Look Ma, No Latent Variables: Accurate Cutset Networks via Compilation

Tahrima Rahman [*1] **Shasha Jin** [*1] **Vibhav Gogate** [1]

Abstract

Tractable probabilistic models obviate the need for unreliable approximate inference approaches and as a result often yield accurate query answers in practice. However, most tractable models that achieve state-of-the-art generalization performance (measured using test set likelihood score) use latent variables. Such models admit poly-time marginal (MAR) inference but do not admit poly-time (full) maximum-a-posteriori (MAP) inference. To address this problem, in this paper, we propose a novel approach for inducing cutset networks, a well-known tractable, highly interpretable representation that does not use latent variables and admits linear time MAR as well as MAP inference. Our approach addresses a major limitation of existing techniques that learn cutset networks from data in that their accuracy is quite low as compared to latent variable models such as ensembles of cutset networks and sum-product networks. The key idea in our approach is to construct deep cutset networks by not only learning them from data but also compiling them from a more accurate latent tractable model. We show experimentally that our new approach yields more accurate MAP estimates as compared with existing approaches and significantly improves the test set log-likelihood score of cutset networks bringing them closer in terms of generalization performance to latent variable models.

1. Introduction

A fundamental shortcoming of probabilistic graphical models (PGMs) (Pearl, 1988) such as Bayesian and Markov networks is that probabilistic inference—the process used to answer queries—on most models used in practice is intractable. This is not surprising since even the most basic inference task of computing the marginal probability of a variable given observations or evidence—posterior marginal estimation—is #P-hard (Roth, 1996). To circumvent this issue, one can either use approximate inference in lieu of exponential time exact inference approaches or learn models such that exact inference is tractable, namely takes time that is polynomial in the number of parameters. Both approaches are widely used in practice and each has their own pros and cons. Tractable models are desirable because the user can always recover accurate, reliable answers according to the model. This is a major plus over the approximate inference approach (e.g., sampling, belief propagation, etc.) because the latter often exhibits high variability; different runs and algorithms often yield widely different query answers. However, tractable models typically have slightly worse test set log-likelihood scores, namely they generalize poorly, as compared to arbitrary latent PGMs. Therefore improving the accuracy/fit of tractable models is an active area of research (Rahman & Gogate, 2016a;b; Di Mauro et al., 2016; 2017; Rashwan et al., 2016; Liang et al., 2017).

An often overlooked and one of the earliest goals of PGM research is learning interpretable models (Pearl, 1988; Darwiche, 2009). By interpretable models, we mean PGMs whose random variables, dependencies (structure) and parameters are interpretable. To this end, sparse Bayesian networks having no latent variables are interpretable but Markov networks are not because parameters of the latter are not interpretable. Interpretability is desirable, especially in interactive settings (Kulesza et al., 2015) and explainable AI applications (Gunning, 2017) such as activity recognition and medical diagnosis because it helps explain the model, specifically its assumptions to the user. This allows the user to update the model if the assumptions are wrong based on his/her prior knowledge. When the interpretable model is also tractable, it helps explain why a model made a particular decision as opposed to a different one to the user. Thus, learning tractable interpretable models is an important research endeavor, especially for building high quality explainable AI systems (Gunning, 2017).

In this paper, we focus on a particular class of tractable, interpretable models called cutset networks (Rahman et al., 2014). At a high level, these networks represent how a cutset conditioning method (Pearl, 1988) that takes advantage of determinism (Gogate, 2009), dynamic orderings, context-

[*]Equal contribution [1]Department of Computer Science, The University of Texas at Dallas, United States. Correspondence to: Tahrima Rahman <tahrima.rahman@utdallas.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

specific independence (Boutilier et al., 1996) and similar probability values (Chavira & Darwiche, 2008) would approximate an intractable PGM. Graphically, a cutset network is an OR tree (Dechter & Mateescu, 2007) in which each node is labeled with a random variable and each edge is labeled with the conditional probability of the child given an assignment of values from the root to the parent. The leaves of the cutset network are tree Bayesian networks defined over the variables not present on the path between the root and the leaf and therefore can be learned efficiently using the Chow-Liu algorithm (Chow & Liu, 1968).

