Total Variation Distance Meets Probabilistic Inference

Arnab Bhattacharyya [1] **Sutanu Gayen** [2] **Kuldeep S. Meel** [3] **Dimitrios Myrisiotis** [4] **A. Pavan** [5] **N. V. Vinodchandran** [6]

Abstract

In this paper, we establish a novel connection between total variation (TV) distance estimation and probabilistic inference. In particular, we present an efficient, structure-preserving reduction from relative approximation of TV distance to probabilistic inference over directed graphical models. This reduction leads to a fully polynomial randomized approximation scheme (FPRAS) for estimating TV distances between same-structure distributions over any class of Bayes nets for which there is an efficient probabilistic inference algorithm. In particular, it leads to an FPRAS for estimating TV distances between distributions that are defined over a common Bayes net of small treewidth. Prior to this work, such approximation schemes only existed for estimating TV distances between product distributions. Our approach employs a new notion of *partial* couplings of high-dimensional distributions, which might be of independent interest.

1. Introduction

Substantial research has been devoted to developing models that represent high-dimensional probability distributions succinctly. One prevalent approach is through graphical models. In a graphical model, a graph describes the conditional dependencies among variables and the probability distribution is factorized according to the adjacency relationships in the graph (Koller & Friedman, 2009). When the underlying graph is a directed graph, the model is known as a Bayesian network or Bayes net.

Two fundamental computational tasks on distributions are *distance computation* and *probabilistic inference*. In this work, we establish a novel connection between these two seemingly different computational tasks. Using this connection, we design new relative error approximation algorithms for estimating the statistical distance between Bayes net distributions with small treewidth.

Total Variation Distance Computation. The distance computation problem is the following: Given descriptions of two probability distributions P and Q, compute $\rho(P, Q)$ for a distance measure ρ. A distance measure of central importance is the *total variation (TV) distance* (also known as the *statistical distance*). Let P and Q be distributions over a finite domain $\mathcal{D}$. The total variation distance between P and Q, denoted by $d_{\mathrm{TV}}(P, Q)$, is defined as

$$d_{\mathrm{TV}}(P, Q) = \max_{S \subseteq \mathcal{D}} (P(S) - Q(S)) \,.$$

The total variation distance satisfies many basic properties which makes it a versatile and fundamental measure for quantifying the dissimilarity between probability distributions. First, it has an explicit probabilistic interpretation: The TV distance between two distributions is the maximum gap between the probabilities assigned to a single event by the two distributions. Second, it satisfies many mathematically desirable properties: It is bounded and lies in $[0, 1]$, it is a metric, and it is invariant with respect to bijections. Total variation distance also measures the minimum probability that $X \neq Y$ among all couplings (X, Y) between P and Q. Because of these reasons, the total variation distance is a central distance measure employed in a wide range of areas including probability and statistics (Mitzenmacher & Upfal, 2005), machine learning (Shalev-Shwartz & Ben-David, 2014), information theory (Cover & Thomas, 2006), cryptography (Stinson, 1995), data privacy (Dwork, 2006), and pseudorandomness (Vadhan, 2012).

Probabilistic Inference. Probabilistic inference in graphical models is a fundamental computational task with a wide range of applications that spans disciplines including statistics, machine learning, and artificial intelligence

The author list has been sorted alphabetically by last name; this order should not be used to determine the extent of authors' contributions. [1]School of Computing, National University of Singapore, Singapore [2]Department of Computer Science and Engineering, Indian Institute of Technology Kanpur, India [3]Department of Computer Science, University of Toronto, Canada [4]CNRS@CREATE LTD., Singapore [5]Department of Computer Science, Iowa State University, USA [6]School of Computing, University of Nebraska - Lincoln, USA. Correspondence to: Sutanu Gayen <sutanu@cse.iitk.ac.in>, Dimitrios Myrisiotis <dimitrios.myrisiotis@cnrsatcreate.sg>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

(e.g., see Wainwright et al. (2008)). Various algorithms have been proposed for this problem, encompassing both exact approaches like message passing (Pearl, 1988), variable elimination (Dechter, 1999), and junction-tree propagation (Lauritzen & Spiegelhalter, 1988), as well as approximate techniques such as loopy belief propagation, variational inference-based methods (Wainwright et al., 2008), and particle-based algorithms (refer to Chapter 13 of Koller & Friedman (2009) and the references therein). In this work, we rely on the following formulation of probabilistic inference: Given (a representation of) random variables $X_1, \ldots, X_n$ and (a representation of) sets $S_1, \ldots, S_n$ such that for all i the set S_i is a subset of the range of X_i, compute

$$\mathbf{Pr}[X_1 \in S_1, \ldots, X_n \in S_n] .$$

1.1. Our Contributions

The problems of total variation distance computation and probabilistic inference have been studied for nearly four decades on their own, but there is no known relationship between these two fundamental yet seemingly different computational tasks. The primary goal of our paper is to initiate an investigation to determine such a relationship. Surprisingly, we demonstrate that there is a *structure-preserving* reduction from the TV distance estimation problem to the probabilistic inference problem over Bayes nets: In particular, we exhibit an efficient probabilistic reduction that, given two Bayes nets P and Q defined over a directed acyclic graph (DAG) G, makes probabilistic inference queries to a Bayes net $\mathcal{L}$ defined over the *same* DAG G and returns a relative approximation of $d_{\mathrm{TV}}(P, Q)$.

Theorem 1.1 (Informal)**.** *There is a polynomial-time randomized algorithm that takes a DAG G, two Bayes nets P and Q over G, and parameters ε, δ as inputs and behaves as follows. The algorithm makes probabilistic inference oracle queries to a Bayes net over the same DAG G and outputs a $(1+\varepsilon)$-relative approximation of $d_{\mathrm{TV}}(P, Q)$ with probability at least $1-\delta$.*

It is known that probabilistic inference computation over Bayes nets in general is a #P-hard problem and hence exact d_{TV} computation reduces to probabilistic inference over Bayes nets (Cooper, 1990). However, a salient feature of our reduction is that it *preserves the structure* of the Bayes net. Note that exact d_{TV} computation is #P-complete even for product distributions for which inference computation is straightforward (Bhattacharyya et al., 2023).

A conceptual contribution of our work is the introduction of a new notion of *partial coupling* between two probability distributions, which is a relaxation of the classical notion of coupling (Definition 3.2). Specifically, we illustrate that while establishing a computationally efficient coupling for distributions such as Bayesian networks may not be possible, it is possible to define a computationally efficient *partial coupling*. Remarkably, we show that a *partial coupling* is adequate for approximating the total variation distance. The technique of coupling, introduced by Doeblin in 1938 (Doeblin, 1938), has been fundamental in the realms of computer science and statistics for over four decades, underpinning some of the most seminal results (Lindvall, 2002; Levin et al., 2006; Meyn & Tweedie, 2012). In a similar vein, we believe the notion of *partial coupling* possesses the potential to become an essential tool in the toolkit of these domains.

The aforementioned reduction from total variation distance and probabilistic inference leads to efficient d_{TV} estimation algorithms for any class of Bayes nets that admits efficient probabilistic inference algorithms. In particular, it leads to the first polynomial-time randomized approximation scheme for calculating the total variation distance between two Bayes nets of treewidth $O(\log n)$, since the well-known variable elimination algorithm can be used for efficient probabilistic inference for such Bayes nets.

Theorem 1.2 (Informal)**.** *There is an FPRAS for estimating the TV distance between two Bayes nets of treewidth $O(\log n)$ that are defined over the same DAG of n nodes.*

Prior to our work, such approximation schemes were known only for product distributions, which are Bayes nets over a graph with no edges (Feng et al., 2023). In particular, designing an FPRAS for estimating TV distance between Bayes nets over trees (which are graphs with treewidth 1) was an open question. Our result resolves this open question. In fact, Theorem 1.2 shows that it is indeed possible to obtain an FPRAS for a large class of Bayes nets, namely Bayes nets of $O(\log n)$ treewidth.

Note that the setting of Theorem 1.2 (whereby the Bayes net distributions considered are over the same DAG) is practically relevant where one learns parameters of a Bayes net for a fixed structure from different batches and a natural question is whether the two models are close to each other or not. The TV distance-based approaches have played a significant role in the testing and improvement of constrained samplers (Golia et al., 2021).

Our next set of results focuses on the case when one of the distributions is the uniform distribution. We first prove that the exact computation of the TV distance between a Bayes net distribution and the uniform distribution is #P-complete.

Theorem 1.3. *It is #P-complete to exactly compute the TV distance between a Bayes net that has bounded in-degree and the uniform distribution.*

To complement this result, we show that there is an FPRAS that estimates the TV distance between the uniform distribution and *any* Bayes net distribution.

Theorem 1.4 (Informal)**.** *There is an FPRAS for estimating the TV distance between a Bayes net and the uniform distribution.*

1.2. Related Work

Koller and Friedman (Koller & Friedman, 2009) provide a comprehensive overview of probabilistic graphical models (such as Bayes nets).

TV Distance Computation. Recently, Bhattacharyya et al. (2023) initiated the study of the computational complexity aspects of TV distance over graphical models. In that work, they proved that exactly computing the TV distance between product distributions is #P-complete, that it is NP-hard to decide whether the TV distance between two Bayes nets of in-degree 2 is equal to 0 or not, and also gave an FPTAS for approximating the TV distance between an arbitrary product distribution and any product distribution that has a constant number of distinct marginals (note that this includes the uniform distribution). In a subsequent work, Feng et al. (2023) gave an FPRAS for approximating the TV distance between two arbitrary product distributions. Later, Feng et al. (2024) gave a *deterministic* approximation algorithm (FPTAS) for the same task.

TV distance estimation was also studied previously from a more complexity-theoretic and cryptographic viewpoint. Sahai & Vadhan (2003) established in a seminal work that additively approximating the TV distance between two distributions that are samplable by Boolean circuits is hard for SZK (Statistical Zero Knowledge). Goldreich et al. (1999) showed that the problem of deciding whether a distribution samplable by a Boolean circuit is close or far from the uniform distribution is complete for the complexity class NISZK (Non-Interactive Statistical Zero Knowledge).

Additive approximation of TV distance is much easier. Canonne & Rubinfeld (2014) showed how to additively estimate TV distance between distributions that can be efficiently sampled and whose probability mass functions can be efficiently evaluated. Clearly, Bayes nets satisfy both conditions (where "efficient" means as usual polynomial in the number of parameters). Bhattacharyya et al. (2020) extended this idea to develop polynomial-time algorithms for additively approximating the TV distance between two bounded in-degree Bayes nets using a polynomial number of samples from each.

Probabilistic Inference. There is a significant body of work dedicated to exact probabilistic inference. As we mentioned earlier, some algorithmic paradigms that have been developed for the task of probabilistic inference are message passing (Pearl, 1988), variable elimination (Dechter, 1999), and junction-tree propagation (Lauritzen & Spiegelhalter, 1988). Recently, Klinkenberg et al. (2023) presented an exact Bayesian inference method for inferring posterior distributions encoded by probabilistic programs. Zaiser et al. (2023) present an exact Bayesian inference method for discrete statistical models, by introducing a probabilistic programming language (based on probability generating functions) that supports discrete and continuous sampling, and conditioning on discrete events (among others). Holtzen et al. (2020) develop a domain-specific probabilistic programming language, called Dice, that exploits program structure in order to factorize inference, enabling them to perform exact inference on large probabilistic programs. Saad et al. (2021) present the Sum-Product Probabilistic Language (SPPL), a new probabilistic programming language that automatically delivers exact solutions to a broad range of probabilistic inference queries enabling them to give exact algorithms for conditioning on and computing probabilities of events.

