$.

Proof. Let $h(z) = (1+z)\log(1+z) - z$ and $g(z) = \min(z^2, |z|\log(2+|z|))$. Then, $g(z) = z^2$ for $-1 \le z \le 1.314$ and $g(z) = z\log(2+z)$ otherwise. We can show that

$$\begin{aligned} z^2/3 \le & h(z) \le z^2 && \text{(for } -1 \le z \le 1.5\text{)} \\ z\log(2+z)/3 \le & h(z) \le z\log(2+z) && \text{(for } z \ge 1\text{)} \end{aligned}$$

using calculus and Taylor expansion. Hence $g(z)/3 \le h(z) \le g(z)$. We get the result upon choosing $z = a/b$. □

Claim 4.3. *For any* x, y*, the following holds:*

1. $f(\Delta_{xy}, P_xP_y) \le 1$.
2. $\min(P_x, P_y, |\Delta_{xy}|) \gtrsim f(\Delta_{xy}, P_xP_y)/\log(3/f(\Delta_{xy}, P_xP_y))$.

Proof. Of (1): WLOG $P_x \le P_y$. Note that $-P_xP_y \le \Delta_{xy} \le P_x - P_xP_y$, and $f(x,a)$ is convex in x, so that:

$$\begin{aligned} f(\Delta_{xy}, P_xP_y) &\le \max(f(-P_xP_y, P_xP_y), f(P_x - P_xP_y, P_xP_y)) \\ &\le \max(P_xP_y, P_x\log(1+1/P_y) - P_x + P_xP_y) \\ &\le \max(1, P_x\log(1+1/P_x)) \\ &\le 1. \end{aligned}$$

Of (2): We have

$$f(\Delta_{xy}, P_xP_y) \lesssim |\Delta_{xy}|\log\left(2 + \frac{|\Delta_{xy}|}{P_xP_y}\right) \le P_x\log\left(2 + \frac{1}{P_y}\right) \le P_x\log\left(\frac{3}{P_x}\right)$$

and hence

$$\min(P_x, P_y) \gtrsim f(\Delta_{xy}, P_xP_y)/\log(3/f(\Delta_{xy}, P_xP_y)).$$

But then

$$|\Delta_{xy}| \gtrsim f(\Delta_{xy}, P_xP_y)/\log\left(\frac{3}{P_y}\right) \gtrsim f(\Delta_{xy}, P_xP_y)/\log(3/f(\Delta_{xy}, P_xP_y)),$$

finishing the result. □

4.2 Properties of the Empirical Distribution

By Chernoff bounds, the empirical distribution is close to the actual one:

Claim 4.4. *Let* $\widehat{P}_x, \widehat{P}_y$*, and* $\widehat{P}_{xy}$ *be empirical distributions over* $N > 1$ *samples. Then with* $1-3\delta$ *probability, all of the following bounds hold:*

$$\begin{aligned} |\widehat{P}_x - P_x| &\lesssim \sqrt{\frac{P_x\log\frac{2}{\delta}}{N}} + \frac{\log\frac{2}{\delta}}{N} \\ |\widehat{P}_y - P_y| &\lesssim \sqrt{\frac{P_x\log\frac{2}{\delta}}{N}} + \frac{\log\frac{2}{\delta}}{N} \\ |\widehat{P}_{xy} - P_{xy}| &\lesssim \sqrt{\frac{P_{xy}\log\frac{2}{\delta}}{N}} + \frac{\log\frac{2}{\delta}}{N} \\ |\widehat{P}_x\widehat{P}_y - P_xP_y| &\lesssim \sqrt{P_xP_y\frac{\log\frac{2}{\delta}}{N}} + (P_x+P_y)\frac{\log\frac{2}{\delta}}{N} + \frac{\log^2\frac{2}{\delta}}{N^2} \\ |\widehat{\Delta}_{xy} - \Delta_{xy}| &\lesssim \sqrt{|\Delta_{xy}|\frac{\log\frac{2}{\delta}}{N}} + \sqrt{P_xP_y\frac{\log\frac{2}{\delta}}{N}} + \frac{\log\frac{2}{\delta}}{N}. \end{aligned}$$

Proof. By the multiplicative Chernoff bound,

$$\mathbf{Pr}[|\widehat{P}_x - P_x| > \varepsilon P_x] < 2\exp(-C\min(\varepsilon, \varepsilon^2)P_x N).$$

Rearranging, with probability $1-\delta$,

$$|\widehat{P}_x - P_x| \lesssim \max\left(\frac{\log\frac{2}{\delta}}{P_x N}, \sqrt{\frac{\log\frac{2}{\delta}}{P_x N}}\right) P_x \le \sqrt{\frac{P_x \log\frac{2}{\delta}}{N}} + \frac{\log\frac{2}{\delta}}{N}.$$

Similarly for $\widehat{P}_y$ and $\widehat{P}_{xy}$. Then

$$\begin{aligned}
|\widehat{P}_x\widehat{P}_y - P_xP_y| &\lesssim \sqrt{\frac{\log\frac{2}{\delta}}{N}}(P_x\sqrt{P_y} + P_y\sqrt{P_x}) + \sqrt{P_xP_y}\frac{\log\frac{2}{\delta}}{N} + (P_x+P_y)\frac{\log\frac{2}{\delta}}{N} + (\sqrt{P_x}+\sqrt{P_y})\frac{\log^{1.5}\frac{2}{\delta}}{N^{1.5}} + \frac{\log^2\frac{2}{\delta}}{N^2} \\
&\le 2\sqrt{P_xP_y\frac{\log\frac{2}{\delta}}{N}} + 2(P_x+P_y)\frac{\log\frac{2}{\delta}}{N} + 2(\sqrt{P_x+P_y})\frac{\log^{1.5}\frac{2}{\delta}}{N^{1.5}} + \frac{\log^2\frac{2}{\delta}}{N^2}. \\
&\le 2\sqrt{P_xP_y\frac{\log\frac{2}{\delta}}{N}} + 3(P_x+P_y)\frac{\log\frac{2}{\delta}}{N} + 2\frac{\log^2\frac{2}{\delta}}{N^2}.
\end{aligned}$$

This implies:

$$\begin{aligned}
|\widehat{\Delta}_{xy} - \Delta_{xy}| &\le |\widehat{P}_{xy} - P_{xy}| + |\widehat{P}_x\widehat{P}_y - P_xP_y| \\
&\lesssim \sqrt{P_{xy}\frac{\log\frac{2}{\delta}}{N}} + \sqrt{P_xP_y\frac{\log\frac{2}{\delta}}{N}} + \frac{\log\frac{2}{\delta}}{N}.
\end{aligned}$$

The result follows because $P_{xy} \le 2\Delta_{xy}$ whenever that term is largest. □

4.3 Completeness

Lemma 4.5. *Let $\widehat{P}$ be the empirical distribution over N samples. Then for every $\delta > 0$ there exist constants $C, C' > 0$ such that: if $f(\Delta_{xy}, P_xP_y) \ge C\frac{\log N}{N}\log\frac{2}{\delta}$, then*

$$\mathbf{Pr}[f(\widehat{\Delta}_{xy}, \widehat{P}_x\widehat{P}_y) > C' f(\Delta_{xy}, P_xP_y)] > 1 - 3\delta.$$

Proof. By Claim 4.3 $f(\Delta_{xy}, P_xP_y) \le 1$. Claim 4.3 also implies

$$\min(|\Delta_{xy}|, P_x, P_y) \gtrsim \frac{f(\Delta_{xy}, P_xP_y)}{\log(2/f(\Delta_{xy}, P_xP_y))} \gtrsim \frac{C}{N}\log\frac{2}{\delta}. \tag{4.2}$$

Suppose that the Claim 4.4 statements hold, as happens with $1-3\delta$ probability. We will show that this implies the result.

We split into cases, based on whether $\Delta_{xy} > 8P_xP_y$.

Large Δ_{xy}. This case of $\Delta_{xy} > 8P_xP_y$ implies

$$f(\Delta_{xy}, P_xP_y) \eqsim \Delta_{xy}\log\frac{\Delta_{xy}}{P_xP_y}.$$

In this regime, we have by Claim 4.4 holding that

$$|\widehat{\Delta}_{xy} - \Delta_{xy}| \lesssim \sqrt{\frac{\Delta_{xy}}{N}\log\frac{2}{\delta}} + \frac{\log\frac{2}{\delta}}{N}.$$

Since $N \gtrsim C\log\frac{2}{\delta}/\Delta_{xy}$ by (4.2), this implies

$$|\widehat{\Delta}_{xy} - \Delta_{xy}| \lesssim \Delta_{xy}/C$$

and hence $|\widehat{\Delta}_{xy} - \Delta_{xy}| < \Delta_{xy}/10$ for a sufficiently large C.