Although simple and efficient to learn, a major shortcoming of cutset networks is that their test set log-likelihood score is worse, often by a significant amount, as compared to arbitrary PGMs. To alleviate this shortcoming, (Rahman & Gogate, 2016a; Di Mauro et al., 2016) proposed to learn mixtures of cutset networks (i.e., introduce a latent variable) by using various approaches such as the expectation maximization algorithm, Bagging, random forests and Boosting. Latent variables are also employed to improve generalization performance by other, related tractable models such as sum-product networks (SPNs) (Poon & Domingos, 2011) and ensembles of probabilistic sentential decision diagrams (Kisa et al., 2014; Liang et al., 2017) . Unfortunately, despite improved generalization performance, in presence of latent variables (full) maximum-a-posteriori (MAP) inference is no longer tractable (while marginal inference (MAR) is still tractable). As a result, application designers have to use unreliable approximate inference approaches to compute MAP estimates on these models. In other words, existing tractable models use latent variables to trade the accuracy of MAP inference with the accuracy of MAR inference.

The main contribution of this paper is to propose a new method for improving the accuracy of cutset networks without using latent variables. We propose a two-step, anytime approach. In the first step, we learn a tractable model having latent variables from data. In the second step, we take a generic, heuristic algorithm for learning the structure and parameters of cutset networks and replace the empirical statistics computed from data alone by a convex combination of the empirical statistics and the statistics computed from a latent tractable model (compilation). The key reason for using latent tractable models instead of Markov and Bayesian networks is that sufficient statistics can be inferred efficiently and accurately in the former (by performing polytime exact inference). Although, our approach is simple, it is quite powerful. In particular, we prove that the accuracy of cutset networks constructed using our approach will increase over time. We demonstrate via experiments on 20 benchmark datasets that our new method achieves test set log-likelihood scores that are only slightly inferior to state-of-the-art latent models while substantially outperforming the latter as well as existing methods that induce tractable models having no latent variables for MAP inference.

2. Related Work

Our work is related to work in the classification community by (Towell & Shavlik, 1994; Craven, 1996) who proposed to learn accurate, interpretable decision trees by overfitting a decision tree to a highly accurate deep neural network. Our work is different in that we are interested in modeling probability distributions that can solve much harder tasks such as structured prediction and posterior marginal inference. Our work is also related to the work on approximate compilation approaches (Gogate & Dechter, 2008; Lowd & Domingos, 2010; Gogate & Dechter, 2012; Friedman & Van den Broeck, 2018) that seek to construct an approximate tractable model from an intractable graphical model. The difference is that the aforementioned approaches do not use data and often learn tractable models having latent variables. Another line of work that is related to our work is work on learning tractable models without latent variables directly from data (Rahman et al., 2014; Lowd & Domingos, 2008; Rahman & Gogate, 2016b). Our work is different in that we leverage more accurate latent models to improve the performance of non-latent models.

3. Notation and Background

We assume that all random variables used in this paper are Boolean or binary valued. Note that we make this assumption for simplicity of exposition and the techniques presented in this paper can be easily extended to multi-valued random variables. Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ denote the set of n Boolean variables where each variable $X_i \in \mathbf{X}$ takes values from the domain $\{0, 1\}$. Let $\mathbf{x}$ denote an assignment of values to all variables in $\mathbf{X}$. We will denote the assignment $X_i = 1$ by x_i and $X_i = 0$ by $\overline{x_i}$. Given a subset $\mathbf{U}$ of $\mathbf{X}$, we denote by $\mathbf{x}_{\mathbf{U}}$ as the projection of $\mathbf{x}$ on $\mathbf{U}$.