With the advent of big data and the increasing complexity of models, traditional exact inference methods may become computationally infeasible. Approximate inference techniques, such as variational inference and sampling methods like Markov Chain Monte Carlo, provide efficient and scalable alternatives to tackle these challenges. Minka (2001) introduces the expectation propagation algorithm for approximate Bayesian inference. Hoffman et al. (2013) propose a stochastic variational inference algorithm for large-scale Bayesian inference. Murphy et al. (2013) investigate the effectiveness of loopy belief propagation. Ranganath et al. (2014) introduce black box variational inference, a flexible and scalable approach for approximate Bayesian inference. Rezende & Mohamed (2015) propose a variational inference method using normalizing flows, a class of flexible and expressive transformations. Blei et al. (2017) provide a comprehensive review of variational inference, a family of methods for approximate Bayesian inference.

1.3. Organization

The rest of the paper is organized as follows. We provide some background material in Section 2 and a technical overview of our results in Section 3. We prove the main results as follows: We show Theorem 1.1 in Section 4; Theorem 1.2 in Section 4.3; Theorem 1.3 in Appendix B.1; Theorem 1.4 in Appendix B.2. We conclude in Section 5. Appendix A contains all of the proofs that are not presented in the main body.

2. Preliminaries

We use $[n]$ to denote the set $\{1, \ldots, n\}$ and $\log$ to denote $\log_2$. Throughout the paper, we shall assume that all probabilities are represented as rational numbers of the form a/b. We denote the uniform distribution by $\mathbb{U}$.

The following concentration inequality will be useful in our proofs.

Lemma 2.1 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$ for all $1 \leq i \leq n$. Then $\mathbf{Pr}[|\sum_{i=1}^n X_i - \mathbf{E}[\sum_{i=1}^n X_i]| \geq t]$ is at most $2\exp\left(-2t^2/\sum_{i=1}^n (b_i - a_i)^2\right)$.*

We shall use the following notion of an approximation algorithm.

Definition 2.2 (FPRAS)**.** A function $f : \{0,1\}^* \to \mathbb{R}$ admits a *fully polynomial-time randomized approximation scheme (FPRAS)* if there is a *randomized* algorithm $\mathcal{A}$ such that for all n and all inputs $x \in \{0,1\}^n$, $\varepsilon > 0$, and $\delta > 0$, $\mathcal{A}$ outputs a $(1+\varepsilon)$-relative approximation of $f(x)$, i.e., a value v that lies in the interval $[f(x)/(1+\varepsilon), (1+\varepsilon)f(x)]$, with probability $1-\delta$. The running time of $\mathcal{A}$ is polynomial in $n, 1/\varepsilon, 1/\delta$.

2.1. Bayes Nets

For a directed acyclic graph (DAG) G and a node v in G, let $\Pi(v)$ denote the set of parents of v.

Definition 2.3 (Bayes nets)**.** A *Bayes net* is specified by a directed acyclic graph (DAG) over a vertex set $[n]$ and a collection of probability distributions over symbols in $[\ell]$, as follows. Each vertex i is associated with a random variable X_i whose range is $[\ell]$. Each node i of G has a Conditional Probability Table (CPT) that describes the following: For every $x \in [\ell]$ and every $y \in [\ell]^k$, where k is the size of $\Pi(i)$, the CPT has the value of $\mathbf{Pr}[X_i = x | X_{\Pi(i)} = y]$ stored. Given such a Bayes net, its associated probability distribution P is given by the following: For all $x \in [\ell]^n$, $P(x)$ is equal to

$$\mathop{\mathbf{Pr}}_{P}[X = x] = \prod_{i=1}^{n} \mathop{\mathbf{Pr}}_{P}\left[X_i = x_i | X_{\Pi(i)} = x_{\Pi(i)}\right].$$

Here X is the joint distribution $(X_1, \ldots, X_n)$ and $x_{\Pi(i)}$ is the projection of x to the indices in $\Pi(i)$.

Note that $P(x)$ can be computed in linear time by using the CPTs of P to retrieve each $\mathbf{Pr}_P\left[X_i = x_i | X_{\Pi(i)} = x_{\Pi(i)}\right]$. We also use $P_{i|\Pi(i)}\left(x_i | x_{\Pi(i)}\right)$ to denote this probability.

An important notion is that of the moralization of a Bayes net.

Definition 2.4 (Moralization of Bayes nets)**.** Let B be a Bayes net over a DAG G. The *moralization of B* is the undirected graph that is obtained from G as follows. For every node u of G and any pair (v, w) of its parents $\Pi(u)$ if v and w are not connected by some edge in G, then add the edge (v, w). (Note that after this step the parents of every node of G form a clique.) Finally, make all edges of G undirected.

We shall require the following simple observation.

Lemma 2.5. *Given a Bayes net over n nodes, its moralization can be computed in time $O(\mathrm{poly}(n))$.*

2.2. Total Variation Distance

The following notion of distance is central in this work.

Definition 2.6 (Total variation distance)**.** For probability distributions P, Q over a finite sample space $\mathcal{D}$, the *total variation distance* of P and Q is

$$d_{\mathrm{TV}}(P, Q) = \max_{S \subseteq \mathcal{D}} (P(S) - Q(S)).$$

Note that $d_{\mathrm{TV}}(P, Q) = \frac{1}{2}\sum_{w \in \mathcal{D}} |P(w) - Q(w)|$. Equivalently,

$$\begin{aligned} d_{\mathrm{TV}}(P, Q) &= \sum_{w \in \mathcal{D}} \max(0, P(w) - Q(w)) \\ &= \sum_{w \in \mathcal{D}} (P(w) - \min(P(w), Q(w))). \end{aligned}$$

2.3. Probabilistic Inference

In this work, *probabilistic inference* is the following computational task: Given (a representation of) random variables $X_1, \ldots, X_n$ and (a representation of) sets $S_1, \ldots, S_n$ such that for all i the set S_i is a subset of the range of X_i, compute $\mathbf{Pr}[X_1 \in S_1, \ldots, X_n \in S_n]$.[1]

Let us now define probabilistic inference (oracle) queries.

Definition 2.7 (Probabilistic inference query over Bayes nets)**.** A *probabilistic inference query* takes a description of a Bayes net distribution P over n nodes and alphabet size ℓ and descriptions of sets $S_1, \ldots, S_n$, where for all $1 \leq i \leq n$, $S_i \subseteq [\ell]$, and returns the value of $\mathbf{Pr}_P[X_1 \in S_1, \ldots, X_n \in S_n]$ in one time step.

2.4. Treewidth and Tree Decompositions

We require the definition of treewidth.

Definition 2.8. A *tree decomposition of an undirected graph $G = (V, E)$* is a tree T with nodes $X_1, \ldots, X_n$, where each X_i is a subset of V, satisfying the following properties (the term node is used to refer to a vertex of T to avoid confusion with vertices of G): (a) The union of all sets X_i equals V. That is, each graph vertex is contained in at least one tree

[1] Note that a notion of probabilistic inference that has previously been considered (Kwisthout et al., 2010) is the following: Given random variables $X_1, \ldots, X_n$, a set $I = \{i_1, \cdots, i_k\} \subseteq [n]$, values $x_{i_1}, \ldots, x_{i_k}$ that belong to the ranges of $X_{i_1}, \ldots, X_{i_k}$, respectively, and an event E, compute the probability $\mathbf{Pr}[(X_{i_1}, \ldots, X_{i_k}) = (x_{i_1}, \ldots, x_{i_k}) | E]$. However, algorithms such as variable elimination for inference in this sense also works for the notion we consider.

node. (b) If X_i and X_j both contain a vertex v, then all nodes X_k of T in the (unique) path between X_i and X_j contain v as well. Equivalently, the tree nodes containing vertex v form a connected subtree of T. (c) For every edge (v_1, v_2) in the graph, there is a subset X_i that contains both v_1 and v_2. That is, vertices are adjacent in the graph only when the corresponding subtrees have a node in common. The *width of a tree decomposition* is the size of its largest set X_i minus one. The *treewidth* $\mathrm{tw}(G)$ is the minimum width among all possible tree decompositions of G.

We shall also extend the notion of treewidth to Bayes nets, as follows.

Definition 2.9. The *treewidth of a Bayes net* is defined to be equal to the treewidth of its moralization.

We require the following two theorems, Theorem 2.10 and Theorem 2.11, respectively; Theorem 2.10 is about a tree decomposition algorithm and Theorem 2.11 is about the variable elimination algorithm.

Theorem 2.10 (Tree decomposition (Robertson & Seymour, 1984))**.** *There is a $O\big(w3^{3w}n^2\big)$-time algorithm that finds a tree decomposition of width $4w+1$, if the treewidth of the input graph is at most w.*

We will make use of the variable elimination algorithm to efficiently implement probabilistic inference queries for bounded treewidth Bayes nets.

Theorem 2.11 (Variable elimination; following Zhang and Poole (Zhang & Poole, 1994))**.** *There is an algorithm, called the* variable elimination algorithm*, for the following task: Given a Bayes net B over variables $X_1, \ldots, X_n \in [\ell]$, sets $S_1, \ldots, S_n \subseteq [\ell]$, the moralization M_B of B, and a tree decomposition $\mathcal{T}$ of width w of M_B, compute the probability $\mathbf{Pr}_B[X_1 \in S_1, \ldots, X_n \in S_n]$. The running time of this algorithm is $O(n\ell^w)$.*

3. Technical Overview

We present in this section some intuition regarding the technical aspects of our results.

3.1. Proof of Theorem 1.1

For the sake of simplicity of exposition, in this overview we assume that both the Bayes net distributions P and Q are defined over a directed path of length n (over a finite alphabet of size ℓ). However, the ideas can be generalized to arbitrary Bayes nets. Refer to Section 2 for definitions and notation that we use here.

Our approach relies on the well-known importance sampling technique. The high-level approach is to define an estimator function f and a distribution π so that $\mathbf{E}_\pi[f] = d_{\mathrm{TV}}(P,Q)/Z$ where Z is a normalization constant.

We start with the following characterization of d_{TV}: $d_{\mathrm{TV}}(P,Q) = \sum_w g^*(w)$ where $g^*(w)$ is defined to be $P(w) - \min(P(w), Q(w))$. We define an auxiliary function $g(w)$ which is an overestimate of $g^*(w)$. Define $h(w,i)$ as

$$h(w,i) = \min(P_{i|i-1}(w_i|w_{i-1}), Q_{i|i-1}(w_i|w_{i-1}))$$

and let $h(w) = \prod_{i=1}^n h(w,i)$. Finally, set $g(w) = P(w) - h(w)$. Comparing $g(w)$ with $g^*(w)$, note that $\min(P(w), Q(w)$ is the *minimum of the products* where product is of the form $\prod_i P_{i|i-1}(w_i|w_{i-1})$ (similarly for Q). Whereas $h(w)$ is the product of the minimums. Thus, it can be seen that $g(w) \geq g^*(w)$.