We also have by Claim 4.4 holding that

$$|\widehat{P}_x\widehat{P}_y - P_xP_y| \lesssim \sqrt{P_xP_y\frac{\log\frac{2}{\delta}}{N}} + (P_x+P_y)\frac{\log\frac{2}{\delta}}{N} + \frac{\log^2\frac{2}{\delta}}{N^2} \lesssim \sqrt{\frac{\Delta_{xy}}{N}\log\frac{2}{\delta}} + \frac{\log\frac{2}{\delta}}{N}$$

and hence (by (4.2)) $|\widehat{P}_x\widehat{P}_y - P_xP_y| \le \Delta_{xy}/10$ for a sufficiently large C. This implies $\widehat{P}_x\widehat{P}_y \le 0.23\Delta_{xy}$. Therefore:

$$\frac{\widehat{\Delta}_{xy}}{\widehat{P}_x\widehat{P}_y} \ge \frac{0.9\Delta_{xy}}{0.23\Delta_{xy}} > 3.9,$$

so that (in Lemma 4.2),

$$f(\widehat{\Delta}_{xy}, \widehat{P}_x\widehat{P}_y) \eqsim \widehat{\Delta}_{xy}\log\frac{\widehat{\Delta}_{xy}}{\widehat{P}_x\widehat{P}_y} \gtrsim \Delta_{xy}\log\frac{\Delta_{xy}}{\widehat{P}_x\widehat{P}_y}$$

Now,

$$\begin{aligned}
|\widehat{P}_x\widehat{P}_y - P_xP_y| &\lesssim \sqrt{P_xP_y\frac{\log\frac{2}{\delta}}{N}} + (P_x+P_y)\frac{\log\frac{2}{\delta}}{N} + \frac{\log^2\frac{2}{\delta}}{N^2} \\
&\lesssim \sqrt{P_xP_y\Delta_{xy}/C} + 2P_xP_y/C + P_xP_y/C^2 && \text{(Using (4.2))} \\
&\lesssim \frac{1}{\sqrt{C}}(P_xP_y + \sqrt{P_xP_y\Delta_{xy}}).
\end{aligned}$$

For sufficiently large constant C the constant factor is overcome, so that

$$\begin{aligned}
f(\widehat{\Delta}_{xy}, \widehat{P}_x\widehat{P}_y) &\gtrsim \Delta_{xy}\log\frac{\Delta_{xy}}{2P_xP_y + \sqrt{P_xP_y\Delta_{xy}}} \\
&\ge \Delta_{xy}\min(\log\frac{\Delta_{xy}}{4P_xP_y}, \log\frac{\Delta_{xy}}{2\sqrt{P_xP_y\Delta_{xy}}}) \\
&\eqsim \Delta_{xy}\min(\log\frac{\Delta_{xy}}{P_xP_y}, \frac{1}{2}\log\frac{\Delta_{xy}}{P_xP_y}) \\
&\eqsim f(\Delta_{xy}, P_xP_y)
\end{aligned}$$

as desired.

Small Δ_{xy}. This case of $-P_xP_y \le \Delta_{xy} \le 8P_xP_y$ implies

$$f(\Delta_{xy}, P_xP_y) \eqsim \Delta_{xy}^2/(P_xP_y) \le 64P_xP_y.$$

Now, the condition that $f(\Delta_{xy}, P_xP_y) > C\frac{\log N}{N}\log\frac{2}{\delta}$ implies

$$P_xP_y \ge \frac{1}{64}\Delta_{xy}^2/(P_xP_y) \gtrsim C\frac{\log N}{N}\log\frac{2}{\delta} \tag{4.3}$$

and hence $N \gtrsim C\log\frac{2}{\delta}/\min(P_x, P_y)$. Therefore, for a sufficiently large C, we have by Claim 4.4 that both:

$$\begin{aligned}
|\widehat{P}_x - P_x| &\le P_x/10 \\
|\widehat{P}_y - P_x| &\le P_y/10.
\end{aligned}$$

Furthermore, the condition (4.3) also implies:

$$|\Delta_{xy}| \gtrsim \sqrt{\frac{CP_xP_y}{N}\log\frac{2}{\delta}}. \tag{4.4}$$

Hence by Claim 4.4 and the conditions (4.3) and (4.4),

$$|\widehat{\Delta}_{xy} - \Delta_{xy}| \lesssim \sqrt{\frac{P_x P_y}{N} \log \frac{2}{\delta}} + \frac{\log \frac{2}{\delta}}{N} \lesssim |\Delta_{xy}|/\sqrt{C} + \Delta_{xy}^2/(P_x P_y C).$$

Using $|\Delta_{xy}| \le 8 P_x P_y$, we get $|\widehat{\Delta}_{xy} - \Delta_{xy}| < |\Delta_{xy}|/10$ for a sufficiently large constant C. Therefore:

$$f(\widehat{\Delta}_{xy}, \widehat{P}_x \widehat{P}_y) \gtrsim \frac{\widehat{\Delta}_{xy}^2}{\widehat{P}_x \widehat{P}_y} \eqsim \frac{\Delta_{xy}^2}{P_x P_y} \eqsim f(\Delta_{xy}, P_x P_y).$$

□

Corollary 4.6. *Let $\widehat{P}$ be the empirical distribution over $N > 1$ samples. Then for every $\delta > 0$ there exist universal constants $C_1, C_2 > 0$ such that:*

$$I(\widehat{X}; \widehat{Y}) \ge C_1 I(X;Y) - C_2 |\Sigma|^2 \frac{\log N}{N} \log \frac{|\Sigma|}{\delta}$$

with probability at least $1 - \delta$.

Proof. Lemma 4.5 has a condition on f being large. But in general, since $f \ge 0$ always, it shows that with probability $1 - 3\delta$,

$$f(\widehat{\Delta}_{xy}, \widehat{P}_x \widehat{P}_y) > C' f(\Delta_{xy}, P_x P_y) - C' C \frac{\log N}{N} \log \frac{2}{\delta}.$$

Taking a union bound over the sum (4.1), and rescaling δ by $3|\Sigma|^2$, we get the result. □

4.4 Soundness

Lemma 4.7. *Let $\widehat{P}$ be the empirical distribution over N samples. Then for every $\delta > 0$ there exists a universal constant $C > 0$ such that: if $\Delta_{xy} = 0$, then*

$$\mathbf{Pr}[f(\widehat{\Delta}_{xy}, \widehat{P}_x \widehat{P}_y) < C \frac{\log N}{N} \log \frac{2}{\delta}] > 1 - 3\delta.$$

Proof. As in the completeness section, we suppose that the equations of Claim 4.4 all hold. In particular, this implies that

$$|\widehat{\Delta}_{xy}| \lesssim \sqrt{\frac{P_x P_y}{N} \log \frac{2}{\delta}} + \frac{\log \frac{2}{\delta}}{N}.$$

We again split into cases depending on $P_x P_y < \frac{\log \frac{2}{\delta}}{CN}$ or not for a large enough constant C.

Small $P_x P_y$. Suppose $P_x P_y < \frac{\log \frac{2}{\delta}}{CN}$, so that

$$|\widehat{\Delta}_{xy}| \lesssim \frac{1}{N} \log \frac{2}{\delta}.$$

First note that, if either of $\widehat{P}_x = 0$ or $\widehat{P}_y = 0$, then $\widehat{P}_{xy} = 0$ and $\widehat{\Delta}_{xy} = 0$, so $f(\widehat{\Delta}_{xy}, \widehat{P}_x \widehat{P}_y) = 0$. Therefore, in order for $f(\widehat{\Delta}_{xy}, \widehat{P}_x \widehat{P}_y)$ to be nonzero, we must sample x and y in our set, in which case $\widehat{P}_x \widehat{P}_y \ge 1/N^2$.

Therefore

$$f(\widehat{\Delta}_{xy}, \widehat{P}_x \widehat{P}_y) \lesssim |\widehat{\Delta}_{xy}| \log(1 + \widehat{\Delta}_{xy}/(\widehat{P}_x \widehat{P}_y)) \lesssim |\widehat{\Delta}_{xy}| \log(1 + N^2) \lesssim \frac{1}{N} \log \frac{2}{\delta} \log N.$$

Large P_xP_y. If $P_xP_y \geq \frac{\log\frac{2}{\delta}}{CN}$, then

$$\min(P_x, P_y) \geq \frac{\log\frac{2}{\delta}}{CN}$$

and hence by Claim 4.4 holding we have

$$|\widehat{P}_x - P_x| \leq O\left(\sqrt{\frac{P_x}{N}\log\frac{2}{\delta}} + \frac{\log\frac{2}{\delta}}{N}\right) \leq P_x/2,$$

for a large enough C and similarly for $\widehat{P_y}$. Therefore:

$$f(\widehat{\Delta}_{xy}, \widehat{P}_x\widehat{P}_y) \lesssim \frac{\widehat{\Delta}^2_{xy}}{\widehat{P}_x\widehat{P}_y} \leq 4\frac{\widehat{\Delta}^2_{xy}}{P_xP_y} \lesssim \frac{1}{N}\log\frac{2}{\delta} + \frac{\log^2\frac{2}{\delta}}{N^2P_xP_y} \lesssim \frac{C}{N}\log\frac{2}{\delta}.$$

Therefore the result holds regardless of the case, as long as Claim 4.4 holds. □

Corollary 4.8. *Let $\widehat{P}$ be the empirical distribution over $N > 1$ samples. If P is a product distribution, then for every $\delta > 0$ there exist a universal constant $C_3 > 0$ such that:*

$$I(\widehat{X};\widehat{Y}) \leq \frac{\log N}{N}C_3|\Sigma|^2\log\frac{|\Sigma|}{\delta}$$

with probability at least $1-\delta$.