3.1. Bayesian Networks

Bayesian networks (BNs) (Pearl, 1988; Darwiche, 2009) are often used in practice to represent and reason about uncertainty. At a high level, they can be understood as a compact graphical representation of a joint probability distribution over a large number of random variables. Formally, a Bayesian network, denoted by B is a triple $\langle \mathbf{X}, G, \mathbf{P} \rangle$ where $G = (\mathbf{V}, \mathbf{E})$ is a directed acyclic graph such that $\mathbf{V}$ has one node for each variable in $\mathbf{X}$, $\mathbf{E}$ is a set of directed edges, and $\mathbf{P}$ is a set of conditional probability tables (CPTs). Each CPT $P_i \in \mathbf{P}$ is defined as $P_i(X_i | pa(X_i))$ where $pa(X_i)$ is the set of parents of X_i in G. A Bayesian network represents the following probability distribution $P_B(\mathbf{x}) = \prod_{i=1}^{n} P_i(\mathbf{x}_{\{X_i\}} | \mathbf{x}_{pa(X_i)})$. Since the size of each CPT is exponential in the number of its parents, in practice,

Figure 1. An example cutset network de ned over the set of variables $\{X_1; X_2; X_3; X_4; X_5; X_6\}$.

for computational reasons, we assume that the number of parents is bounded by a constant.

The two key tasks in Bayesian networks are learning the structure and parameters from data and inference. We focus on two typical inference tasks in this paper: marginal inference (MAR) given evidence and full maximum-a-posteriori (MAP) inference[1]. It is well known that most tasks of interest in practice can be reduced to either MAR or MAP. Let the variables be partitioned into three (possibly empty) sets: evidence (or observed), non-evidence (or unobserved) and latent (or hidden). MAR is de ned as nding the marginal probability distribution over each or a small subset of non-evidence variables given evidence while MAP is de ned as nding the most probable assignment to all non-evidence (unobserved) variables given evidence. Both tasks are at least NP-hard and are computationally infeasible in practice. We say that a Bayesian network, and a probabilistic model in general, is MAR-tractable (MAP-tractable) when the MAR (MAP) task can be solved in time that scales polynomially with the number of variables. Examples of tractable Bayesian networks include tree Bayesian networks (Chow-Liu trees (Chow & Liu, 1968)), bounded treewidth Bayesian networks (Elidan & Gould, 2008) and networks having polynomial-sized arithmetic circuits (Darwiche, 2000).

3.2. Cutset Networks

Cutset networks (CNs) combine tree Bayesian networks with OR trees (probabilistic decision trees) (Dechter & Mateescu, 2007). Graphically, they can be depicted using a rooted OR tree with a tree Bayesian network at each leaf of the OR tree. Formally, a CN C is a pair $\langle O; T \rangle$ where O is an OR Tree having L leaves and T is a collection of L tree Bayesian networks attached to each of the L leaves. An OR tree is a rooted binary tree in which internal nodes are labeled with variables and each of the two edges emanating from a node represents conditioning of the variable with an appropriate value (either 0 or 1). We follow the convention that the left branch of a node labeled by X_i represents $\bar{x}_i$ while the right branch represents x_i. Each edge is labeled with the conditional probability of the variable taking the value given an assignment of values from the root to the parent node. C represents the following probability distribution

$$P_C(x) = \prod_{(v_i;v_j) \in path_O(x)} \left(p_{i;j} \; T_{l(x)}(x_{V(T_{l(x)})}) \right) \quad (1)$$

where $path_O(x)$ is a unique path (sequence of edges) from the root to the leaf in O corresponding to the assignment x, $p_{i;j}$ is the conditional probability attached to the edge between nodes v_i and v_j, $l(x)$ is the index of the leaf node along $path_O(x)$, $T_{l(x)}$ is the tree Bayesian network in T at index $l(x)$ and $V(T_{l(x)})$ denotes the subset of variables over which $T_{l(x)}$ is de ned.