Recall that our goal is to estimate $d_{\mathrm{TV}}(P,Q) = \sum_w g^*(w)$. To estimate this we appeal to the classic importance sampling technique (for example, see (Chen & Shao, 1997)). Define function $f(w) = \frac{g^*(w)}{g(w)}$ and distribution $\pi(w) = \frac{g(w)}{Z}$, where $Z = \sum_w g(w)$ is the normalizing constant. Observe that

$$\mathop{\mathbf{E}}_{w\sim\pi}[f(w)] = \frac{\sum_w g^*(w)}{Z} = \frac{d_{\mathrm{TV}}(P,Q)}{Z}.$$

Now the algorithm to estimate d_{TV} works as follows: Empirically estimate $\mathbf{E}_\pi[f(w)]$ by drawing samples from the distribution π and multiply the empirical estimate by Z. We appeal to the standard Chernoff bounds to obtain the guarantee on the quality of the approximation as well as the run time.

This approach will work if the following conditions are satisfied: (i) It is the case that $f(w) = \frac{g^*(w)}{g(w)}$ lies between 0 and 1. This follows from the definition of $g(w)$ and the earlier discussion. (ii) The expectation $\mathbf{E}_{w\sim\pi}[f(w)]$ is large enough (inverse polynomial) so that an additive approximation will lead to a multiplicative approximation. (iii) Sampling from the distribution π and computing $Z = \sum_w g(w)$ can be done efficiently.

The *key insight* that we bring in is that sampling from the distribution π and computing Z reduces to inference queries to a Bayes net distribution over the same directed graph as that of P and Q. Thus (iii) becomes efficient for all Bayes net distributions for which inference queries are feasible. We also show that $\frac{1}{2n} \leq \mathbf{E}_{w\sim\pi}[f(w)]$ thus satisfying (ii). The rest of this subsection is devoted to explaining how inference queries can be used to compute Z. We start with the known connection between d_{TV} and *coupling* between distributions where the quantity $\min(P(w), Q(w))$ naturally arises.

Definition 3.1. Let P and Q are two arbitrary distributions on a common symbol set $[\ell]$ (where $\ell > 0$). A *coupling* of P and Q is a distribution on pairs (X,Y) such that $X \sim P$ and $Y \sim Q$. An *optimal coupling* of P and Q is a distribution on pairs (X,Y) such that (1) $X \sim P$, $Y \sim Q$, and (2) for any $w \in [\ell]$, $\mathbf{Pr}[X = Y = w] = \min(P(w), Q(w))$.

It is well known that for any coupling (X, Y) between P and Q, $d_{\mathrm{TV}}(P, Q) \leq \mathbf{Pr}[X \neq Y]$. Additionally, for an optimal coupling as defined above, $\mathbf{Pr}[X \neq Y]$ exactly equals $d_{\mathrm{TV}}(P, Q)$. Note that,

$$\begin{aligned}\mathbf{Pr}[X \neq Y] &= \sum_w \mathbf{Pr}[X = w, X \neq Y] \\ &= \sum_w \left(\mathbf{Pr}[X = w] - \mathbf{Pr}[X = Y = w]\right).\end{aligned}$$

Thus the term $\mathbf{Pr}[X = w] - \mathbf{Pr}[X = Y = w]$ equals $P(w) - \mathbf{Pr}[X = Y = w]$. By the definition of optimal coupling, $\mathbf{Pr}[X = Y = w]$ is precisely $\min(P(w), Q(w))$. Thus $g^*(w) := P(w) - \min(P(w), Q(w))$ equals $\Pr[X = w \wedge X \neq Y]$. Thus $g^*(w)$ has an interpretation using optimal coupling.

Since $g^*(w)$ can be expressed as the probability of an event over the optimal coupling, it is natural to ask whether there is a coupling $\mathcal{L}$ such that $\mathbf{Pr}_{\mathcal{L}}[X = w \wedge X \neq Y] = g(w)$. In a very recent work, Feng et al. (2023) showed that when P and Q are product distributions, $g(w)$ admits such a characterization using couplings. They define $\mathcal{L}$ as *local coupling* between X and Y: A joint distribution $\mathcal{L}$ on $(X, Y) = (X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ where each (X_i, Y_i) is independently sampled from an optimal coupling of the i-th marginals of P and Q.

Generalizing this approach to Bayes nets poses several obstacles. As in the case of product distributions, suppose we seek a coupling $\mathcal{L}$ of P and Q that also forms a Bayes net over the directed path. In other words, we would like a coupling $\mathcal{L}$ generating the tuple $(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ such that each (X_i, Y_i) is independent of (X_{i-2}, Y_{i-2}) conditioned on (X_{i-1}, Y_{i-1}). However, there is an immediate problem: Namely, X_i and X_{i-2} may be dependent given X_{i-1} through the path $X_{i-2} \to Y_{i-1} \to X_i$, and similarly Y_i and Y_{i-2} may be dependent given Y_{i-1} through the path $Y_{i-2} \to X_{i-1} \to Y_i$. Hence, it may not be possible[2] to ensure that $(X_1, \ldots, X_n)$ form a copy of P and $(Y_1, \ldots, Y_n)$ form a copy of Q, as is required for a coupling.

In light of obstacles faced by the coupling-based approach, we introduce a new notion of *partial couplings*. The introduction of this notion is a primary conceptual contribution of our work.

Definition 3.2. A *partial coupling* of distributions P and Q is a distribution on pairs (X, Y) such that (i) $X \sim P$ and (ii) it is the case that $\mathbf{Pr}[X = Y = w] \leq \min(P(w), Q(w))$.

With the above definition in hand, we will show that it is possible to construct a partial coupling $\mathcal{L}$ of distributions P and Q such that $\mathcal{L}$ can be expressed as a Bayes net distribution over a graph that has the same *structure* as P and Q. We illustrate this for the case when P and Q are Bayes net distributions over a directed path. We define a partial coupling $\mathcal{L}$ that is local. The CPTs are defined as follows: For any $b, c_1, c_2 \in [\ell]$, $\mathbf{Pr}[X_i = Y_i = b | X_{i-1} = c_1, Y_{i-1} = c_2]$ is equal to $\min(P_{i|i-1}(b|c_1), Q_{i|i-1}(b|c_2))$. We will adjust the rest of the entries of the CPT and ensure that for all b, c_1, c_2: $\mathbf{Pr}[X_i = b | X_{i-1} = c_1, Y_{i-1} = c_2] = P_{i|i-1}(b|c_1)$. This ensures that $X \sim P$. It can be shown that $\mathcal{L}$ is a *partial coupling*.

With this, we can indeed connect the function g to the local partial coupling distribution $\mathcal{L}$. We will show that $g(w) = \mathbf{Pr}_{\mathcal{L}}[X = w \wedge X \neq Y]$ and $Z = \mathbf{Pr}_{\mathcal{L}}[X \neq Y]$. Note that $\mathbf{Pr}_{\mathcal{L}}[X \neq Y] = 1 - \mathbf{Pr}_{\mathcal{L}}[X = Y]$. Let E_i denote the event that $(X_i, Y_i) \in \{(1, 1), (2, 2), \ldots (\ell, \ell)\}$. Note that $\mathbf{Pr}_{\mathcal{L}}[X = Y]$ is $\mathbf{Pr}_{\mathcal{L}}[E_1 \cap E_2 \cap \ldots \cap E_n]$ which is an inference query to the Bayes net distribution $\mathcal{L}$.

To summarize: We have shown that the quantity Z can be computed by making an inference query to the distribution $\mathcal{L}$ which is expressible as a Bayes net over a straight line graph. We can build on this idea to show that sampling from the distribution π can also be done by making inference calls to the distribution $\mathcal{L}$. What remains is to show that $\mathbf{E}_\pi[f(w)] = d_{\mathrm{TV}}(P, Q)/Z$ is large enough. We will establish that $Z \leq 2n \cdot d_{\mathrm{TV}}(P, Q)$. The proof of this inequality is somewhat technical and crucially uses properties in the definition of $\mathcal{L}$ mentioned above.

3.2. Proofs of the Rest of the Results

We outline here the main proof ideas of the rest of our results.

Proof of Theorem 1.2. The proof of Theorem 1.2 is an application of Theorem 1.1. To make use of Theorem 1.1, we establish that probabilistic inference (i.e., computing $\mathbf{Pr}[X_1 \in S_1, \ldots, X_n \in S_n]$) can be efficiently implemented for Bayes nets of constant alphabet size and logarithmic treewidth (Lemma 4.8). It is known that a tree decomposition of graphs that have logarithmic treewidth can be computed in polynomial time (Robertson & Seymour, 1984). The variable elimination algorithm of Zhang & Poole (1994) shows that inference can be done in polynomial time given a tree decomposition, provided that the treewidth of the Bayes net is logarithmic in the dimension of the distribution.

Proof of Theorem 1.3. Theorem 1.3 is proved by showing a reduction from #SAT to computing the TV distance between an appropriately defined Bayes net and the uniform distribution. This is achieved by creating a Bayes net that captures the circuit structure of a Boolean formula F of which we want to compute its number of satisfying assignments. The CPTs of this Bayes net mimic the function of

[2] Note that this issue does not arise for product distributions as there are no paths to speak of.

the logical gates (AND, OR, NOT) of F.

Proof of Theorem 1.4. Theorem 1.4 is proved by giving an algorithm that exploits the following property of TV distance. Let P be a Bayes net over n variables that has maximum in-degree d and alphabet size ℓ. In this case $d_{\mathrm{TV}}(P, \mathbb{U})$ is equal to

$$\begin{aligned}\frac{1}{2}\sum_x |P(x) - \mathbb{U}(x)| &= \sum_x \max(0, P(x) - \mathbb{U}(x)) \\ &= \sum_x \mathbb{U}(x) \max\left(0, \frac{P(x)}{\mathbb{U}(x)} - 1\right) \\ &= \mathop{\mathbf{E}}_{x\sim\mathbb{U}}\left[\max\left(0, \frac{P(x)}{\mathbb{U}(x)} - 1\right)\right] \\ &= \mathop{\mathbf{E}}_{x\sim\mathbb{U}}[\max(0, P(x)\,\ell^n - 1)]\,.\end{aligned}$$

This yields a natural estimator for $d_{\mathrm{TV}}(P, \mathbb{U})$, whereby we draw samples $x_1, \ldots, x_m \sim \mathbb{U}$ and then compute and output

$$\frac{1}{m}\sum_{i=1}^{m} \max(0, P(x_i)\,\ell^n - 1)\,.$$

The crux of our analysis is to show that $\max(0, P(x)\,\ell^n - 1)$ is between 0 and $1 + O\big(d_{\mathrm{TV}}(P, \mathbb{U})\,\ell^{d+1} n\big)$. This enables us to use a value of m that is in $O\big(\mathrm{poly}\big(n\ell^d, 1/\varepsilon, \log(1/\delta)\big)\big)$, whereby ε is the accuracy error and δ is the confidence error of the FPRAS. Note that the running time is polynomial in the input length, as any description of the Bayes net P has size at least $n + \ell^{d+1}$.

4. From TV Distance to Probabilistic Inference

In this section, we prove Theorem 1.1 and Theorem 1.2. In the following, let $T(G, \ell)$ be the running time of some implementation of a probabilistic inference oracle for a Bayes net over a DAG G that has alphabet size ℓ. We will first state the formal version of Theorem 1.1.