Proof. Follows from taking the sum (4.1) and applying a union bound over the events in Lemma 4.7 for all possible x, y. □

4.5 Conditional independence testing

Theorem 1.3 (Conditional Mutual Information Tester)**.** *Let (X,Y,Z) be three random variables over Σ, and $(\widehat{X},\widehat{Y},\widehat{Z})$ be the empirical distribution over a size N sample of (X,Y,Z). There exists a universal constant $0 < C < 1$ so that for any*

$$N \geq \Theta\left(\frac{|\Sigma|^3}{\varepsilon}\log\frac{|\Sigma|}{\delta}\log\frac{|\Sigma|\log(1/\delta)}{\varepsilon}\right),$$

the following results hold with probability $1-\delta$:

1. *If $I(X;Y \mid Z) = 0$, then $I(\widehat{X};\widehat{Y} \mid \widehat{Z}) < \varepsilon$.*
2. *If $I(X;Y \mid Z) \geq \varepsilon$, then $I(\widehat{X};\widehat{Y} \mid \widehat{Z}) > C \cdot I(X;Y \mid Z)$.*

Proof. For any $z \in \Sigma$ let N_z be the number of samples with $Z = z$.

Proof of (1): If $I(X;Y \mid Z) = 0$ then $I(X;Y \mid Z = z) = 0$ for each z. Then Corollary 4.8 gives us that, with probability at least $1-\delta$,

$$I(\widehat{X};\widehat{Y} \mid \widehat{Z} = z) \lesssim \frac{|\Sigma|^2}{N_z}\log\frac{|\Sigma|}{\delta}\log N_z \leq \frac{|\Sigma|^2}{N_z}\log\frac{|\Sigma|}{\delta}\log N.$$

Let $S \subseteq \Sigma$ contain the set of z such that $\mathbf{Pr}[|\widehat{P}_z - P_z| > P_z/2] \leq \delta$. By a Chernoff bound, this consists of all z with $P_z \geq O(\frac{\log 1/\delta}{N})$. With probability $1 - 2|\Sigma|\delta$, then,

$$\sum_{z\in S}\widehat{P}_z I(\widehat{X};\widehat{Y} \mid \widehat{Z} = z) \lesssim \sum_{z\in S}P_z\frac{|\Sigma|^2}{NP_z/2}\log\frac{|\Sigma|}{\delta}\log N \leq \frac{\log N}{N}2|\Sigma|^3\log\frac{|\Sigma|}{\delta}.$$

On the other hand, $z \notin S$ will have $\widehat{P}_z \lesssim \frac{\log(1/\delta)}{N}$ with probability $1-\delta$, so that with probability $1 - |\Sigma|\delta$

$$\sum_{z\notin S}\widehat{P}_z I(\widehat{X};\widehat{Y} \mid \widehat{Z} = z) \lesssim |\Sigma|\frac{\log(1/\delta)}{N}\log|\Sigma| \lesssim \frac{|\Sigma|^2\log(1/\delta)}{N}$$

is even smaller. Rescaling δ, we get with probability $1-\delta$ that

$$I(\widehat{X};\widehat{Y} \mid \widehat{Z}) = \sum_z \widehat{P}_z I(\widehat{X};\widehat{Y} \mid \widehat{Z}=z) \lesssim \frac{\log N}{N}|\Sigma|^3 \log\frac{|\Sigma|}{\delta}$$

which is at most ε for the desired N.

Proof of (2): Consider the set S of $z \in \Sigma$ which satisfy $P_z \times I(X;Y \mid Z=z^*) \geq \frac{I(X;Y|Z)}{2|\Sigma|}$. Note that this implies $P_z \geq \frac{\varepsilon}{2|\Sigma|\log|\Sigma|}$. We also have,

$$\begin{aligned}\sum_{z\in S} P_z \times I(X;Y \mid Z=z) &= \sum_z P_z \times I(X;Y \mid Z=z) - \sum_{z\notin S} P_z \times I(X;Y \mid Z=z)\\ &\geq I(X;Y \mid Z) - |\Sigma|\frac{I(X,Y \mid Z)}{2|\Sigma|}\\ &\geq I(X,Y \mid Z)/2.\end{aligned}$$

Our N is large enough that for $z \in S$, $\mathbb{E}[N_z] = P_z N > O(\log(|\Sigma|/\delta))$. Hence, with probability $1-\delta$, we have $N_z \geq N_z/2$ for all $z \in S$. Then Corollary 4.6 gives us, with probability $1-\delta$,

$$I(\widehat{X};\widehat{Y} \mid \widehat{Z}=z) \geq C_1 I(X;Y \mid Z=z) - 2C_2|\Sigma|^2\frac{\log(0.5NP_z)}{NP_z}\log\frac{|\Sigma|}{\delta},$$

for all $z \in S$. Multiplying P_z and summing over all $z \in S$ gives us:

$$\begin{aligned}I(\widehat{X};\widehat{Y} \mid \widehat{Z}) &\geq \sum_{z\in S} \widehat{P}_z I(\widehat{X};\widehat{Y} \mid \widehat{Z}=z)\\ &= \sum_{z\in S}\frac{P_z}{2}\left(C_1 I(X;Y \mid Z=z) - 2C_2|\Sigma|^2\frac{\log(0.5NP_z)}{NP_z}\log\frac{|\Sigma|}{\delta}\right)\\ &\geq \frac{C_1}{4}I(X;Y \mid Z) - C_2|\Sigma|^3\frac{\log N}{N}\log\frac{|\Sigma|}{\delta}.\end{aligned}$$

For N as large as given, the term being subtracted is at most $\frac{\varepsilon C_1}{8}$, which is at most half the first term. □

5 Tree Structure Recovery

5.1 Non-realizable Case

Let P be an unknown distribution over Σ^n and $\hat{P}$ be the empirical distribution of P for a certain number of samples to be fixed later. Our algorithm Chow-Liu returns a maximum spanning tree $\hat{T}$ of the complete graph whose edge weights for every pair of variables are given by the estimated mutual informations with respect to $\hat{P}$.

Let T^* be the tree minimizing $D(P \parallel P_{T^*})$. Recall that $\mathrm{wt}_P(T)$ is defined as the sum of the pairwise mutual informations across T. By Corollary 3.4, the Chow-Liu algorithm will return a tree $\widehat{T}$ satisfying

$$D(P \parallel P_{\widehat{T}}) \leq D(P \parallel P_{T^*}) + \varepsilon$$

if $\mathrm{wt}_P(\widehat{T}) \geq \mathrm{wt}_P(T^*) - \varepsilon$. Since $\widehat{T}$ maximizes $\mathrm{wt}_{\widehat{P}}(\widehat{T})$, it would suffice to ensure $\mathrm{wt}_{\widehat{P}}(T) = \mathrm{wt}_P(T) \pm \varepsilon/2$ for all T; and therefore it would suffice for

$$I(\widehat{X};\widehat{Y}) = I(X;Y) \pm \frac{\varepsilon}{2n}$$

for all pairs of variables (X,Y). The following result is standard, which analyzes the the plug-in estimator $H(\widehat{X})$ for estimating a single discrete entropy $H(X)$ to $\pm\varepsilon$ with probability $1-\delta$.

Theorem 5.1 ([AK01, Pan03]). *For* $N > O\left(\frac{|\Sigma|}{\varepsilon} + \frac{1}{\varepsilon^2}\log\frac{1}{\delta}\log^2(\frac{|\Sigma|}{\varepsilon}\log\frac{1}{\delta})\right)$, $|H(\widehat{X}) - H(X)| \leq \varepsilon$ *with probability at least* $(1-\delta)$.

Since $I(X;Y) = H(X) + H(Y) - H(X,Y)$, Theorem 5.1 tells us that once

$$N \geq O\left(\frac{|\Sigma|^2 n}{\varepsilon} + \frac{n^2}{\varepsilon^2}\log\frac{n}{\delta}\log^2(\frac{n|\Sigma|}{\varepsilon}\log\frac{n}{\delta})\right),$$

all the pairwise mutual informations of the variables of P will be estimated to within $\frac{\varepsilon}{2n}$. In that case, Chow-Liu would return a tree T, the best distribution on which would be close to the closest tree Bayes net of P.

Lemma 5.2. *Let P be any unknown distribution over Σ^n. Let Q be the tree Bayes net which is closest to P in* KL *distance. Then* Chow-Liu*, when run with $O\left(\frac{|\Sigma|^2 n}{\varepsilon} + \frac{n^2}{\varepsilon^2}\log\frac{n}{\delta}\log^2(\frac{n|\Sigma|}{\varepsilon}\log\frac{n}{\delta})\right)$ samples, returns a tree T such that there exists a T-structured R with $D(P \parallel R) \leq D(P \parallel Q) + \varepsilon$.*

We conclude this section by noting that when P itself is a tree Bayes net (realizable case) $D(P \parallel Q) = 0$ and the best Bayes net R on the tree returned by Chow-Liu with the sample complexity of Lemma 5.2 would satisfy $D(P \parallel R) \leq \varepsilon$ with probability at least $(1-\delta)$. In the next section, we show how to bring the sample complexity analysis down from $\widetilde{O}(n^2/\varepsilon^2)$ to $\widetilde{O}(n/\varepsilon)$ when P is actually tree-structured for some unknown tree.