Example 1. Figure 1 shows an example cutset network de ned over six variables $\{X_1; \ldots; X_6\}$. The probability of the assignment $(0; 1; 1; 1; 0; 1)$ to the six variables (where X_1 is assigned to 0, X_2 is assigned to 1, etc.) equals $0.3 \cdot 0.8 \cdot P(X_3 = 1; X_4 = 1; X_5 = 0; X_6 = 1 | X_1 = 0; X_2 = 1)$ where 0.3 equals the probability $P(X_1 = 0)$ (left branch of root node labeled by X_1), 0.8 equals the conditional probability $P(X_2 = 1 | X_1 = 0)$ (right branch of X_2) respectively and $P(X_3 = 1; X_4 = 1; X_5 = 0; X_6 = 1 | X_1 = 0; X_2 = 1)$ is computed from the tree Bayesian network T_3 (right child of the OR node labeled by X_2).

A key feature of cutset networks is that they are both MAP-tractable and MAR-tractable (Rahman et al., 2014; Dechter & Mateescu, 2007) assuming that all variables in the OR tree as well as all tree Bayesian networks T are known or observed (namely when the network has no latent variables). Other examples of tractable models that are both MAP- and MAR-tractable include bounded treewidth Bayesian networks (Bach & Jordan, 2002), Bayesian networks having polynomial sized arithmetic circuits (Lowd & Domingos, 2008) and probabilistic sentential decision diagrams (Liang et al., 2017). Another important feature of cutset networks is that their parameters have well-de ned probabilistic semantics, i.e., they are interpretable and as a result enable smooth, reliable human-machine interaction (Gunning, 2017).

CNs can be learned from data by adapting typical top-down decision tree induction techniques for selecting a variable at each OR tree node and then using the Chow-Liu algorithm (Chow & Liu, 1968) for learning a tree Bayesian network at each leaf node when a pre-de ned termination condition is satis ed. However, despite advances in learning algorithms (Vergari et al., 2015; Di Mauro et al., 2015), CNs are less accurate than arbitrary, intractable Bayesian networks.

To address this problem, (Di Mauro et al., 2015; Rahman & Gogate, 2016a;b) proposed using latent sum-product mixtures of cutset networks. In numerous empirical studies, it

[1] Full MAP inference is also called most probable explanation (MPE) inference in the Bayesian network literature.

has been shown that these networks achieve state-of-the-art test set log-likelihood scores. However, a key issue with these latent models is that they sacrifice MAP-tractability[2] while maintaining MAR-tractability.

4. Inducing Accurate Cutset Networks: A Novel Approach

The main goal of this paper is to improve the accuracy of cutset networks while maintaining both MAP and MAR tractability. Our main intuition is the following. Improving the model fit, measured using test set log-likelihood score, of MAR tractable models also substantially improves their MAR inference accuracy. In particular, numerous previous studies (Rooshenas & Lowd, 2014; Gens & Domingos, 2013) have shown that given two models, say M_1 (e.g., SPNs) and M_2 (e.g., BNs), such that: (1) M_1 is MAR-tractable; (2) M_2 is MAR-intractable; and (3) M_2 is slightly better than M_1 in terms of model fit, posterior marginal estimates obtained by performing exact inference on M_1 are far superior to those obtained by performing approximate inference on M_2 when both M_1 and M_2 are given the same amount of (reasonably bounded) time. We hope to achieve similar results for MAP inference, namely by improving the fit of MAP-tractable cutset networks (i.e., networks without latent variables), we hope to significantly improve the quality of their MAP estimates.

Next, we present our new method for learning accurate cutset networks. The key idea is to first learn a highly accurate tractable model with latent variables from data. We will denote this model by Q. We then combine statistics computed by performing exact marginal inference on Q with the ones empirically estimated from data to induce a cutset network. The main intuition behind this new approach is that we are unable to learn deep, accurate cutset networks in high dimensions from data alone because the variance increases with increasing depth. In particular, as we increase the depth d of a cutset network by one, the sufficient statistics used to estimate the remaining network at depth $d+1$ are not as reliable as the ones at depth d because the former are based on roughly half the number of training examples as compared to the latter. Therefore, to improve the accuracy of deep cutset networks we need to reduce the variance using an artifact other than the training data. In this paper, we propose to use Q, an auxiliary latent tractable model to achieve this objective.