Theorem 1.1 (Formal)**.** *There is a polynomial-time randomized algorithm that takes a DAG G, two Bayes nets P and Q over G (as CPTs) with alphabet size ℓ, and parameters ε, δ as inputs and behaves as follows. The algorithm constructs a Bayes net distribution $\mathcal{L}$ over the same DAG G with alphabet size ℓ^2, makes probabilistic inference queries to $\mathcal{L}$, and outputs an $(1+\varepsilon)$-relative approximation of $d_{\mathrm{TV}}(P, Q)$ with probability at least $1 - \delta$. The running time of this algorithm is $T\big(G, \ell^2\big) \cdot O\big(n^3\varepsilon^{-2}\ell \log \delta^{-1}\big)$ and the number of its probabilistic inference queries is $O\big(n^3\varepsilon^{-2}\ell \log \delta^{-1}\big)$.*

The rest of the section is devoted to proving Theorem 1.1 and is organized as follows. We first introduce the ingredients that are necessary for describing the algorithm (many of these are defined in Section 3 for path Bayes nets). In Section 4.1, we show how the algorithm can be implemented using probabilistic inference queries. Finally, in Section 4.2 we establish its correctness. Due to space limitation, many of the technical proofs are given in the Appendix.

Let P and Q be two Bayes net distributions defined over a DAG G with n nodes and alphabet $[\ell]$. Without loss of generality, assume that the nodes are topologically ordered as in the sequence $1, 2, \ldots, n$.

Let w be an element of the sample space, i.e., a n-symbol string over $[\ell]$. Given $1 \le i \le n$, $\Pi(i)$ denotes the set of parents of i in G and let $w_{\Pi(i)}$ denote the projection w at the parents of node i in G. We first define a function h over $[\ell]^n \times [n]$ as follows:

$$h(w, i) := \min\big(P_{i|\Pi(i)}\big(w_i|w_{\Pi(i)}\big)\,, Q_{i|\Pi(i)}\big(w_i|w_{\Pi(i)}\big)\big)\,.$$

Descriptions of f, Z, and π. The *estimator function* f is defined as follows: $f(w) := g^*(w)/g(w)$ where

$$g^*(w) = P(w) - \min(P(w), Q(w))$$

and

$$g(w) := P(w) - \prod_{i=1}^{n} h(w, i)$$

for all w. It is straightforward to show that f is computable in time $O(n)$. We define $Z := \sum_{w\in[\ell]^n} g(w)$ to be a normalization constant. Finally, the distribution π is specified by the probability function $\pi(w) := g(w)/Z$ for all w.

Description of $\mathcal{L}$. We now define a Bayes net distribution $\mathcal{L}$ over the graph G which is used to make inference queries by the algorithm. The distribution $\mathcal{L}$ is over the alphabet $[\ell]^2$ and is a joint distribution (X, Y) where X and Y take value over $[\ell]^n$. We specify a CPT for (X, Y). For this, we need to specify for every i and $b, z \in [\ell]$ the probability $\Pr[(X_i, Y_i) = (b, z)]$ conditioned on the values $\Pi(i)$ take. We will first describe the probability where both X_i and Y_i take the same value b. For every $c_1, c_2 \in [\ell]^{|\Pi(i)|}$,

$$\begin{aligned}&\mathbf{Pr}\big[(X_i, Y_i) = (b, b) \mid \big(X_{\Pi(i)}, Y_{\Pi(i)}\big) = (c_1, c_2)\big] \\ &\quad = \min\big(P_{i|\Pi(i)}(b|c_1)\,, Q_{i|\Pi(i)}(b|c_2)\big)\,.\end{aligned}$$

Define the remaining probabilities to ensure that the marginal X is distributed according to P. That is, for every $z \neq b$ assign $\mathbf{Pr}\big[(X_i, Y_i) = (b, z) \mid \big(X_{\Pi(i)}, Y_{\Pi(i)}\big) = (c_1, c_2)\big]$ so that the following holds:

$$\begin{aligned}&\sum_{z:z\neq b} \mathbf{Pr}\big[(X_i, Y_i) = (b, z) \mid \big(X_{\Pi(i)}, Y_{\Pi(i)}\big) = (c_1, c_2)\big] \\ &\quad = P_{i|\Pi(i)}(b|c_1) - \min\big(P_{i|\Pi(i)}(b|c_1)\,, Q_{i|\Pi(i)}(b|c_2)\big)\end{aligned}$$

Description of the Algorithm. Now we are ready to describe the algorithm (see Algorithm 1).

Algorithm 1 FPRAS for d_{TV} estimation using a probabilistic inference oracle.

Require: Bayes nets P, Q over DAG G with n nodes, parameters ε, δ.
Ensure: The output Est is an $(1+\varepsilon)$-approximation of $d_{\mathrm{TV}}(P, Q)$, with probability at least $1-\delta$.

```
Construct the Bayes net distribution 𝓛 over G
Compute Z by making one inference query to 𝓛
if Z = 0 then
    return 0
end if
m ← C n^2 ε^{-2} log δ^{-1}        (for some large C > 0)
F ← 0
for i ← 1 to m do
    Sample w^i ~ π by making inference queries to 𝓛
    F ← F + f(w^i)
end for
Est ← Z · F/m
return Est
```

4.1. The Power of Probabilistic Inference

This subsection is devoted to showing that the sampling from the distribution π and the computation of the normalization constant Z can be done by making probabilistic inference queries. Recall that $\mathcal{L}$ is the joint distribution (X, Y). We start with the following crucial observation which states that the marginal X (in $\mathcal{L}$) is distributed according to the distribution P.

Observation 4.1. *For every $b \in [\ell]$ and $c_1, c_2 \in [\ell]^{|\Pi(i)|}$,*

$$\mathbf{Pr}\left[X_i = b \middle| \left(X_{\Pi(i)}, Y_{\Pi(i)}\right) = (c_1, c_2)\right] = P_{i|\Pi(i)}(b|c_1)\,.$$

Therefore, X factorizes like P with its conditional probabilities matching that of P and hence $X \sim P$. This realizes the notion of a local partial coupling as was earlier discussed in Section 3.1 and satisfies all three properties: (i) $\mathcal{L}$ is a Bayes net distribution over the same DAG G (that is used to describe distributions P and Q), (ii) $X \sim P$, and (iii) in the joint distribution (X, Y), the conditional probabilities are equal to the minimum of the two conditional probabilities associated to P and Q as it is the case in standard couplings.

In Claim 4.2 we relate the normalization constant Z of the distribution π to the marginals X and Y of the distribution $\mathcal{L}$. Moreover, we also relate the generalized normalization constant

$$Z_{b_1,\ldots,b_k} := \sum_{w:(w_1,\ldots,w_k)=(b_1,\ldots,b_k)} g(w)\,,$$

for $b_1, \ldots, b_k \in [\ell]$, to the marginals X and Y of the distribution $\mathcal{L}$. We need this generalized normalization constant to show that sampling from the distribution π (Claim 4.4) can be efficiently done via probabilistic inference queries.

Claim 4.2. *It is the case that $Z = \mathbf{Pr}[X \neq Y]$ and*

$$Z_{b_1,\ldots,b_k} = \mathbf{Pr}[X \neq Y, X_1 = b_1, \ldots, X_k = b_k]$$

for any $b_1, \ldots, b_k \in [\ell]$.

The following claim says that Z and $Z_{b_1,\ldots,b_k}$ can be easily computed given access to a probabilistic inference oracle for $\mathcal{L}$.

Claim 4.3. *It is the case that Z and $Z_{b_1,\ldots,b_k}$ (for any $b_1, \ldots, b_k \in [\ell]$) can be computed efficiently by making $O(1)$ probabilistic inference queries to the Bayes net distribution $\mathcal{L}$.*

Proof. We will use Claim 4.2. Note that $Z = \mathbf{Pr}[X \neq Y]$ is equal to $1 - \mathbf{Pr}[X = Y]$. Therefore it suffices to compute $\mathbf{Pr}[X = Y]$ by using a probabilistic inference oracle. This can done by observing that $\mathbf{Pr}[X = Y]$ is equal to $\mathbf{Pr}[(X_1, Y_1) \in S_1, \ldots, (X_n, Y_n) \in S_n]$ for $S_1 = \cdots = S_n = \{(1,1), \ldots, (\ell, \ell)\}$.

Now note that the quantity $Z_{b_1,\ldots,b_k}$ is equal to $\mathbf{Pr}[X \neq Y, X_1 = b_1, \ldots, X_k = b_k]$ which, in turn, is equal to $\mathbf{Pr}[X_1 = b_1, \ldots, X_k = b_k] - \mathbf{Pr}[X = Y, X_1 = b_1, \ldots, X_k = b_k]$. What is left now, is to show how to compute these two probabilities by using a probabilistic inference oracle. We have that the value of $\mathbf{Pr}[X_1 = b_1, \ldots, X_k = b_k]$ is $\mathbf{Pr}[(X_1, Y_1) \in S_1, \ldots, (X_n, Y_n) \in S_n]$ for $S_i = \{(b_i, 1), \ldots, (b_i, \ell)\}$ for all $1 \leq i \leq k$ and $S_{k+1} = \cdots = S_n = [\ell]^2$.

Similarly, we have that $\mathbf{Pr}[X = Y, X_1 = b_1, \ldots, X_k = b_k]$ equals $\mathbf{Pr}[(X_1, Y_1) \in S_1, \ldots, (X_n, Y_n) \in S_n]$ for $S_i = \{(b_i, b_i)\}$ for all $1 \leq i \leq k$, and $S_{k+1} = \cdots = S_n = \{(1,1), \ldots, (\ell, \ell)\}$. □

We can now show that probabilistic inference queries allow for efficient sampling from π.

Claim 4.4. *Sampling from the distribution π can be implemented in time $O(n\ell)$ by making $O(n\ell)$ probabilistic inference queries.*

4.2. Analysis of the Algorithm

Next, we establish some useful properties of the function f and the distribution π.

Claim 4.5. *For any w, it is the case that $0 \leq f(w) \leq 1$.*

We will also relate the expected value of the function f with respect to the distribution π to $d_{\mathrm{TV}}(P, Q)$.

Claim 4.6. *It is the case that $\mathbf{E}_\pi[f(w)] = d_{\mathrm{TV}}(P, Q)/Z$.*

We need the following lemma that ensures the estimand is large enough to facilitate Monte Carlo sampling.

Lemma 4.7. *It is the case that* $Z \leq 2n \cdot d_{\mathrm{TV}}(P, Q)$.

We are now ready to prove the correctness of and provide a running time bound for Algorithm 1. We have, from Hoeffding's inequality (Lemma 2.1), that

$$\begin{aligned}
&\mathbf{Pr}[|\mathsf{Est} - d_{\mathrm{TV}}(P,Q)| > \varepsilon d_{\mathrm{TV}}(P,Q)] \\
&= \mathbf{Pr}\left[\left|\frac{Z}{m}\sum_{i=1}^{m} f\left(w^i\right) - Z \mathop{\mathbf{E}}_{\pi}[f(w)]\right| > \varepsilon d_{\mathrm{TV}}(P,Q)\right] \\
&= \mathbf{Pr}\left[\left|\sum_{i=1}^{m} f\left(w^i\right) - m \mathop{\mathbf{E}}_{\pi}[f(w)]\right| > \frac{m\varepsilon}{Z} d_{\mathrm{TV}}(P,Q)\right] \\
&\leq 2\exp\left(-\frac{2m^2\varepsilon^2 d_{\mathrm{TV}}^2(P,Q)}{Z^2\sum_{i=1}^{m}(0-1)^2}\right) \\
&\leq 2\exp\left(-\frac{2m^2\varepsilon^2 d_{\mathrm{TV}}^2(P,Q)}{4n^2 d_{\mathrm{TV}}^2(P,Q)\, m}\right) \\
&= 2\exp\left(-\frac{m\varepsilon^2}{2n^2}\right),
\end{aligned}$$

which is at most δ whenever $m = \Omega\left(n^2\varepsilon^{-2}\log\delta^{-1}\right)$. The second inequality follows from Lemma 4.7.