5.2 Realizable Case

We will need the following fact which follows from the chain rule of mutual information.

Fact 5.3. *For three random variables X, Y, and Z*

$$I(X;Y) - I(X;Z) = I(X;Y \mid Z) - I(X;Z \mid Y).$$

Proof. Follows from observing $I(X;Y) + I(X;Z \mid Y) = I(X;Y,Z) = I(X;Z) + I(X;Y \mid Z)$. □

We also use the following fact about spanning trees:

Fact 5.4. *Let T_1 and T_2 be two spanning trees on n vertices such that their symmetric difference consists of the edges $E = \{e_1, e_2 \ldots, e_\ell\} \in T_1 \setminus T_2$ and $F = \{f_1, f_2 \ldots, f_\ell\} \in T_2 \setminus T_1$. Then E and F can be paired up, without loss of generality say $\langle e_i, f_i\rangle$, such that for all i, $T_1 \cup \{f_i\} \setminus \{e_i\}$ is a spanning tree.*

Proof. We use induction on ℓ. Base case of $\ell = 0$ is trivial.

Assume it holds for any two trees T_1 and T_2 so that $|T_1 \setminus T_2| = |T_2 \setminus T_1| = (l-1)$. Now, pick an arbitrary $e = (u,v) \in T_1 \setminus T_2$. $T_1 \setminus \{e\}$ has two connected components, $L \ni u$ and $R = (V \setminus L) \ni v$. In T_2, there is some path connecting u to v. This path starts in L and ends in R, so it must have some edge f connecting L to R. But $f \notin T_1$, since $e \notin T_2$ is the only edge connecting L and R in T_1.

Because f connects L and R, which are otherwise unconnected in $T_1 \setminus \{e\}$, $T_1 \cup \{f\} \setminus \{e\}$ is a spanning tree. Thus it is valid to pair $\langle e, f\rangle$. Furthermore, because f lies on the path connecting e in T_2, $T_3 := T_2 \cup \{e\} \setminus \{f\}$ is also a spanning tree, and it differs in only $\ell - 1$ edges from T_1. Therefore by induction, T_3 can be paired with T_1 in the desired way. Adding $\langle e, f\rangle$ to this pairing means that T_2 can be paired with T_1. □

Theorem 5.5. *Let N be such that the bound in Theorem 1.3 holds for a given $\varepsilon, \delta > 0$. Then in the realizable case with n nodes, with probability $1 - 4n\delta$* Chow-Liu *returns a tree $\widehat{T}$ with $D(P \parallel P_T) \leq \varepsilon n$.*

Proof. Let P be the unknown distribution on the true tree T^*. Let $\widehat{T}$ be the tree returned by the Chow-Liu algorithm with N samples. For any set of vertex pairs S, let $\mathrm{wt}(S)$ and $\widehat{\mathrm{wt}}(S)$ denote the sum of mutual information over all pairs in S with respect to the true and empirical distributions respectively, so $\mathrm{wt}(T^*)$ and $\widehat{\mathrm{wt}}(\widehat{T})$ are each maximal over spanning trees.

For our analysis, we make at most $4n$ invocations of Theorem 1.3. We will assume the conclusion holds in all cases, as happens with at least $1 - 4n\delta$ probability.

Let $\{\langle e_i, f_i\rangle\}_i$ be a pairing given by Fact 5.4 for T^* and $\widehat{T}$. By (2.1),

$$D(P \parallel P_{\widehat{T}}) = \mathrm{wt}(T^*) - \mathrm{wt}(\widehat{T}) = \sum_i (\mathrm{wt}(\{e_i\}) - \mathrm{wt}(\{f_i\})).$$

For any i, let $e_i = (X_i, Y_i) \in T^*$ and $f_i = (W_i, Z_i) \in \widehat{T}$. Because $T^* \cup \{f_i\} \setminus \{e_i\}$ is a spanning tree, the path connecting W_i and Z_i in T^* must go through e_i. Without loss of generality let $W \rightsquigarrow X \to Y \rightsquigarrow Z$ be this path (it is possible that $W = X$ or $Y = Z$). Hence, from Theorem 3.2, we have for the true distribution

$$I(X_i; Z_i \mid Y_i) = I(Z_i; W_i \mid X_i) = 0$$

and so

$$\begin{aligned} I(X_i; Y_i) - I(W_i; Z_i) &= I(X_i; Y_i) - I(X_i; Z_i) + I(X_i; Z_i) - I(W_i; Z_i) \\ &= I(X_i; Y_i \mid Z_i) - I(X_i; Z_i \mid Y_i) + I(X_i; Z_i \mid W_i) - I(Z_i; W_i \mid X_i) \\ &= I(X_i; Y_i \mid Z_i) + I(Z_i; X_i \mid W_i). \end{aligned} \tag{5.1}$$

On the other hand the empirical distribution will have

$$I(\widehat{X_i}; \widehat{Y_i}) - I(\widehat{W_i}; \widehat{Z_i}) = I(\widehat{X_i}; \widehat{Y_i} \mid \widehat{Z_i}) - I(\widehat{X_i}; \widehat{Z_i} \mid \widehat{Y_i}) + I(\widehat{Z_i}; \widehat{X_i} \mid \widehat{W_i}) - I(\widehat{Z_i}; \widehat{W_i} \mid \widehat{X_i}). \tag{5.2}$$

Because $\widehat{T}$ is maximal under $\widehat{\mathrm{wt}}$,

$$\begin{aligned} 0 \geq \widehat{\mathrm{wt}}(T^*) - \widehat{\mathrm{wt}}(\widehat{T}) &= \sum_i (I(\widehat{X_i}; \widehat{Y_i}) - I(\widehat{W_i}; \widehat{Z_i})) \\ &= \sum_i \left(I(\widehat{X_i}; \widehat{Y_i} \mid \widehat{Z_i}) + I(\widehat{Z_i}; \widehat{X_i} \mid \widehat{W_i}) \right) - \sum_i \left(I(\widehat{X_i}; \widehat{Z_i} \mid \widehat{Y_i}) + I(\widehat{Z_i}; \widehat{W_i} \mid \widehat{X_i}) \right). \end{aligned} \tag{5.3}$$

We invoke Theorem 1.3 with $\varepsilon' := C\varepsilon/10$ where $C < 1$ is the constant given by the theorem. As a consequence of Theorem 1.3 and the fact that each of $I(X_i; Z_i \mid Y_i) = I(Z_i; W_i \mid X_i) = 0$, the second sum is at most $C\varepsilon n/5$.

On the other hand, Theorem 1.3 implies that

$$I(\widehat{X_i}; \widehat{Y_i} \mid \widehat{Z_i}) \geq C I(X_i; Y_i \mid Z_i) - C\varepsilon'$$

and similarly for $I(\widehat{Z_i}; \widehat{X_i} \mid \widehat{W_i})$. As a result, the first sum has

$$\begin{aligned} \sum_i \left(I(\widehat{X_i}; \widehat{Y_i} \mid \widehat{Z_i}) + I(\widehat{Z_i}; \widehat{X_i} \mid \widehat{W_i}) \right) &\geq C \sum_i (I(X_i; Y_i \mid Z_i) + I(Z_i; X_i \mid W_i) - 2\varepsilon') \\ &\geq C(\mathrm{wt}(T^*) - \mathrm{wt}(\widehat{T})) - 2C\varepsilon' n \end{aligned}$$

by (5.1). Combining these bounds into (5.3),

$$0 \geq C(\mathrm{wt}(T^*) - \mathrm{wt}(\widehat{T})) - \frac{1}{5}(C^2 + C)\varepsilon n$$

or

$$D(P \parallel P_{\widehat{T}}) = \mathrm{wt}(T^*) - \mathrm{wt}(\widehat{T}) \leq \frac{1}{5}(C + 1)\varepsilon n \leq \varepsilon n.$$

□

6 Distribution Recovery

This section shows how, for a fixed tree T, to find a T-structured distribution Q with $D(P \parallel Q) \leq D(P \parallel P_T) + \varepsilon$. We start by analyzing how to learn an arbitrary distribution over Σ.

6.1 KL Learning of Discrete Distributions

Given N samples from a distribution P over Σ, the "add-1" empirical estimator is based on Laplace's rule of succession. This distribution Q is defined by: for each item $i \in \Sigma$, if i appears T_i times in the samples, then $Q_i = \frac{T_i+1}{N+|\Sigma|}$. Kamath, Orlitsky, Pichapati and Suresh [KOPS15] have analyzed the expected behavior of the add-1 empirical estimator. In this section, we analyze its behavior in the high-probability regime.