Algorithm 1 describes the main steps in our approach. We call the algorithm LC-CN which stands for Learn/Compile Cutset Networks. The algorithm takes as input a dataset

Algorithm 1 LC-CN (D, Q)

Input : Training examples D defined over a set of variables X and a tractable latent model representing a distribution Q
Output : A Cutset network
1 **begin**
2 Compute pairwise marginal distribution $P(X_i, X_j)$ for all pairs (X_i, X_j) from Q and D (see Eq. (4)).
3 if the termination condition is satisfied **then**
4 return ChowLiuTree(P)
5 **else**
6 // Variable selection Heuristic
7 Use P to compute $Score(X)$ for each variable X using Eq. (3).
8 X_i = variable with the highest Score.
9 Create a new internal node o labeled by X_i
10 Let l and r be the left and right child nodes of o respectively.
11 $Label(o, l) = P(\overline{x_i})$
12 l = LC-CN($D|\overline{x_i}$, $Q|\overline{x_i}$).
13 $Label(o, r) = P(x_i)$
14 r = LC-CN($D|x_i$, $Q|x_i$).
15 return o
16 **end**

D defined over a collection of observed variables X and a tractable latent model representing a distribution Q (learned from D). The algorithm builds a cutset network via a top-down decision-tree style induction. At each recursive call, it outputs a Chow-Liu tree (steps 2-3) if the termination condition is met. Otherwise, it heuristically selects a variable X_i to condition on (steps 6-7) and then recurses on 0 and 1 value assignments to X_i (steps 9-13).

Variable Selection Heuristic. Following previous work (Rahman et al., 2014; Di Mauro et al., 2015), we propose to use pairwise mutual information score to select the variable X_i in steps 6 and 7. This scoring function is based on the following intuition. Ideally, we should condition on a variable having strong dependencies with other variables, since conditioning on that variable would help us reach the termination condition faster. In other words, the conditioning operation is highly likely to yield a sparse, tree-like Bayesian network having fewer dependencies (edges). A popular approach for measuring dependence between two sets of variables is the mutual information score:

$$I(X;Y) = \sum_x \sum_y P(x;y) \log \frac{P(x;y)}{P(x)P(y)} \qquad (2)$$

However, estimating mutual information between a variable X_i and the remaining variables $X \setminus \{X_i\}$ is computationally intractable because we need to sum over exponential number of combinations. Therefore, we approximate it using

[2] MAP inference in latent models is also called marginal MAP (MMAP) inference. MMAP is substantially harder because the latent variables need to be marginalized out before performing maximization (Liu & Ihler, 2013; Park & Darwiche, 2004).

the following, computationally efficient, pairwise mutual information score and select a variable having the largest score (breaking ties arbitrarily).

$$\text{Score}(X_i) = \sum_{j : j \neq i} I(X_i; X_j) \tag{3}$$

A key sub-step in computing $\text{Score}(X_i)$ is computing the marginal distribution $P(X_i; X_j)$ for all pairs of variables. When cutset networks are learned from data, we estimate $P(X_i; X_j)$ from data. Namely, counts of $(X_i = a; X_j = b)$ where $a, b \in \{0, 1\}$ in the data is the sufficient statistic. Since these counts will exhibit high variance as the depth increases, we propose to estimate $P(X_i; X_j)$ by combining the counts with estimates computed from the latent tractable model. More formally, we use

$$P(X_i; X_j) = \alpha\, Q(X_i; X_j) + (1 - \alpha)\, S(X_i; X_j) \tag{4}$$

where $\alpha \in [0, 1]$ is a constant (hyper-parameter), $Q(X_i; X_j)$ is computed from the distribution Q represented by the latent tractable model and $S(X_i; X_j)$ is the empirical distribution computed from the data counts. In general, computing $Q(X_i; X_j)$ is NP-hard. However, on tractable models, computing $Q(X_i; X_j)$ is poly-time and thus efficient. The hyper-parameter α controls the relative importance of S and Q. When $\alpha = 0$, P equals S and the latent model is not used. Similarly, when $\alpha = 1$, P equals Q and the data is ignored. In other words, the cutset network is learned from data when $\alpha = 0$ and compiled from Q when $\alpha = 1$. In practice, α can be set using the validation set.