Thus the running time of Algorithm 1 is $O(mn\ell)$, which equals $O\left(n^3\varepsilon^{-2}\ell\log\delta^{-1}\right)$, since we draw m samples from π, we can sample from π in time $O(n\ell)$, and evaluate f in time $O(n)$. Finally, the number of probabilistic inference queries is at most $O\left(n^3\varepsilon^{-2}\ell\log\delta^{-1}\right)$.

4.3. Application: Bayes Nets of Small Treewidth

We are now ready to prove Theorem 1.2.

Theorem 1.2 (Formal)**.** *There is an FPRAS for estimating the TV distance between two Bayes nets of treewidth* $w = O(\log n)$ *and alphabet size* $\ell = O(1)$*, which are defined over the same DAG of* n *nodes. In particular, if* ε *and* δ *are the accuracy and confidence errors of the FPRAS, respectively, the FPRAS runs in time* $\mathrm{poly}(n) \cdot O\left(\varepsilon^{-2}\log\delta^{-1}\right)$.

The proof of Theorem 1.2 will follow from the lemma below, Lemma 4.8, and Theorem 1.1 for $\ell = O(1)$ and $T\left(G, \ell^2\right) = O(\mathrm{poly}(n))$.

Lemma 4.8. *Probabilistic inference is efficient for all Bayes nets over* n *variables which have alphabet size* $\ell = O(1)$ *and treewidth* $O(\log n)$.

5. Conclusion

We have established a general connection between probabilistic inference and TV distance computation. In particular, we proved that TV distance estimation can be reduced to probabilistic inference in a structure preserving manner. This enables us to prove the existence of a novel FPRAS for estimating the TV distance between Bayes nets of small treewidth.

The notion of *partial couplings* introduced in this work is of independent interest. It would be fruitful to explore applications of this notion in other contexts.

We outline the following open problems: Can we prove similar results for TV distance estimation between undirected graphical models? Another problem of interest is to study other notions of distance, such as Wasserstein metrics.

Acknowledgements

The work of AB was supported in part by National Research Foundation Singapore under its NRF Fellowship Programme (NRF-NRFFAI-2019-0002) and an Amazon Faculty Research Award. The work of SG was supported by an initiation grant from IIT Kanpur and a SERB award CRG/2022/007985. Pavan's work is partly supported by NSF award 2130536. Vinodchandran's work is partly supported by NSF award 2130608. This work was supported in part by National Research Foundation Singapore under its NRF Fellowship Programme [NRF-NRFFAI1-2019-0004] and an Amazon Research Award. Part of the work was done during Meel, Pavan, and Vinodchandran's visit to the Simons Institute for the Theory of Computing.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Bhattacharyya, A., Gayen, S., Meel, K. S., and Vinodchandran, N. V. Efficient distance approximation for structured high-dimensional distributions via learning. In *Proc. of NeurIPS*, 2020.

Bhattacharyya, A., Gayen, S., Meel, K. S., Myrisiotis, D., Pavan, A., and Vinodchandran, N. V. On approximating total variation distance. In *Proc. of IJCAI*, pp. 3479–3487. ijcai.org, 2023.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Canonne, C. L. and Rubinfeld, R. Testing probability distributions underlying aggregated data. In Esparza, J., Fraigniaud, P., Husfeldt, T., and Koutsoupias, E. (eds.), *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I*, volume 8572 of *Lec-*

ture Notes in Computer Science, pp. 283–295. Springer, 2014.

Chen, M.-H. and Shao, Q.-M. On Monte Carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics*, 25(4):1563–1594, 1997.

Cooper, G. F. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial intelligence*, 42(2-3):393–405, 1990.

Cover, T. M. and Thomas, J. A. *Elements of information theory (2. ed.)*. Wiley, 2006.

Dechter, R. Bucket elimination: A unifying framework for reasoning. *Artificial Intelligence*, 113(1-2):41–85, 1999.

Doeblin, W. Exposé de la théorie des chaınes simples constantes de markova un nombre fini d'états. *Mathématique de l'Union Interbalkanique*, 2(77-105):78–80, 1938.

Dwork, C. Differential privacy. In Bugliesi, M., Preneel, B., Sassone, V., and Wegener, I. (eds.), *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, volume 4052 of *Lecture Notes in Computer Science*, pp. 1–12. Springer, 2006.

Feng, W., Guo, H., Jerrum, M., and Wang, J. A simple polynomial-time approximation algorithm for the total variation distance between two product distributions. *TheoretiCS*, 2, 2023.

Feng, W., Liu, L., and Liu, T. On deterministically approximating total variation distance. In Woodruff, D. P. (ed.), *Proceedings of the 2024 ACM-SIAM Symposium on Discrete Algorithms, SODA 2024, Alexandria, VA, USA, January 7-10, 2024*, pp. 1766–1791. SIAM, 2024.

Goldreich, O., Sahai, A., and Vadhan, S. P. Can statistical zero knowledge be made non-interactive? or On the relationship of SZK and NISZK. In *Proc. of CRYPTO*, pp. 467–484, 1999.

Golia, P., Soos, M., Chakraborty, S., and Meel, K. S. Designing samplers is easy: The boon of testers. In *Formal Methods in Computer Aided Design, FMCAD 2021, New Haven, CT, USA, October 19-22, 2021*, pp. 222–230. IEEE, 2021.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. In *JMLR*, volume 14, pp. 1303–1347, 2013.

Holtzen, S., den Broeck, G. V., and Millstein, T. D. Dice: Compiling discrete probabilistic programs for scalable inference. *CoRR*, abs/2005.09089, 2020.

Klinkenberg, L., Blumenthal, C., Chen, M., and Katoen, J. Exact Bayesian inference for loopy probabilistic programs. *CoRR*, abs/2307.07314, 2023.

Koller, D. and Friedman, N. *Probabilistic graphical models: Principles and techniques*. MIT press, 2009.

Kwisthout, J., Bodlaender, H. L., and van der Gaag, L. C. The necessity of bounded treewidth for efficient inference in Bayesian networks. In *ECAI*, volume 215, pp. 237–242, 2010.

Lauritzen, S. L. and Spiegelhalter, D. J. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988.

Levin, D. A., Peres, Y., and Wilmer, E. L. *Markov chains and mixing times*. American Mathematical Society, 2006.

Lindvall, T. *Lectures on the coupling method*. Courier Corporation, 2002.

Meyn, S. P. and Tweedie, R. L. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

Minka, T. Expectation propagation for approximate Bayesian inference. In *UAI*, 2001.

Mitzenmacher, M. and Upfal, E. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

Murphy, K. P., Weiss, Y., and Jordan, M. I. Loopy belief propagation for approximate inference: An empirical study. *CoRR*, abs/1301.6725, 2013.

Pearl, J. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 1988.

Ranganath, R., Gerrish, S., and Blei, D. M. Black box variational inference. *Arxiv preprint arXiv:1401.0118*, 2014. URL https://arxiv.org/abs/1401.0118.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. *Arxiv preprint arXiv:1505.05770*, 2015. URL https://arxiv.org/abs/1505.05770.

Robertson, N. and Seymour, P. D. Graph minors III. Planar tree-width. *J. Comb. Theory, Ser. B*, 36(1):49–64, 1984.

Saad, F. A., Rinard, M. C., and Mansinghka, V. K. SPPL: probabilistic programming with fast exact symbolic inference. In Freund, S. N. and Yahav, E. (eds.), *PLDI '21: 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation, Virtual Event, Canada, June 20-25, 2021*, pp. 804–819. ACM, 2021.

Sahai, A. and Vadhan, S. P. A complete problem for statistical zero knowledge. *J. ACM*, 50(2):196–249, 2003.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.

Stinson, D. R. *Cryptography - theory and practice*. Discrete mathematics and its applications series. CRC Press, 1995.

Vadhan, S. P. Pseudorandomness. *Found. Trends Theor. Comput. Sci.*, 7(1-3):1–336, 2012.

Wainwright, M. J., Jordan, M. I., et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Zaiser, F., Murawski, A. S., and Ong, C. L. Exact bayesian inference on discrete models via probability generating functions: A probabilistic programming approach. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Zhang, N. L. and Poole, D. L. A simple approach to Bayesian network computations, 1994.

A. Deferred Proofs

We present here all of the deferred proofs.

A.1. Proof of Lemma 2.5

Let B be a Bayes net over a DAG G that has n nodes. Let v be a node of G and let $\Pi(v)$ be the set of the parents of v. We can construct a clique among the nodes of $\Pi(v)$ in time $O\big(n^2\big)$, since $|\Pi(v)| \leq n$. Therefore we can construct all of the required cliques in time $n \cdot O\big(n^2\big) = O\big(n^3\big)$. Finally, we can make all directed edges of G undirected in time $O\big(n^2\big)$. This yields a total running time of $O\big(n^3\big)$.

A.2. Proof of Observation 4.1

We have

$$\begin{aligned}
\mathbf{Pr}\big[X_i = b \big| \left(X_{\Pi(i)}, Y_{\Pi(i)}\right) = (c_1, c_2)\big] &= \sum_{z \in [\ell]} \mathbf{Pr}\big[(X_i, Y_i) = (b, z) \,\big|\, \left(X_{\Pi(i)}, Y_{\Pi(i)}\right) = (c_1, c_2)\big] \\
&= \mathbf{Pr}\big[(X_i, Y_i) = (b, b) \,\big|\, \left(X_{\Pi(i)}, Y_{\Pi(i)}\right) = (c_1, c_2)\big] \\
&\quad + \sum_{z : z \neq b} \mathbf{Pr}\big[(X_i, Y_i) = (b, z) \,\big|\, \left(X_{\Pi(i)}, Y_{\Pi(i)}\right) = (c_1, c_2)\big] \\
&= \min\big(P_{i|\Pi(i)}(b|c_1)\,, Q_{i|\Pi(i)}(b|c_2)\big) + P_{i|\Pi(i)}(b|c_1) \\
&\quad - \min\big(P_{i|\Pi(i)}(b|c_1)\,, Q_{i|\Pi(i)}(b|c_2)\big) \\
&= P_{i|\Pi(i)}(b|c_1)
\end{aligned}$$

as desired.