Theorem 6.1. *Let P be a distribution over Σ and $N \geq 1$. Let Q be the empirical add-1 estimator from N samples of P. There is an universal constant $C > 0$ such that, with probability $1 - \delta$,*

$$D(P \parallel Q) \leq \frac{C|\Sigma| \log \frac{|\Sigma|}{\delta} \log N}{N}.$$

Proof. Let $C' > 1$ be a large constant to be determined later. If $N \leq C'|\Sigma|$, the result follows from $D(P \parallel Q) \leq \log \frac{1}{\min_i Q_i} \leq \log(N + |\Sigma|) \lesssim \log |\Sigma|$, so we may assume $N \geq C'|\Sigma|$. Then

$$\begin{aligned}
D(P \parallel Q)) &= \sum_i P_i \log \frac{P_i}{Q_i} \\
&= \sum_i f(P_i - Q_i, Q_i) && \text{(where } f(x,a) = a[(1+\tfrac{x}{a})\log(1+\tfrac{x}{a}) - \tfrac{x}{a}]\text{)} \\
&\lesssim \sum_i \min\left(\frac{(P_i - Q_i)^2}{Q_i}, |P_i - Q_i| \log\left(1 + \frac{|P_i - Q_i|}{Q_i}\right)\right) && \text{(From Lemma 4.2)} \\
&\lesssim \sum_i \min\left(\frac{(P_i - Q_i)^2}{Q_i}, |P_i - Q_i| \log N\right). && \text{(Since } Q_i \geq \tfrac{1}{N+|\Sigma|}\text{)}
\end{aligned}$$

We also know from Claim 4.4 that with probability at least $1 - \delta$ for each i,

$$\begin{aligned}
|P_i - \frac{T_i}{N}| &\lesssim \sqrt{\frac{P_i \log \frac{1}{\delta}}{N}} + \frac{\log \frac{1}{\delta}}{N} \\
\implies |P_i - Q_i| &\lesssim \sqrt{\frac{P_i \log \frac{1}{\delta}}{N}} + \frac{\log \frac{1}{\delta}}{N} + \left|\frac{T_i}{N} - \frac{T_i + 1}{N + |\Sigma|}\right| \\
&= \sqrt{\frac{P_i \log \frac{1}{\delta}}{N}} + \frac{\log \frac{1}{\delta}}{N} + \frac{||\Sigma|T_i/N - 1|}{N + |\Sigma|} \\
&\lesssim \sqrt{\frac{P_i \log \frac{1}{\delta}}{N}} + \frac{\log \frac{1}{\delta}}{N} + \frac{|\Sigma|}{N + |\Sigma|}\frac{T_i}{N} \\
&\lesssim \sqrt{\frac{P_i \log \frac{1}{\delta}}{N}} + \frac{\log \frac{1}{\delta}}{N} + \frac{|\Sigma|}{N + |\Sigma|}\left(P_i + \sqrt{\frac{P_i \log \frac{1}{\delta}}{N}} + \frac{\log \frac{1}{\delta}}{N}\right) \\
&\lesssim \sqrt{\frac{P_i \log \frac{1}{\delta}}{N}} + \frac{\log \frac{1}{\delta}}{N} + \frac{|\Sigma|}{N + |\Sigma|}P_i
\end{aligned}$$

If $P_i \leq \frac{C' \log \frac{1}{\delta}}{N}$, then $|P_i - Q_i| \log N \lesssim \frac{\log \frac{1}{\delta}}{N} \log N$.

If $P_i > \frac{C' \log \frac{1}{\delta}}{N}$, then $|P_i - Q_i| \lesssim P_i(\frac{1}{\sqrt{C'}} + \frac{1}{C'} + \frac{1}{C'+1})$ is at most $\frac{P_i}{2}$ for sufficiently large C', so $Q_i \geq P_i/2$ and hence $\frac{(P_i - Q_i)^2}{Q_i} \lesssim \frac{\log \frac{1}{\delta}}{N} + P_i\left(\frac{|\Sigma|}{N+|\Sigma|}\right)^2$.

By a union bound, with probability at least $1 - |\Sigma|\delta$ we have that these equations hold for all i. If true, then

$$D(P \parallel Q) \lesssim \frac{|\Sigma|}{N} \log \frac{1}{\delta} \log N + \left(\frac{|\Sigma|}{N + |\Sigma|}\right)^2 \lesssim \frac{|\Sigma|}{N} \log \frac{1}{\delta} \log N$$

Rescaling δ gives the desired bound. □

Algorithm 2: Learning closest T-structured distribution

input : Samples access to P over Σ^n; Rooted tree T on n nodes labeled in topological order.
output: n row-stochastic $|\Sigma| \times |\Sigma|$ matrices $Q_1, \ldots, Q_n$ that define a T-structured distribution Q by $Q(x) = \prod_{i=1}^{n} Q_i[x_{\mathrm{pa}(i)}, x_i]$ (where $x_{\mathrm{pa}(1)}$ is arbitrary).

Draw N i.i.d. samples $X^{(1)}, \ldots, X^{(N)}$ from P;
for $i \leftarrow 1$ **to** n **do**
  for $x \in \Sigma$ **do**
    $k \leftarrow \sum_{j=1}^{N} \mathbf{1}\left[X^{(j)}_{\mathrm{pa}(i)} = x\right]$; `// condition on parent satisfied vacuously if` $i = 1$
    for $y \in \Sigma$ **do**
      $t \leftarrow \sum_{j=1}^{N} \mathbf{1}\left[X^{(j)}_{\mathrm{pa}(i)} = x, X^{(j)}_{i} = y\right]$; `// condition on parent sat. vacuously if` $i = 1$
      $Q_i[x, y] \leftarrow (t+1)/(k+|\Sigma|)$;
    end
  end
end
return $(Q_1, \ldots, Q_n)$

Figure 3: Pseudocode for algorithm analyzed in Theorem 1.4.

6.2 Learning Trees

We are ready to prove the main result of this section. Figure 3 shows the algorithm we analyze below.

Theorem 1.4. *Let P be a discrete distribution over Σ^n. Let T be a tree on n vertices, and Q be a T-structured distribution with conditional probabilities at each node estimated using the empirical add-1 estimator on*

$$N = \Theta\left(\frac{n|\Sigma|^2}{\varepsilon} \log \frac{n|\Sigma|}{\delta} \log\left(\frac{n|\Sigma|}{\varepsilon} \log \frac{1}{\delta}\right)\right)$$

samples from P. Then

$$D(P \parallel Q) - D(P \parallel P_T) \le \varepsilon$$

with probability $1 - \delta$.

Proof. Note that $D(P \parallel Q)$ is bounded, because $Q(x) > 0$ for all x. By Lemma 3.3 and (2.1), the learned T-structured distribution Q satisfies

$$D(P \parallel Q) - D(P \parallel P_T) = \sum_{i \in [n]} \sum_{x \in \Sigma} \mathbf{Pr}[X_{\mathrm{pa}(i)} = x] \cdot D(X_i \mid X_{\mathrm{pa}(i)} = x \parallel X'_i \mid X'_{\mathrm{pa}(i)} = x) \tag{6.1}$$

where $X \sim P$ and $X' \sim Q$. Now, the node-wise conditional probabilities of Q are the add-1 empirical distribution of the conditional probabilities of P. Therefore by Theorem 6.1, if we have k samples of $(X \mid X_{\mathrm{pa}(i)} = x)$, we will have with probability $1 - \delta$ that

$$D(X_i \mid X_{\mathrm{pa}(i)} = x \parallel X'_i \mid X'_{\mathrm{pa}(i)} = x) \lesssim \frac{|\Sigma| \log(|\Sigma|/\delta) \log k}{k} \le \frac{|\Sigma| \log(|\Sigma|/\delta) \log N}{k}.$$

If $\mathbb{E}[k] = N \mathbf{Pr}[X_{\mathrm{pa}(i)} = x] > 15 \log \frac{1}{\delta}$, then by a Chernoff bound, with probability $1 - \delta$ we have $k > \frac{1}{2}\mathbb{E}[k]$ and

$$\mathbf{Pr}[X_{\mathrm{pa}(i)} = x] \cdot D(X_i \mid X_{\mathrm{pa}(i)} = x \parallel X'_i \mid X'_{\mathrm{pa}(i)} = x) \lesssim \frac{|\Sigma| \log(|\Sigma|/\delta) \log N}{N} \tag{6.2}$$

On the other hand, if $\mathbb{E}[k] \le 15 \log \frac{1}{\delta}$, then $D(X_i \mid X_{\mathrm{pa}(i)} = x \parallel X'_i \mid X'_{\mathrm{pa}(i)} = x) \le \log(k + |\Sigma|)$ [because Q is the add-1 estimator, so the minimum probability is $\frac{1}{k+|\Sigma|}$] and we have

$$\mathbf{Pr}[X_{\mathrm{pa}(i)} = x] \cdot D(X_i \mid X_{\mathrm{pa}(i)} = x \parallel X'_i \mid X'_{\mathrm{pa}(i)} = x) \le \frac{\mathbb{E}[k]}{N} \log(k + |\Sigma|) \lesssim \frac{\log \frac{1}{\delta} \log N}{N}$$

Regardless, each term in (6.1) is bounded by (6.2). Taking a union bound, with probability $1 - n|\Sigma|\delta$ we have

$$D(P \parallel Q) - D(P \parallel P_T) \lesssim n|\Sigma| \frac{|\Sigma| \log(|\Sigma|/\delta) \log N}{N}.$$

Rescaling δ and choosing N appropriately gives the result. □

The algorithm and analysis in Theorem 1.4 straightforwardly generalizes to Bayes nets. This shows that if G is a directed acyclic graph with in-degree bounded by d, we can obtain a G-structured distribution Q using $\widetilde{O}(n|\Sigma|^{d+1}/\varepsilon)$ samples from P which satisfies $D(P \parallel Q) - D(P \parallel P_G) \leq \varepsilon$, where $P_G = \arg\min_{G\text{-structured } R} D(P \parallel R)$.