Note that the Chow-Liu algorithm also uses pairwise mutual information to construct a tree Bayesian network. We propose to use $P(X_i; X_j)$ (see Eq(4)) to compute pairwise mutual information for use in the Chow-Liu algorithm.

After selecting a variable X_i having the highest score, the algorithm conditions on it by creating an internal node in the OR tree (step 8) and then recursively builds its left and right subtrees (steps 10–13). In the algorithm, the notation $D|x_i$ (similarly $D|\bar{x}_i$) denotes the dataset obtained by deleting all examples in which X_i equals 0 (similarly 1) and then removing the column for X_i. Also, the notation $Q|x_i$ (similarly $Q|\bar{x}_i$) denotes the conditional distribution obtained by setting X_i to 1 (similarly 0) in Q. Finally in step 14, the algorithm returns the new internal node constructed in steps 6-13.

Termination Condition. Algorithm 1 can be easily modified to yield an anytime scheme by carefully setting the termination condition. In particular, we can perform an iterative deepening search by using a parameter and progressively increasing or decreasing the value of the parameter until a user-defined time bound is reached. One option is to use a bound on the maximum depth of the OR tree as a parameter and terminate the recursion when the bound is reached. A second option is to stop the recursion when the KL divergence between Q and P at a particular node is bounded by a small constant. In our experiments, we use the former approach.

Theoretical Properties of LC-CN. Next we formally describe the anytime properties of Algorithm 1 in Theorem 1. Specifically, the theorem shows that under the assumption that the latent tractable model (its distribution Q) is a more accurate approximation of the data generating distribution than the cutset network learned from data, the accuracy of the cutset network output by LC-CN can only increase or remain the same with increasing depth.

Theorem 1. Let $\alpha = 1$ and R_d denote the distribution associated with the cutset network output by Algorithm LC-CN having maximum depth d, then $D(Q; R_d) \geq D(Q; R_j)$ where $j \geq d$ and $D(Q; R)$ denotes the KL divergence between the distributions Q and R.

Proof of Theorem 1 is presented in the supplementary material. Theorem 1 shows that learning deep cutset networks is a good idea because as the depth increases their performance will approach that of a superior latent model Q.

We conclude this section by showing that each recursive step of LC-CN has only polynomial time complexity. In other words, LC-CN is a general-purpose, scalable algorithm for learning cutset networks.

Proposition 1. The time complexity of Algorithm LC-CN is $O(n^2 \cdot V \cdot (I + N))$ where n is the number of variables, V is the number of nodes in the cutset network, I is the inference complexity to compute $Q(X_i; X_j)$, and N is the number of training examples.

5. Experiments

5.1. Setup

We evaluated the performance of cutset networks along two dimensions: (1) Model fit measured using the test set log-likelihood score and (2) MAP estimation quality measured using the log-likelihood of the MAP assignment. We used 20 benchmark datasets used in numerous prior studies (Rooshenas & Lowd, 2014; 2013; Gens & Domingos, 2013; Vergari et al., 2015) to evaluate our new algorithm. All the datasets are defined over binary valued variables with the number of variables ranging from 16 to 1556 (see Table 1).

We used mixtures of Chow-Liu trees (MTs) (Meila & Jordan, 2000) and bags of cutset networks (BCNs) as our choice of latent tractable models $Q(\cdot)$ in Algorithm 1. Although any tractable model could have been used in Algorithm 1, for example SPNs, we chose MTs and BCNs because they admit faster inference and learning algorithms and on many datasets are as accurate as state-of-the-art methods. Moreover, since the complexity of each recursive call to our