A.3. Proof of Claim 4.2

Since $X \sim P$ and for that matter $P(w) = \mathbf{Pr}[X = w]$, we have

$$\begin{aligned}
g(w) &= P(w) - \prod_{i=1}^{n} \min\big(P_{i|\Pi(i)}\big(w_i|w_{\Pi(i)}\big)\,, Q_{i|\Pi(i)}\big(w_i|w_{\Pi(i)}\big)\big) \\
&= P(w) - \mathbf{Pr}[X = Y = w] \\
&= \mathbf{Pr}[X = w] - \mathbf{Pr}[X = Y = w] \\
&= \mathbf{Pr}[X = w] - \mathbf{Pr}[Y = w | X = w] \cdot \mathbf{Pr}[X = w] \\
&= \mathbf{Pr}[X = w] \Big(1 - \mathbf{Pr}[Y = w | X = w]\Big) \\
&= \mathbf{Pr}[X = w]\, \mathbf{Pr}[Y \neq w | X = w] \\
&= \mathbf{Pr}[X = w, Y \neq w]\,.
\end{aligned}$$

Therefore, we have that $Z = \sum_w g(w) = \mathbf{Pr}[X \neq Y]$ and

$$\begin{aligned}
Z_{b_1,\ldots,b_k} &= \sum_{w : (w_1,\ldots,w_k) = (b_1,\ldots,b_k)} g(w) \\
&= \sum_{w : (w_1,\ldots,w_k) = (b_1,\ldots,b_k)} \mathbf{Pr}[X = w, Y \neq w] \\
&= \mathbf{Pr}[X \neq Y, X_1 = b_1, \ldots, X_k = b_k]\,.
\end{aligned}$$

A.4. Proof of Claim 4.4

We describe how to sample from π iteratively, symbol by symbol. Assume that we have sampled the first $k-1$ symbols, that is, assume that we have already sampled $w_1, \ldots, w_{k-1}$ to be equal to $b_1, \ldots, b_{k-1} \in [\ell]$. We describe now how to

sample w_k. For every possible value $b \in [\ell]$ of w_k, we compute the marginal

$$\mu_b := \pi(b_1, \ldots, b_{k-1}, b) = \frac{\sum_{w:(w_1 \cdots w_k)=(b_1 \cdots b_{k-1} b)} g(w)}{Z} = \frac{Z_{b_1 \cdots b_{k-1} b}}{Z}$$

by two invocations of Claim 4.3. Then, we sample w_k based on the values $\{\mu_b\}_{b=1}^{\ell}$.

Let $S(n)$ be the number of steps to sample n symbols from π. The above procedure gives the recurrence relation $S(n) = O(\ell) + S(n-1)$ which yields $S(n) = O(n\ell)$. Since we perform at most two probabilistic inference queries for every coordinate and every symbol, the total number of probabilistic inference queries is at most $S(n) = O(n\ell)$.

A.5. Proof of Claim 4.5

We separately show $0 \leq f(w)$ and $f(w) \leq 1$. To establish $0 \leq f(w)$, since the numerator is non-negative, it suffices to show that $g(w) = P(w) - \prod_{i=1}^{n} h(w,i) \geq 0$ or equivalently $P(w) \geq \prod_{i=1}^{n} h(w,i)$.

We have

$$P(w) = \prod_{i=1}^{n} P_{i|\Pi(i)}\left(w_i | w_{\Pi(i)}\right) \geq \prod_{i=1}^{n} \min\left(P_{i|\Pi(i)}\left(w_i | w_{\Pi(i)}\right), Q_{i|\Pi(i)}\left(w_i | w_{\Pi(i)}\right)\right) = \prod_{i=1}^{n} h(w,i).$$

For showing $f(w) \leq 1$, it suffices to show that $P(w) - Q(w) \leq g(w)$ (since $0/g(w) = 0 \leq 1$). Since $g(w) = P(w) - \prod_{i-1}^{n} h(w,i)$, it suffices to show that $Q(w) \geq \prod_{i=1}^{n} h(w,i)$. An argument identical to the above, where we showed that $P(w) \geq \prod_{i=1}^{n} h(w,i)$, will show this.

A.6. Proof of Claim 4.6

We have that

$$\begin{aligned}
\mathop{\mathbf{E}}_{\pi}[f(w)] &= \mathop{\mathbf{E}}_{\pi}\left[\frac{\max(0, P(w) - Q(w))}{g(w)}\right] \\
&= \sum_{w} \pi(w) \frac{\max(0, P(w) - Q(w))}{g(w)} \\
&= \sum_{w} \frac{g(w)}{Z} \frac{\max(0, P(w) - Q(w))}{g(w)} \\
&= \frac{1}{Z} \sum_{w} \max(0, P(w) - Q(w)) \\
&= \frac{d_{\mathrm{TV}}(P, Q)}{Z}.
\end{aligned}$$

A.7. Proof of Lemma 4.7

By Claim 4.2, it suffices to show that $\mathbf{Pr}[X \neq Y] \leq 2n \cdot d_{\mathrm{TV}}(P, Q)$. We split the event $(X \neq Y)$ into n disjoint events $\{E_i\}_{i=1}^{n}$. Without loss of generality, assume that $1, 2, \ldots, n$ is the topological ordering of the vertices of G. Event E_i is defined as $(\bigwedge_{1 \leq j \leq i-1} X_j = Y_j) \wedge (X_i \neq Y_i)$. Note that the E_i's are disjoint. Thus $\mathbf{Pr}[X \neq Y] = \sum_i \mathbf{Pr}[E_i]$. We have that

$$\mathbf{Pr}[E_i] \leq \mathbf{Pr}\left[(X_i \neq Y_i) \wedge (X_{\Pi(i)} = Y_{\Pi(i)})\right] = \sum_{\sigma} \mathbf{Pr}\left[(X_i \neq Y_i) \wedge (X_{\Pi(i)}, Y_{\Pi(i)}) = (\sigma, \sigma)\right]$$

where σ is an assignment for $\Pi(i)$ (note that the length of σ is equal to the in-degree of i). Henceforth, for notational brevity, we shall omit the dependence on i. Thus,

$$\begin{aligned}
\mathbf{Pr}[X \neq Y] &= \sum_{i} \mathbf{Pr}[E_i] \\
&\leq \sum_{i} \sum_{\sigma} \mathbf{Pr}\left[X_i \neq Y_i \wedge \left(X_{\Pi(i)}, Y_{\Pi(i)}\right) = (\sigma, \sigma)\right]
\end{aligned}$$

$$= \sum_i \sum_\sigma \mathbf{Pr}\left[X_i \neq Y_i \middle| \left(X_{\Pi(i)}, Y_{\Pi(i)}\right) = (\sigma,\sigma)\right] \cdot \mathbf{Pr}\left[\left(X_{\Pi(i)}, Y_{\Pi(i)}\right) = (\sigma,\sigma)\right].$$

We require the following claim.

Claim A.1. *For any σ, it is the case that*

$$\mathbf{Pr}\left[X_i \neq Y_i \middle| \left(X_{\Pi(i)}, Y_{\Pi(i)}\right) = (\sigma,\sigma)\right] = d_{\mathrm{TV}}\left(P_{i|\Pi(i)}(\cdot|\sigma), Q_{i|\Pi(i)}(\cdot|\sigma)\right).$$

Proof. We have that

$$\begin{aligned}
\mathbf{Pr}\left[X_i \neq Y_i \middle| \left(X_{\Pi(i)}, Y_{\Pi(i)}\right) = (\sigma,\sigma)\right] &= 1 - \mathbf{Pr}\left[X_i = Y_i \middle| \left(X_{\Pi(i)}, Y_{\Pi(i)}\right) = (\sigma,\sigma)\right] \\
&= 1 - \sum_{c\in[\ell]} \mathbf{Pr}\left[(X_i, Y_i) = (c,c) \,\middle|\, \left(X_{\Pi(i)}, Y_{\Pi(i)}\right) = (\sigma,\sigma)\right] \\
&= 1 - \sum_{c\in[\ell]} \min\left(P_{i|\Pi(i)}(c|\sigma), Q_{i|\Pi(i)}(c|\sigma)\right) \\
&= \sum_{c\in[\ell]} P_{i|\Pi(i)}(c|\sigma) - \sum_{c\in[\ell]} \min\left(P_{i|\Pi(i)}(c|\sigma), Q_{i|\Pi(i)}(c|\sigma)\right) \\
&= \sum_{c\in[\ell]} \left(P_{i|\Pi(i)}(c|\sigma) - \min\left(P_{i|\Pi(i)}(c|\sigma), Q_{i|\Pi(i)}(c|\sigma)\right)\right) \\
&= \sum_{c\in[\ell]} \max\left(0, P_{i|\Pi(i)}(c|\sigma) - Q_{i|\Pi(i)}(c|\sigma)\right) \\
&= d_{\mathrm{TV}}\left(P_{i|\Pi(i)}(\cdot|\sigma), Q_{i|\Pi(i)}(\cdot|\sigma)\right).
\end{aligned}$$

□

By Claim A.1 we have that $\mathbf{Pr}[X \neq Y]$ is at most

$$\begin{aligned}
&\sum_i \sum_\sigma \mathbf{Pr}\left[\left(X_{\Pi(i)}, Y_{\Pi(i)}\right) = (\sigma,\sigma)\right] d_{\mathrm{TV}}\left(P_{i|\Pi(i)}(\cdot|\sigma), Q_{i|\Pi(i)}(\cdot|\sigma)\right) \\
&\quad \leq \sum_i \sum_\sigma \mathbf{Pr}\left[X_{\Pi(i)} = \sigma\right] \frac{1}{2} \sum_c \left|P_{i|\Pi(i)}(c|b) - Q_{i|\Pi(i)}(c|\sigma)\right| \\
&\quad \leq \sum_i \sum_\sigma P_{\Pi(i)}(\sigma) \frac{1}{2} \sum_c \left|P_{i|\Pi(i)}(c|\sigma) - Q_{i|\Pi(i)}(c|\sigma)\right| \quad \text{(since } X \sim P \text{ by Observation 4.1)} \\
&\quad = \sum_i \sum_\sigma \frac{1}{2} \sum_c \left|P_{\Pi(i)}(\sigma)\, P_{i|\Pi(i)}(c|\sigma) - P_{\Pi(i)}(\sigma)\, Q_{i|\Pi(i)}(c|\sigma)\right| \\
&\quad = \sum_i \sum_\sigma \frac{1}{2} \sum_c \left|P_{\Pi(i)}(\sigma)\, P_{i|\Pi(i)}(c|\sigma) - Q_{\Pi(i)}(\sigma)\, Q_{i|\Pi(i)}(c|\sigma)\right. \\
&\qquad \left. + Q_{\Pi(i)}(\sigma)\, Q_{i|\Pi(i)}(c|\sigma) - P_{\Pi(i)}(\sigma)\, Q_{i|\Pi(i)}(c|\sigma)\right| \\
&\quad \leq \sum_i \sum_\sigma \frac{1}{2} \sum_c \left|P_{\Pi(i)}(\sigma)\, P_{i|\Pi(i)}(c|\sigma) - Q_{\Pi(i)}(\sigma)\, Q_{i|\Pi(i)}(c|\sigma)\right| \\
&\qquad + \sum_i \sum_\sigma \frac{1}{2} \sum_c \left|Q_{\Pi(i)}(\sigma)\, Q_{i|\Pi(i)}(c|\sigma) - P_{\Pi(i)}(\sigma)\, Q_{i|\Pi(i)}(c|\sigma)\right| \\
&\quad = \sum_i \sum_\sigma \frac{1}{2} \sum_c \left|P_{i,\Pi(i)}(c,\sigma) - Q_{i,\Pi(i)}(c,\sigma)\right| + \sum_i \sum_\sigma \frac{1}{2} \left|Q_{\Pi(i)}(\sigma) - P_{\Pi(i)}(\sigma)\right| \sum_c Q_{i|\Pi(i)}(c|\sigma) \\
&\quad = \sum_i \frac{1}{2} \sum_\sigma \sum_c \left|P_{i,\Pi(i)}(c,\sigma) - Q_{i,\Pi(i)}(c,\sigma)\right| + \sum_i \sum_\sigma \frac{1}{2} \left|Q_{\Pi(i)}(\sigma) - P_{\Pi(i)}(\sigma)\right| \\
&\quad = \sum_i d_{\mathrm{TV}}\left(P_{i,\Pi(i)}, Q_{i,\Pi(i)}\right) + \sum_i d_{\mathrm{TV}}(P_i, Q_i) \\
&\quad \leq 2n \cdot d_{\mathrm{TV}}(P, Q).
\end{aligned}$$

The last inequality follows because the inequalities

$$d_{\mathrm{TV}}\big(P_{i,\Pi(i)}, Q_{i,\Pi(i)}\big) \leq d_{\mathrm{TV}}(P,Q)\,, \qquad d_{\mathrm{TV}}(P_i, Q_i) \leq d_{\mathrm{TV}}(P,Q)$$

hold.