7 Lower Bounds for Structure Recovery

7.1 Non-Realizable Case

This section focuses on the non-realizable case, i.e., the input distribution is not necessarily a tree structured distribution. We prove that $\Omega(n^2/\varepsilon^2)$ samples from a distribution P over $\{0,1\}^n$ are required to find an ε-approximate tree for P.

First, we prove the following lemma for $n = 3$.

Lemma 7.1. *There exist two distributions R and R' over $\{0,1\}^3$ such that:*

(i) $\Omega(1/\varepsilon^2)$ samples are required to distinguish R from R' with probability at least $2/3$.

(ii) Every tree T on 3 vertices is not ε-approximate for either R or R'.

Observe that Lemma 7.1 implies that if the distribution that generates the samples is chosen uniformly at random between R and R', then any algorithm that outputs an ε-approximate tree with error probability $< 1/3$ must draw $\Omega(\varepsilon^{-2})$ samples.

Proof of Lemma 7.1. Let $B \sim \text{Ber}(1/2)$ be a random bit. R is a distribution over $\{0,1\}^3$ where independently: $\mathbf{Pr}[R_1 = B] = 0.75 + \varepsilon$, $\mathbf{Pr}[R_2 = B] = 0.75 + \varepsilon$, and $\mathbf{Pr}[R_3 = B] = 0.75 - \varepsilon$. Similarly, R' is a distribution over $\{0,1\}^3$ where independently: $\mathbf{Pr}[R'_1 = B] = 0.75 + \varepsilon$, $\mathbf{Pr}[R'_2 = B] = 0.75 - \varepsilon$, and $\mathbf{Pr}[R'_3 = B] = 0.75 + \varepsilon$.

For part (i), a routine calculation (Claim 7.2) shows that $D(R \parallel R') = O(\varepsilon^2)$. Hence, $D(R^k \parallel R'^k) = O(k\varepsilon^2)$. By Pinsker's inequality, if $k < c\varepsilon^{-2}$ for a sufficiently small c, the total variation distance between k i.i.d. samples of R and of R' is at most $1/3$, proving the claim.

For part (ii), consider graphical models on three nodes X, Y, Z, corresponding to the first, second and third coordinates respectively of $\{0,1\}^3$. Let G be the tree X—Y—Z, and H the tree X—Z—Y. Let R_G denote the closest G-structured distribution to R, and similarly define R_H, R'_G, R'_H. Using Corollary 3.4, $D(R \parallel R_G) - D(R \parallel R_H) = I(R_1; R_2) - I(R_1; R_3)$. A calculation (Claim 7.3) shows that $I(R_1; R_2) - I(R_1; R_3) \approx -1.17\varepsilon$, which means $D(R \parallel R_H) = \Omega(\varepsilon)$. Similarly, $D(R' \parallel R'_G) - D(R' \parallel R'_H) = I(R'_1; R'_2) - I(R'_1; R'_3) \approx 1.17\varepsilon$, which means $D(R' \parallel R'_G) = \Omega(\varepsilon)$. The situation is analogous for the other tree Y—X—Z. □

Claim 7.2.

$$D(R \parallel R') = O(\varepsilon^2).$$

Proof. It is straightforward to check that $\mathbf{Pr}[R = 010] = \mathbf{Pr}[R = 101] = \frac{3}{32} - \frac{\varepsilon}{4} + O(\varepsilon^2)$ and $\mathbf{Pr}[R = 001] = \mathbf{Pr}[R = 110] = \frac{3}{32} + \frac{\varepsilon}{4} + O(\varepsilon^2)$ while $\mathbf{Pr}[R' = 010] = \mathbf{Pr}[R' = 101] = \frac{3}{32} + \frac{\varepsilon}{4} + O(\varepsilon^2)$ and $\mathbf{Pr}[R' = 001] = \mathbf{Pr}[R' = 110] = \frac{3}{32} - \frac{\varepsilon}{4} + O(\varepsilon^2)$. All other assignments have equal probability for R and R'. The result then follows from the definition of relative entropy. □

Claim 7.3.

$$I(R_1; R_2) - I(R_1; R_3) \approx -1.17\varepsilon.$$

Proof. In Appendix. □

We can now prove the main result of this section.

Theorem 7.4. *The sample complexity of computing an ε-approximate tree on n variables with error probability less than $1/3$ is $\Omega(n^2\varepsilon^{-2})$.*

Proof. Suppose $n = 3m$ is a multiple of 3. We consider the n variables as being divided into m blocks, each of size 3. Let P be a random distribution on $\{0,1\}^n$, defined by setting the distribution of the i'th block to be R or R' with probability $1/2$ each independently, where R and R' satisfy Lemma 7.1.

For the sake of contradiction, suppose we have an algorithm that draws $cn^2\varepsilon^{-2}$ samples from P (for a sufficiently small constant c) and outputs an ε-approximate tree T with probability at least $2/3$ (over the choice of P as well as the algorithm's randomness). Since each block is independent, without loss of generality, T is a union of disjoint trees $T_1, \ldots, T_m$ for each block. By Lemma 7.1, each T_i is not $10\varepsilon/m$-approximate with probability at least $1/3$. Hence, by a Chernoff bound, with probability $> 2/3$, for at least $\frac{m}{10}$ trees, T_i is not $10\varepsilon/m$-approximate. Therefore, for any T-structured distribution Q, $D(P \parallel Q) > \frac{m}{10} \cdot \frac{10\varepsilon}{m} = \varepsilon$. □

7.2 Realizable Case

We now show that if P is a tree-structured distribution on n variables, then $\Omega(n/\varepsilon)$ samples are required to find an ε-approximate tree. As with the non-realizable case, we first show the construction for $n = 3$.

Lemma 7.5. *There exist two tree-structured distributions R and R' over $\{0,1\}^3$ such that:*

(i) $\Omega(1/\varepsilon)$ samples are required to distinguish R from R' with probability at least $2/3$.

(ii) Every tree T on 3 vertices is not ε-approximate for either R or R'.

The same argument used for Theorem 7.4 implies:

Theorem 7.6. *The sample complexity of computing an ε-approximate tree for a tree-structured distribution on n variables with error probability less than $1/3$ is $\Omega(n/\varepsilon)$.*

Proof of Lemma 7.5. Let R be a distribution over $\{0,1\}^3$, defined as: $\mathbf{Pr}[R_1 = 1] = \frac{1}{2}$, $\mathbf{Pr}[R_2 = R_1] = 1-\varepsilon$, and $\mathbf{Pr}[R_3 = R_1] = \frac{3}{4}$. Also, define R' as: $\mathbf{Pr}[R'_2 = 1] = \frac{1}{2}$, $\mathbf{Pr}[R'_1 = R'_2] = 1-\varepsilon$, and $\mathbf{Pr}[R'_3 = R'_2] = \frac{3}{4}$. Clearly, both R and R' are tree-structured.

We can calculate (Claim 7.7) that $D(R \parallel R') = O(\varepsilon)$. Hence, by the same argument as in the proof of Lemma 7.1, the total variation distance between R^k and R'^k is less than $1/3$ if $k = O(1/\varepsilon)$, implying (i).

For part (ii), consider graphical models over nodes X, Y, and Z, representing the three coordinates of $\{0,1\}^3$. Let G be the tree Y—X—Z, and H be the tree X—Y—Z. Note that R is H-structured, and R' is G-structured. Hence, by Corollary 3.4, $D(R \parallel R_G) = I(R_1; R_3) - I(R_2; R_3)$ where R_G is the closest G-structured distribution to R. Applying Claim 7.8 below, $D(R \parallel R_G) = \Omega(\varepsilon)$. The same arguments apply to $D(R' \parallel R'_H)$. Finally, it is easy to check that the tree Y—Z—X is not $O(\varepsilon)$-approximate for either R or R'.

Claim 7.7.

$$D(R \parallel R') = O(\varepsilon).$$

Proof. Note that $\mathbf{Pr}[R = 010] = \mathbf{Pr}[R = 101] = \frac{3\varepsilon}{8}$ and $\mathbf{Pr}[R = 011] = \mathbf{Pr}[R = 100] = \frac{\varepsilon}{8}$, while $\mathbf{Pr}[R' = 010] = \mathbf{Pr}[R' = 101] = \frac{\varepsilon}{8}$ and $\mathbf{Pr}[R' = 011] = \mathbf{Pr}[R' = 100] = \frac{\varepsilon}{8}$. For all other assignments, R and R' have the same value. Hence, $D(R \parallel R') = \frac{3\varepsilon}{8} \log 3 + \frac{\varepsilon}{8} \log \frac{1}{3} = \frac{\varepsilon}{2} \log 3$. □

Claim 7.8.

$$I(R_1; R_3) - I(R_2; R_3) = \Omega(\varepsilon).$$

Proof. Each of R_1, R_2, R_3 is an unbiased bit. Also, $\mathbf{Pr}[R_3 = R_2] = (1-\varepsilon) \cdot \frac{3}{4} + \varepsilon \cdot \frac{1}{4} = \frac{3}{4} - \frac{\varepsilon}{2}$. Hence:

$$\begin{aligned} I(R_1; R_3) - I(R_2; R_3) &= -H(R_3 \mid R_1) + H(R_3 \mid R_2) \\ &= -\left(\frac{3}{4}\log\frac{4}{3} + \frac{1}{4}\log 4\right) + \left(\left(\frac{3}{4} - \frac{\varepsilon}{2}\right)\log\frac{1}{\frac{3}{4}-\frac{\varepsilon}{2}} + \left(\frac{1}{4} + \frac{\varepsilon}{2}\right)\log\frac{1}{\frac{1}{4}+\frac{\varepsilon}{2}}\right) \end{aligned}$$

$$= \frac{\varepsilon}{2} \log 3 - O(\varepsilon^2)$$

□

□

Acknowledgements

AB and SG thank Rishi Gajjala for contributing to related discussions and Sanjoy Dasgupta for clarifying some questions about [Das97]. Also, EP thanks Alex Dimakis for pointing us to [DP20].

References

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3591–3599, 2015. 5

[AHHK12] Anima Anandkumar, Daniel J Hsu, Furong Huang, and Sham M Kakade. Learning mixtures of tree graphical models. In *Advances in Neural Information Processing Systems*, pages 1052–1060, 2012. 4

[AK01] András Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, 2001. 17

[AKN06] Pieter Abbeel, Daphne Koller, and Andrew Y Ng. Learning factor graphs in polynomial time and sample complexity. *Journal of Machine Learning Research*, 7(Aug):1743–1788, 2006. 4

[BCD20] Johannes Brustle, Yang Cai, and Constantinos Daskalakis. Multi-item mechanisms without item-independence: Learnability via robustness. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 715–761, 2020. 4, 8

[BFF+01] Tugkan Batu, Lance Fortnow, Eldar Fischer, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *42nd Annual Symposium on Foundations of Computer Science, FOCS 2001, 14-17 October 2001, Las Vegas, Nevada, USA*, pages 442–451. IEEE Computer Society, 2001. 5

[BGMV20] Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, and N. V. Vinodchandran. Efficient distance approximation for structured high-dimensional distributions via learning. *CoRR*, abs/2002.05378, 2020. 5

[BK20] Guy Bresler and Mina Karzand. Learning a tree-structured ising model in order to make predictions. *Ann. Statist.*, 48(2):713–737, 04 2020. 4

[BMS13] Guy Bresler, Elchanan Mossel, and Allan Sly. Reconstruction of markov random fields from samples: Some observations and algorithms. *SIAM Journal on Computing*, 42(2):563–578, 2013. 4

[Bre15] Guy Bresler. Efficiently learning ising models on arbitrary graphs. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 771–782, 2015. 4

[Can15] Clément L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, 2015. 5

[Can20] Clément L. Canonne, Jan 2020. Personal communication. 5

[CDKS18] Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing conditional independence of discrete distributions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018*, pages 735–748. ACM, 2018. 3, 5

[CDKS20] Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing bayesian networks. *IEEE Trans. Inf. Theory*, 66(5):3132–3170, 2020. 5

[CG08] Anton Chechetka and Carlos Guestrin. Efficient principled learning of thin junction trees. In *Advances in Neural Information Processing Systems*, pages 273–280, 2008. 4

[Chi95] David Maxwell Chickering. Learning bayesian networks is np-complete. In Doug Fisher and Hans-Joachim Lenz, editors, *Learning from Data - Fifth International Workshop on Artificial Intelligence and Statistics, AISTATS 1995, Key West, Florida, USA, January, 1995. Proceedings*, pages 121–130. Springer, 1995. 4

[CL68] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theory*, 14(3):462–467, 1968. 2, 5, 8, 9, 10

[CW73] C Chow and T Wagner. Consistency of an estimate of tree-dependent probability distributions. *IEEE Transactions on Information Theory*, 19(3):369–371, 1973. 2, 4

[Das97] Sanjoy Dasgupta. The sample complexity of learning fixed-structure bayesian networks. *Mach. Learn.*, 29(2-3):165–180, 1997. 5, 8, 9, 23

[Das13] Sanjoy Dasgupta. Learning polytrees. *CoRR*, abs/1301.6688, 2013. 4

[DDK19] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing ising models. *IEEE Trans. Inf. Theory*, 65(11):6829–6852, 2019. 3, 5

[DGPP18] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018. 8

[DL97] Paul Dagum and Michael Luby. An optimal approximation algorithm for bayesian inference. *Artificial Intelligence*, 93(1):1–28, 1997. 4

[DMR20] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The minimax learning rates of normal and ising undirected graphical models. *Electronic Journal of Statistics*, 14(1):2338–2361, 2020. 5

[DP17] Constantinos Daskalakis and Qinxuan Pan. Square hellinger subadditivity for bayesian networks and its applications to identity testing. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 697–703. PMLR, 2017. 5

[DP20] Constantinos Daskalakis and Qinxuan Pan. Tree-structured ising models can be learned efficiently. *CoRR*, abs/2010.14864, 2020. 5, 23

[Goe20] Surbhi Goel. Learning ising and potts models with latent variables. volume 108 of *Proceedings of Machine Learning Research*, pages 3557–3566, Online, 26–28 Aug 2020. PMLR. 4

[Gol17] Oded Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017. 5

[GR11] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, pages 68–75. Springer, 2011. 5

[Höf93] Klaus-U Höffgen. Learning and robust learning of product distributions. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 77–83, 1993. 4

[KF09] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 2

[KFL01] Frank R Kschischang, Brendan J Frey, and H-A Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on information theory*, 47(2):498–519, 2001. 2

[KM17] Adam R. Klivans and Raghu Meka. Learning graphical models using multiplicative weights. In Chris Umans, editor, *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 343–354. IEEE Computer Society, 2017. 4

[KOPS15] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, 2015. 4, 5, 8, 19

[KS94] Michael J Kearns and Robert E Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994. 2

[KS01] David R. Karger and Nathan Srebro. Learning markov networks: maximum bounded tree-width graphs. In S. Rao Kosaraju, editor, *Proceedings of the Twelfth Annual Symposium on Discrete Algorithms, January 7-9, 2001, Washington, DC, USA*, pages 392–401. ACM/SIAM, 2001. 4

[KSS94] Michael J Kearns, Robert E Schapire, and Linda M Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994. 2

[Lap95] PS Laplace. Philosophical essays on probabilities, from 5th french edition published 1825, translated ai dale, 1995. 3

[Lau96] Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996. 2

[LXG+11] Han Liu, Min Xu, Haijie Gu, Anupam Gupta, John Lafferty, and Larry Wasserman. Forest density estimation. *The Journal of Machine Learning Research*, 12:907–951, 2011. 4

[Mee01] Christopher Meek. Finding a path is harder than finding a tree. *Journal of Artificial Intelligence Research*, 15:383–389, 2001. 4

[Mei99] Marina Meila. An accelerated chow and liu algorithm: Fitting tree distributions to high-dimensional sparse data. In Ivan Bratko and Saso Dzeroski, editors, *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999), Bled, Slovenia, June 27 - 30, 1999*, pages 249–257. Morgan Kaufmann, 1999. 4

[MJ00] Marina Meila and Michael I Jordan. Learning with mixtures of trees. *Journal of Machine Learning Research*, 1(Oct):1–48, 2000. 4

[NB04] Mukund Narasimhan and Jeff Bilmes. Pac-learning bounded tree-width graphical models. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 410–417, 2004. 4

[Pan03] Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003. 17

[Rub12] Ronitt Rubinfeld. Taming big probability distributions. *ACM Crossroads*, 19(1):24–28, 2012. 5

[Sre03] Nathan Srebro. Maximum likelihood bounded tree-width markov networks. *Artificial intelligence*, 143(1):123–138, 2003. 4

[TATW11] Vincent YF Tan, Animashree Anandkumar, Lang Tong, and Alan S Willsky. A large-deviation analysis of the maximum-likelihood learning of markov tree structures. *IEEE Transactions on Information Theory*, 57(3):1714–1735, 2011. 2, 4

[TTZ20] Anshoo Tandon, Vincent YF Tan, and Shiyao Zhu. Exact asymptotics for learning tree-structured graphical models with side information: Noiseless and noisy samples. *arXiv preprint arXiv:2005.04354*, 2020. 2, 4

[Val84] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984. 2

[VP90] Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *UAI '90: Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence, 1990*, pages 255–270. Elsevier, 1990. 9

[Wai06] Martin J Wainwright. Estimating the"wrong" graphical model: Benefits in the computation-limited setting. *Journal of Machine Learning Research*, 7(Sep):1829–1859, 2006. 4

[WJ08] Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008. 2

[WSD19] Shanshan Wu, Sujay Sanghavi, and Alexandros G Dimakis. Sparse logistic regression learns all discrete pairwise graphical models. In *Advances in Neural Information Processing Systems*, pages 8071–8081, 2019. 4

[WSN13] Rui Wu, R Srikant, and Jian Ni. Learning loosely connected markov random fields. *Stochastic Systems*, 3(2):362–404, 2013. 4

A Proofs of Background

Proof of Lemma 3.3. We have that

$$\begin{aligned}\sum_{v\in V} I(X_v; X_{\mathrm{pa}(v)}) &= \sum_{v\in V} \big(H(X_v) - H(X_v \mid X_{\mathrm{pa}(v)})\big) \\ &= \sum_{v\in V} H(X_v) + \sum_{x\in\Sigma^n} \mathbf{Pr}[X = x] \sum_{v\in V} \log \mathbf{Pr}[X_v = x_v \mid X_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}].\end{aligned}$$