A.8. Proof of Lemma 4.8

Let B be a Bayes net over variables $X_1, \ldots, X_n$ that has alphabet size $\ell = O(1)$ and treewidth $w = O(\log n)$. Let $S_1, \ldots, S_n \subseteq [\ell]$ be sets. The probabilistic inference task that we want to perform is to compute the probability $\mathbf{Pr}_B[X_1 \in S_1, \ldots, X_n \in S_n]$.

First, we construct the moralization of B (see Definition 2.4), namely M_B, in time $O(\mathrm{poly}(n))$ by invoking Lemma 2.5. Then, we use Theorem 2.10 to compute a tree decomposition $\mathcal{T}$ of M_B of width at most $4w + 1 \leq 5w$ in time $O\big(w 3^{3w} n^2\big)$. Finally, we use the variable elimination algorithm of Theorem 2.11 on B, $S_1, \ldots, S_n$, M_B, and $\mathcal{T}$ to compute $\mathbf{Pr}_B[X_1 \in S_1, \ldots, X_n \in S_n]$ in time $O\big(n\ell^{5w}\big)$.

The running time of this procedure is $O(\mathrm{poly}(n)) + O\big(w 3^{3w} n^2\big) + O\big(n\ell^{5w}\big) = O(\mathrm{poly}(n))$, whereby we have used the facts that $\ell = O(1)$ and $w = O(\log n)$. This concludes the proof.

B. TV Distance Between a Bayes Net and the Uniform Distribution

Here, we prove Theorem 1.3 and Theorem 1.4.

B.1. #P-Completeness

The main result of this subsection is Theorem 1.3. Recall that a function f from $\{0,1\}^*$ to non-negative integers is in the class #P if there is a polynomial time non-deterministic Turing machine M so that for any x, it is the case that $f(x)$ is equal to the number of accepting paths of $M(x)$.

We now prove Theorem 1.3.

Proof of Theorem 1.3. In what follows, we separately show membership in #P and #P-hardness.

Membership in #P. Let P be a Bayes net distribution over the Boolean domain $\{0,1\}^n$. The goal is to design a nondeterministic machine $\mathcal{N}$ so that the number of accepting paths of $\mathcal{N}$ (normalized by an appropriate quantity) equals $d_{\mathrm{TV}}(P, \mathbb{U})$. We will assume that the probabilities specified in the CPTs of the Bayes net for P are fractions. Let M be equal to 2^n times the product of the denominators of all the probabilities in the CPTs. The non-deterministic machine $\mathcal{N}$ first guesses an element $i \in \{0,1\}^n$ in the sample space of P, computes $|P(i) - 1/2|$ by using the CPTs, then guesses an integer $0 \leq z \leq M$, and finally accepts if and only if $1 \leq z \leq M|P(i) - 1/2|$. (Note that $M|P(i) - 1/2| = |M \cdot P(i) - M/2|$ is an integer.) It follows that

$$d_{\mathrm{TV}}(P, \mathbb{U}) = \frac{1}{2} \sum_{i \in \{0,1\}^n} \left| P(i) - \frac{1}{2} \right| = \frac{\text{number of accepting paths of } \mathcal{N}}{2M}$$

since the number of accepting paths of $\mathcal{N}$ is equal to $\sum_{i \in \{0,1\}^n} (M\,|P(i) - 1/2|)$ which is equal to $M \sum_{i \in \{0,1\}^n} |P(i) - 1/2|$, or $2M d_{\mathrm{TV}}(P, Q)$.

#P-Hardness. For the #P-hardness part, the proof gives a Turing reduction from the problem of counting the satisfying assignments of a CNF formula (which is #P-hard to compute) to computing the total variation distance between a Bayes net distribution and the uniform distribution. In what follows, by a graph of a formula we mean the DAG that captures the circuit structure of F, whereby the nodes are either AND, OR, NOT, or variable gates, and the edges correspond to wires connecting the gates.

Let F be a CNF formula viewed as a Boolean circuit. Assume F has n input variables $x_1, \ldots, x_n$ and m gates $\Gamma = \{y_1, \ldots, y_m\}$, where Γ is topologically sorted with y_m being the output gate. We will define a Bayes net distribution on some DAG G which, intuitively, is the graph of F.

The vertex set of G is split into two sets $\mathcal{X}$ and $\mathcal{Y}$, and a node Z. The set $\mathcal{X} = \{X_i\}_{i=1}^n$ contains n nodes with node X_i corresponding to variable x_i and the set $\mathcal{Y} = \{Y_i\}_{i=1}^m$ contains m nodes with each node Y_i corresponding to gate y_i. So totally there are $n + m + 1$ nodes. There is a directed edge from node V_i to node V_j if the gate/variable corresponding to V_i is an input to V_j.

The distribution P on G is given by a CPT defined as follows. Each X_i is a uniformly random bit. For each Y_i, its CPT is deterministic: For each of the setting of the parents Y_j, Y_k, namely y_j, y_k, the variable Y_i takes the value of the gate y_i for that setting of its inputs y_j, y_k. Finally, let Z be the value of Y_m OR-ed with a random bit.

Note that the formula F computes a Boolean function on the input variables. Let $f : \{0,1\}^n \to \{0,1\}$ be this function. We extend f to $\{0,1\}^m$ (i.e., $f : \{0,1\}^n \to \{0,1\}^m$) to also include the values of the intermediate gates.

With this notation in mind, for any binary string XYZ of length $n + m + 1$ it is the case that P has a probability 0 if $Y \neq f(X)$. Let $A := \{x \mid F(x) = 1\}$ and $R := \{x \mid F(x) = 0\}$.

To finish the proof, we will write the number of satisfying assignments of F, namely $|A|$, as a polynomial-time computable function of $d_{\mathrm{TV}}(P, \mathbb{U})$: We have

$$2 \cdot d_{\mathrm{TV}}(P, \mathbb{U}) = \sum_{X,Y,Z} |P - \mathbb{U}| = \underbrace{\sum_{\substack{X,Y,Z \\ Y \neq f(X)}} |P - \mathbb{U}|}_{(1)} + \underbrace{\sum_{\substack{X,Y,Z \\ Y = f(X)}} |P - \mathbb{U}|}_{(2)}$$

where we have abused the notation $P, \mathbb{U}$ to denote the probabilities $P(X, Y, Z)$, $\mathbb{U}(X, Y, Z)$. We will calculate (1) and (2) separately. For (1) we have:

$$\sum_{\substack{X,Y,Z \\ Y \neq f(X)}} |P - \mathbb{U}| = \sum_{\substack{X,Y,Z \\ Y \neq f(X)}} \left| 0 - \frac{1}{2^{n+m+1}} \right| = \frac{2^{n+1}(2^m - 1)}{2^{n+m+1}} = 1 - \frac{1}{2^m}.$$

For (2), we have

$$\sum_{\substack{X,Y,Z \\ Y = f(X)}} |P - \mathbb{U}| = \underbrace{\sum_{\substack{X,f(X),Z \\ X \in A}} |P - \mathbb{U}|}_{(3)} + \underbrace{\sum_{\substack{X,f(X),Z \\ X \in R}} |P - \mathbb{U}|}_{(4)}$$

and now we calculate the terms (3) and (4) separately. For (3), we have:

$$\begin{aligned}
\sum_{\substack{X,f(X),Z \\ X \in A}} |P - \mathbb{U}| &= \sum_{\substack{X,f(X),0 \\ X \in A}} |P - \mathbb{U}| + \sum_{\substack{X,f(X),1 \\ X \in A}} |P - \mathbb{U}| \\
&= \sum_{\substack{X,f(X),0 \\ X \in A}} \left| 0 - \frac{1}{2^{n+m+1}} \right| + \sum_{\substack{X,f(X),1 \\ X \in A}} \left| \frac{1}{2^n} - \frac{1}{2^{n+m+1}} \right| \\
&= \frac{|A|}{2^{n+m+1}} + \frac{|A| \cdot (2^{m+1} - 1)}{2^{n+m+1}} \\
&= \frac{|A|}{2^n}
\end{aligned}$$

and for (4) we have

$$\begin{aligned}
\sum_{\substack{X,f(X),Z \\ X \in R}} |P - \mathbb{U}| &= \sum_{\substack{X,f(X),0 \\ X \in R}} |P - \mathbb{U}| + \sum_{\substack{X,f(X),1 \\ X \in R}} |P - \mathbb{U}| \\
&= \sum_{\substack{X,f(X),0 \\ X \in R}} \left| \frac{1}{2^{n+1}} - \frac{1}{2^{n+m+1}} \right| + \sum_{\substack{X,f(X),1 \\ X \in R}} \left| \frac{1}{2^{n+1}} - \frac{1}{2^{n+m+1}} \right|
\end{aligned}$$

$$= \frac{|R| \cdot (2^m - 1) \cdot 2}{2^{n+m+1}}.$$

Thus

$$\begin{aligned} 2 \cdot d_{\mathrm{TV}}(P, \mathbb{U}) &= (1) + (2) \\ &= (1) + (3) + (4) \\ &= 1 - \frac{1}{2^m} + \frac{|A|}{2^n} + \frac{|R| \cdot (2^m - 1) \cdot 2}{2^{n+m+1}} \\ &= 2\left(1 - \frac{1}{2^m} + \frac{|A|}{2^{m+n+1}}\right) \end{aligned}$$

since $|A| + |R| = 2^n$. For that matter, $d_{\mathrm{TV}}(P, \mathbb{U}) = \frac{|A|}{2^{n+m+1}} + \left(1 - \frac{1}{2^m}\right)$ or

$$|A| = 2^{n+m+1}\left(d_{\mathrm{TV}}(P, \mathbb{U}) - \left(1 - \frac{1}{2^m}\right)\right).$$

That concludes the proof. □

B.2. Estimation in Randomized Polynomial Time

We prove Theorem 1.4.

Theorem 1.4 (Formal)**.** *There is an FPRAS for estimating the TV distance between a Bayes net P and the uniform distribution. Let n be the number of nodes of P, let ℓ be the size of its alphabet, and let d be its maximum in-degree. Then the running time of this FPRAS is $O\left(n^3 \ell^{2d+2} \varepsilon^{-2} \log \delta^{-1}\right)$ whereby ε is the accuracy error and δ is the confidence error of the FPRAS.*

Remark B.1. Note that the running time of the FPRAS of Theorem 1.4 is polynomial in the input length, as the description of the Bayes net P in terms of the CPTs has size at least $n + \ell^{d+1}$.

We shall now prove Theorem 1.4. We require the following lemma (which we will prove below).

Lemma B.2. *For all x, it is the case that*

$$1 - O\left(d_{\mathrm{TV}}(P, \mathbb{U})\, \ell^{d+1} n\right) \leq P(x)\, \ell^n \leq 1 + O\left(d_{\mathrm{TV}}(P, \mathbb{U})\, \ell^{d+1} n\right)$$

whenever $d_{\mathrm{TV}}(P, \mathbb{U}) \leq \frac{1}{16\ell^{d+1}}$.

The proof of Theorem 1.4 now resumes as follows. First, let us assume that $d_{\mathrm{TV}}(P, \mathbb{U}) \leq \frac{1}{16\ell^{d+1}}$ so that Lemma B.2 holds. We have that $d_{\mathrm{TV}}(P, \mathbb{U})$ is equal to

$$\begin{aligned} \frac{1}{2} \sum_x |P(x) - \mathbb{U}(x)| &= \sum_x \max(0, P(x) - \mathbb{U}(x)) \\ &= \sum_x \mathbb{U}(x) \max\left(0, \frac{P(x)}{\mathbb{U}(x)} - 1\right) \\ &= \mathop{\mathbf{E}}_{x \sim \mathbb{U}}\left[\max\left(0, \frac{P(x)}{\mathbb{U}(x)} - 1\right)\right] \\ &= \mathop{\mathbf{E}}_{x \sim \mathbb{U}}\left[\max(0, P(x)\, \ell^n - 1)\right]. \end{aligned}$$

This yields a natural estimator for $d_{\mathrm{TV}}(P, \mathbb{U})$, namely Est, as follows:

1. Sample $x_1, \ldots, x_m \sim \mathbb{U}$ for some value of m that we will fix later;
2. compute $\max(0, P(x_i)\, \ell^n - 1)$ for all $1 \leq i \leq m$;
3. output $(1/m) \sum_{i=1}^m \max(0, P(x_i)\, \ell^n - 1)$.

We will now prove the correctness and upper bound the running time of this procedure. We have from Hoeffding's inequality (Lemma 2.1) and Lemma B.2 that

$$\begin{aligned}
&\mathbf{Pr}[|\mathsf{Est} - d_{\mathrm{TV}}(P,\mathbb{U})| > \varepsilon d_{\mathrm{TV}}(P,\mathbb{U})] \\
&\quad = \mathbf{Pr}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\max(0, P(x_i)\,\ell^n - 1) - \mathop{\mathbf{E}}_{x\sim\mathbb{U}}[\max(0, P(x)\,\ell^n - 1)]\right| > \varepsilon d_{\mathrm{TV}}(P,\mathbb{U})\right] \\
&\quad = \mathbf{Pr}\left[\left|\sum_{i=1}^{m}\max(0, P(x_i)\,\ell^n - 1) - m\mathop{\mathbf{E}}_{x\sim\mathbb{U}}[\max(0, P(x)\,\ell^n - 1)]\right| > m\varepsilon d_{\mathrm{TV}}(P,\mathbb{U})\right] \\
&\quad \le 2\exp\left(-\frac{2m^2\varepsilon^2 d_{\mathrm{TV}}^2(P,\mathbb{U})}{\sum_{i=1}^{m}\left(0 - O(d_{\mathrm{TV}}(P,\mathbb{U})\,\ell^{d+1}n)\right)^2}\right) \\
&\quad = 2\exp\left(-\frac{2m^2\varepsilon^2 d_{\mathrm{TV}}^2(P,\mathbb{U})}{m\cdot O(d_{\mathrm{TV}}^2(P,\mathbb{U})\,\ell^{2d+2}n^2)}\right) \\
&\quad = 2\exp\left(-\frac{m\varepsilon^2}{O(\ell^{2d+2}n^2)}\right),
\end{aligned}$$

which is at most δ whenever $m = \Omega\big(n^2\ell^{2d+2}\varepsilon^{-2}\log\delta^{-1}\big)$.

The running time of this procedure is $O(mn) = O\big(n^3\ell^{2d+2}\varepsilon^{-2}\log\delta^{-1}\big)$, since we draw m samples and P can be evaluated on any sample in time $O(n)$.

If $d_{\mathrm{TV}}(P,\mathbb{U}) > \frac{1}{16\ell^{d+1}}$, then it suffices to additively approximate $d_{\mathrm{TV}}(P,\mathbb{U})$ up to error $\varepsilon/\left(16\ell^{d+1}\right)$. This can be done by Monte Carlo sampling using $m = \Omega\big(\ell^{2d+2}\varepsilon^{-2}\log\delta^{-1}\big)$ samples and $O(mn) = O\big(n\ell^{2d+2}\varepsilon^{-2}\log\delta^{-1}\big)$ time.

We now prove Lemma B.2.

Proof of Lemma B.2. Let us denote the maximum in-degree of P by d. Let X_0 be an arbitrary node with its parents as $X_1,\ldots,X_d$.

We have that $\gamma := d_{\mathrm{TV}}(P,\mathbb{U})$ is at least

$$\begin{aligned}
&d_{\mathrm{TV}}((X_0,\ldots,X_d),(Y_0,\ldots,Y_d)) \\
&\quad = \frac{1}{2}\sum_{v_0}\cdots\sum_{v_d}|\mathbf{Pr}[(X_0,\ldots,X_d) = (v_0,\ldots,v_d)] - \mathbf{Pr}[(Y_0,\ldots,Y_d) = (v_0,\ldots,v_d)]| \\
&\quad = \frac{1}{2}\sum_{v_0}\cdots\sum_{v_d}\left|\mathbf{Pr}[X_0 = v_0|X_1 = v_1,\ldots,X_d = v_d]\,\mathbf{Pr}[X_1 = v_1,\ldots,X_d = v_d] - \frac{1}{\ell^{d+1}}\right|
\end{aligned}$$

or

$$\frac{1}{2}\sum_{v_0}\cdots\sum_{v_d}\left|\mathbf{Pr}[X_0 = v_0|X_1 = v_1,\ldots,X_d = v_d]\,\mathbf{Pr}[X_1 = v_1,\ldots,X_d = v_d] - \frac{1}{\ell^{d+1}}\right| = \gamma$$

or

$$\left|\mathbf{Pr}[X_0 = v_0|X_1 = v_1,\ldots,X_d = v_d]\,\mathbf{Pr}[X_1 = v_1,\ldots,X_d = v_d] - \frac{1}{\ell^{d+1}}\right| \le 2\gamma, \tag{1}$$

for any $v_0,\ldots,v_d$. We observe the following.

Claim B.3. *We have that* $1/\ell^d - \gamma \le \mathbf{Pr}[X_1 = v_1,\ldots,X_d = v_d] \le 1/\ell^d + \gamma$.

Proof. Since $d_{\mathrm{TV}}(P,\mathbb{U}) = \gamma$ and $\mathbf{Pr}[Y_1 = v_1,\ldots,Y_d = v_d] = 1/\ell^d$, the claim is immediate. □

By Equation (1) and Claim B.3 we have the following.

Corollary B.4. *For* $\gamma < 1/\left(2\ell^d\right)$ *we have that*

$$|\mathbf{Pr}[X_0 = v_0|X_1 = v_1,\ldots,X_d = v_d] - 1/\ell| \le 8\gamma\ell^d.$$

Proof. By Equation (1) we have

$$\frac{1}{\ell^{d+1}} - 2\gamma \leq \mathbf{Pr}[X_0 = v_0 | X_1 = v_1, \ldots, X_d = v_d]\, \mathbf{Pr}[X_1 = v_1, \ldots, X_d = v_d] \leq \frac{1}{\ell^{d+1}} + 2\gamma$$

or

$$\frac{\frac{1}{\ell^{d+1}} - 2\gamma}{\mathbf{Pr}[X_1 = v_1, \ldots, X_d = v_d]} \leq \mathbf{Pr}[X_0 = v_0 | X_1 = v_1, \ldots, X_d = v_d] \leq \frac{\frac{1}{\ell^{d+1}} + 2\gamma}{\mathbf{Pr}[X_1 = v_1, \ldots, X_d = v_d]}$$

or, by making use of Claim B.3,

$$\frac{\frac{1}{\ell^{d+1}} - 2\gamma}{\frac{1}{\ell^{d}} + \gamma} \leq \mathbf{Pr}[X_0 = v_0 | X_1 = v_1, \ldots, X_d = v_d] \leq \frac{\frac{1}{\ell^{d+1}} + 2\gamma}{\frac{1}{\ell^{d}} - \gamma}$$

or

$$\frac{\frac{1}{\ell} - 2\ell^d\gamma}{1 + \ell^d\gamma} \leq \mathbf{Pr}[X_0 = v_0 | X_1 = v_1, \ldots, X_d = v_d] \leq \frac{\frac{1}{\ell} + 2\ell^d\gamma}{1 - \ell^d\gamma}.$$

We now have

$$\begin{aligned}
\mathbf{Pr}[X_0 = v_0 | X_1 = v_1, \ldots, X_d = v_d] &\leq \frac{\frac{1}{\ell} + 2\ell^d\gamma}{1 - \ell^d\gamma} \\
&\leq \left(\frac{1}{\ell} + 2\ell^d\gamma\right)\left(1 + 2\ell^d\gamma\right) \\
&= \frac{1}{\ell} + 2\ell^{d-1}\gamma + 2\ell^d\gamma + 4\ell^{2d}\gamma^2 \\
&\leq \frac{1}{\ell} + 2\ell^d\gamma + 2\ell^d\gamma + 4\ell^d\gamma \\
&= \frac{1}{\ell} + 8\ell^d\gamma,
\end{aligned}$$

since $1/\left(1 - x\right) \leq 1 + 2x$ for $x < 1/2$ (here $x = \ell^d\gamma < 1/2$), and

$$\begin{aligned}
\mathbf{Pr}[X_0 = v_0 | X_1 = v_1, \ldots, X_d = v_d] &\geq \frac{\frac{1}{\ell} - 2\ell^d\gamma}{1 + \ell^d\gamma} \\
&\geq \left(\frac{1}{\ell} - 2\ell^d\gamma\right)\left(1 - \ell^d\gamma\right) \\
&= \frac{1}{\ell} - \ell^{d-1}\gamma - 2\ell^d\gamma + 2\ell^{2d}\gamma^2 \\
&\geq \frac{1}{\ell} - \ell^d\gamma - 2\ell^d\gamma \\
&\geq \frac{1}{\ell} - 8\gamma\ell^d,
\end{aligned}$$

since $1/\left(1 + x\right) \geq 1 - x$ for $x < 1/2$ (here $x = \ell^d\gamma < 1/2$). □

The result now follows from the observation that

$$\left(1/\ell - 8\gamma\ell^d\right)^n \leq P(x) = \prod_{i=1}^{n} \mathbf{Pr}\left[X_i = x_i | X_{\Pi(X_i)} = x_{\Pi(X_i)}\right] \leq \left(1/\ell + 8\gamma\ell^d\right)^n$$

or

$$\left(1 - 8\gamma\ell^{d+1}\right)^n \leq P(x)\,\ell^n \leq \left(1 + 8\gamma\ell^{d+1}\right)^n$$

or

$$1 - 16\gamma\ell^{d+1}n \leq P(x)\,\ell^n \leq 1 + 16\gamma\ell^{d+1}n,$$

whereby we used the facts that $(1 - \alpha)^k \geq (1 - 2\alpha k)$ and $(1 + \alpha)^k \leq (1 + 2\alpha k)$ whenever $\alpha < 1/2$ and $k > 0$, and the fact that $\gamma < 1/\left(16\ell^{d+1}\right)$ or $8\gamma\ell^{d+1} < 1/2$.

Finally, we have

$$1 - 16 d_{\mathrm{TV}}(P, \mathbb{U})\,\ell^{d+1}n \leq P(x)\,\ell^n \leq 1 + 16 d_{\mathrm{TV}}(P, \mathbb{U})\,\ell^{d+1}n,$$

as desired. □