Therefore

$$\begin{aligned}D(P \parallel Q) &= \sum_{x\in\Sigma^n} \mathbf{Pr}[X = x] \log \frac{\mathbf{Pr}[X = x]}{\mathbf{Pr}[X' = x]} \\ &= -H(X) + \sum_{x\in\Sigma^n} \mathbf{Pr}[X = x] \log \frac{1}{\mathbf{Pr}[X' = x]} \\ &= -H(X) + \sum_{x\in\Sigma^n} \mathbf{Pr}[X = x] \sum_{v\in V} \log \frac{1}{\mathbf{Pr}[X'_v = x_v \mid X'_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}]} \\ &= -H(X) + \sum_{v\in V} H(X_v) - \sum_{v\in V} I(X_v; X_{\mathrm{pa}(v)}) \\ &\quad + \sum_{x\in\Sigma^n} \mathbf{Pr}[X = x] \sum_{v\in V} \log \frac{\mathbf{Pr}[X_v = x_v \mid X_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}]}{\mathbf{Pr}[X'_v = x_v \mid X'_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}]} \\ &= -H(X) + \sum_{v\in V} H(X_v) - \sum_{v\in V} I(X_v; X_{\mathrm{pa}(v)}) \\ &\quad + \sum_{v\in V} \sum_{x\in\Sigma^2} \mathbf{Pr}[(X_{\mathrm{pa}(v)}, X_v) = x] \log \frac{\mathbf{Pr}[X_v = x_v \mid X_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}]}{\mathbf{Pr}[X'_v = x_v \mid X'_{\mathrm{pa}(v)} = x_{\mathrm{pa}(v)}]}\end{aligned}$$

which is the desired bound. □

B Proof of Claim 7.3

Proof of Claim 7.3. The following tables show the marginal distribution on R_1, R_2.

R_1	R_2	B	value
0	0	if $B=0$	$\frac{1}{2}\left[\left((0.75+\varepsilon)+(0.25-\varepsilon)\frac{1}{2}\right)\left((0.75+\varepsilon)+(0.25-\varepsilon)\frac{1}{2}\right)\right]$
		if $B=1$	$\frac{1}{2}\left[(0.25-\varepsilon)\frac{1}{2}(0.25-\varepsilon)\frac{1}{2}\right]$
1	1	if $B=0$	$\frac{1}{2}\left[(0.25-\varepsilon)\frac{1}{2}(0.25-\varepsilon)\frac{1}{2}\right]$
		if $B=1$	$\frac{1}{2}\left[\left((0.75+\varepsilon)+(0.25-\varepsilon)\frac{1}{2}\right)\left((0.75+\varepsilon)+(0.25-\varepsilon)\frac{1}{2}\right)\right]$
0	1	if $B=0$	$\frac{1}{2}\left[\left((0.75+\varepsilon)+(0.25-\varepsilon)\frac{1}{2}\right)(0.25-\varepsilon)\frac{1}{2}\right]$
		if $B=1$	$\frac{1}{2}\left[(0.25-\varepsilon)\frac{1}{2}\left((0.75+\varepsilon)+(0.25-\varepsilon)\frac{1}{2}\right)\right]$
1	0	if $B=0$	$\frac{1}{2}\left[(0.25-\varepsilon)\frac{1}{2}\left((0.75+\varepsilon)+(0.25-\varepsilon)\frac{1}{2}\right)\right]$
		if $B=1$	$\frac{1}{2}\left[\left((0.75+\varepsilon)+(0.25-\varepsilon)\frac{1}{2}\right)(0.25-\varepsilon)\frac{1}{2}\right]$

Simplifying, we get:

R_1	R_2	value
0	0	$\frac{1}{2}\left[\left(\frac{7}{8}+\frac{\varepsilon}{2}\right)^2+\left(\frac{1}{8}-\frac{\varepsilon}{2}\right)^2\right]$
1	1	$\frac{1}{2}\left[\left(\frac{7}{8}+\frac{\varepsilon}{2}\right)^2+\left(\frac{1}{8}-\frac{\varepsilon}{2}\right)^2\right]$
0	1	$\frac{1}{2}\left[2\left(\frac{7}{8}+\frac{\varepsilon}{2}\right)\left(\frac{1}{8}-\frac{\varepsilon}{2}\right)\right]$
1	0	$\frac{1}{2}\left[2\left(\frac{7}{8}+\frac{\varepsilon}{2}\right)\left(\frac{1}{8}-\frac{\varepsilon}{2}\right)\right]$

Now,

$$\frac{1}{2}\left[\left(\frac{7}{8}+\frac{\varepsilon}{2}\right)^2+\left(\frac{1}{8}-\frac{\varepsilon}{2}\right)^2\right]=\frac{25}{64}+\frac{\varepsilon^2}{4}+\frac{3\varepsilon}{8}\approx\frac{25}{64}\left(1+\frac{24\varepsilon}{25}\right)$$

$$\left(\frac{7}{8}+\frac{\varepsilon}{2}\right)\left(\frac{1}{8}-\frac{\varepsilon}{2}\right)=\frac{7}{64}+\frac{\varepsilon^2}{4}+\frac{\varepsilon}{2}\approx\frac{7}{64}\left(1+\frac{32\varepsilon}{7}\right)$$

Hence,

$$H(R_1,R_2)\approx 2\left[\frac{25}{64}\left(1+\frac{24\varepsilon}{25}\right)\log\frac{64}{25\left(\frac{1+24\varepsilon}{25}\right)}+\frac{7}{64}\left(1+\frac{32\varepsilon}{7}\right)\log\frac{64}{7\left(\frac{1+32\varepsilon}{7}\right)}\right] \tag{B.1}$$

$$\approx 2\left[\frac{25}{64}\left(1+\frac{24\varepsilon}{25}\right)\left(\log\frac{64}{25}-\log\left(1+\frac{24\varepsilon}{25}\right)\right)+\frac{7}{64}\left(1+\frac{32\varepsilon}{7}\right)\left(\log\frac{64}{7}-\log\left(1+\frac{32\varepsilon}{7}\right)\right)\right]$$

(using $\log(1+x)\approx x$)

$H(R_1,R_2)=C_{12}$ (a constant term corresponding to $\varepsilon=0$) $+1.17\varepsilon+O(\varepsilon^2)$

The following tables show the marginal distribution on R_1,R_3.

R_1	R_3	B	value
0	0	if $B=0$	$\frac{1}{2}\left[\left((0.75+\varepsilon)+(0.25-\varepsilon)\frac{1}{2}\right)\left((0.75-\varepsilon)+(0.25+\varepsilon)\frac{1}{2}\right)\right]$
		if $B=1$	$\frac{1}{2}\left[(0.25-\varepsilon)\frac{1}{2}(0.25+\varepsilon)\frac{1}{2}\right]$
1	1	if $B=0$	$\frac{1}{2}\left[(0.25-\varepsilon)\frac{1}{2}(0.25+\varepsilon)\frac{1}{2}\right]$
		if $B=1$	$\frac{1}{2}\left[\left((0.75+\varepsilon)+(0.25-\varepsilon)\frac{1}{2}\right)\left((0.75-\varepsilon)+(0.25+\varepsilon)\frac{1}{2}\right)\right]$
1	0	if $B=0$	$\frac{1}{2}\left[(0.25-\varepsilon)\frac{1}{2}\left((0.75-\varepsilon)+(0.25+\varepsilon)\frac{1}{2}\right)\right]$
		if $B=1$	$\frac{1}{2}\left[\left((0.75+\varepsilon)+(0.25-\varepsilon)\frac{1}{2}\right)(0.25+\varepsilon)\frac{1}{2}\right]$
0	1	if $B=0$	$\frac{1}{2}\left[\left((0.75+\varepsilon)+(0.25-\varepsilon)\frac{1}{2}\right)(0.25+\varepsilon)\frac{1}{2}\right]$
		if $B=1$	$\frac{1}{2}\left[(0.25-\varepsilon)\frac{1}{2}\left((0.75-\varepsilon)+(0.25+\varepsilon)\frac{1}{2}\right)\right]$

Simplifying, we get:

R_1	R_3	value
0	0	$\frac{25}{64} - \frac{\varepsilon^2}{4}$
1	1	$\frac{25}{64} - \frac{\varepsilon^2}{4}$
0	1	$\frac{7}{64} + \frac{\varepsilon^2}{4}$
1	0	$\frac{7}{64} + \frac{\varepsilon^2}{4}$

Hence,

$$H(R_1, R_3) = C_{13} \text{ (a constant term corresponding to } \varepsilon = 0) + O(\varepsilon^2) \tag{B.2}$$

Finally, we have from (B.1) and (B.2),

$$\begin{aligned} I(R_1; R_2) - I(R_1; R_3) &= H(R_1, R_3) - H(R_1, R_2) \\ &\text{(since the marginal entropies are 0 due to unbiasedness)} \\ &= -1.17\varepsilon + O(\varepsilon^2) \\ &\text{(using } C_{12} = C_{13}\text{, since the three bits are identically distributed when } \varepsilon = 0) \end{aligned}$$

□