Efficient Distance Approximation for Structured High-Dimensional Distributions via Learning*

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

Sutanu Gayen
National University of Singapore
sutanugayen@gmail.com

Kuldeep S. Meel
National University of Singapore
meel@comp.nus.edu.sg

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

Abstract

We design efficient distance approximation algorithms for several classes of well-studied structured high-dimensional distributions. Specifically, we present algorithms for the following problems (where d_{TV} is the total variation distance):

- Given sample access to two Bayesian networks P_1 and P_2 over known directed acyclic graphs G_1 and G_2 having n nodes and bounded in-degree, approximate $d_{\mathrm{TV}}(P_1, P_2)$ to within additive error ε using $\mathrm{poly}(n, \varepsilon^{-1})$ samples and time.
- Given sample access to two ferromagnetic Ising models P_1 and P_2 on n variables with bounded width, approximate $d_{\mathrm{TV}}(P_1, P_2)$ to within additive error ε using $\mathrm{poly}(n, \varepsilon^{-1})$ samples and time.
- Given sample access to two n-dimensional Gaussians P_1 and P_2, approximate $d_{\mathrm{TV}}(P_1, P_2)$ to within additive error ε using $\mathrm{poly}(n, \varepsilon^{-1})$ samples and time.
- Given access to observations from two causal models P and Q on n variables that are defined over known causal graphs, approximate $d_{\mathrm{TV}}(P_a, Q_a)$ to within additive error ε using $\mathrm{poly}(n, \varepsilon^{-1})$ samples and time, where P_a and Q_a are the interventional distributions obtained by the intervention $\mathsf{do}(A = a)$ on P and Q respectively for a particular variable A.

The distance approximation algorithms immediately imply new *tolerant closeness testers* for the corresponding classes of distributions. Prior to our work, only *non-tolerant testers* were known for both Bayes net distributions and Ising models, and no testers with quantitative guarantees were known for interventional distributions. To the best of our knowledge, efficient distance approximation algorithms for Gaussian distributions were not present in the literature. Our algorithms are designed using a conceptually simple but general framework that is applicable to a variety of scenarios.

1 Introduction

Machine learning is primarily concerned with the design of techniques to enable the learning of a generative model $\mathcal{M}$ given access to data $\mathcal{D}$ arising from another distribution, say P [Mur12]. While P is typically an unknown distribution, the design of a new ML technique is often accompanied by empirical and theoretical studies under certain assumptions on P. Let Q be the distribution generated by $\mathcal{M}$; then ideally, one would learn $\mathcal{M}$ such P and Q are as close as possible. Given the widespread

*Authors are in the alphabetical order.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

adoption of machine learning techniques in critical domains, there has been a surge in interest of the design of techniques for rigorous verification of machine learning systems [SSS16]. The development of such verification techniques would necessitate the development of algorithmic techniques for rigorous approximation of the distance between two distributions P and Q.

Distance approximation is also closely related to the topic of *distribution testing* investigated in the statistics and algorithms communities. Two important testing problems are *identity testing* (or, *goodness-of-fit testing*) and *closeness testing* (or, *two-sample testing*). Given samples from an unknown distribution P over a domain $\mathcal{S}$, the problem of identity testing seeks to ask whether P equals a specific reference distribution Q. A sequence of works [Pan08, BFR+13, VV17, CDVV14] in the property testing literature has pinned down the finite sample complexity of this problem. It is known that with $O(|\mathcal{S}|^{1/2}\varepsilon^{-2})$ samples from P, one can, with probability at least $2/3$, distinguish whether $P = Q$ or whether $d_{\mathrm{TV}}(P, Q) > \varepsilon$; also, $\Omega(|\mathcal{S}|^{1/2}\varepsilon^{-2})$ samples are necessary for this task. An important generalization of identity testing is closeness testing: given samples from two unknown distributions P and Q over $\mathcal{S}$, does $P = Q$? Here, $\Theta(|\mathcal{S}|^{2/3}\varepsilon^{-4/3} + |\mathcal{S}|^{1/2}\varepsilon^{-2})$ samples are necessary and sufficient to distinguish $P = Q$ from $d_{\mathrm{TV}}(P, Q) > \varepsilon$ with probability at least $2/3$. The corresponding algorithms for both identity and closeness testing run in time polynomial in $|\mathcal{S}|$ and ε^{-1}. However, in order to solve these testing problems in many real-life settings, there are two issues that need to be surmounted.

- ***High dimensions:*** In typical applications, the data is described using a huge number of (possibly redundant) features; thus, each item in the dataset is represented as a point in a high-dimensional space. If $\mathcal{S} = \Sigma^n$, then from the results quoted above, identity testing or closeness testing for arbitrary probability distributions over $\mathcal{S}$ requires $2^{\Omega(n)}$ many samples, which is clearly unrealistic. Hence, we need to restrict the class of input distributions.
- ***Approximation:*** A high-dimensional distribution requires a large number of parameters to be specified. So, for identity testing, it is unlikely that we can ever hypothesize a reference distribution Q such that it exactly equals the data distribution p. Similarly, for closeness testing, two data distributions P and Q are most likely not exactly equal. Hence, we would like to design *tolerant* testers for identity and closeness that distinguish between the cases $d_{\mathrm{TV}}(P, Q) \leqslant \varepsilon_1$ and $d_{\mathrm{TV}}(P, Q) > \varepsilon_2$ where $\varepsilon_1 < \varepsilon_2$ are user-supplied parameters.

We address both these issues by focusing on designing distance approximation algorithms for certain classes of structured distributions over Σ^n, where Σ is an arbitrary finite set.

Definition 1.1. *Let $\mathcal{D}_1, \mathcal{D}_2$ be two families of distributions over Σ^n. A* distance approximation algorithm for $(\mathcal{D}_1, \mathcal{D}_2)$ *is a randomized algorithm $\mathcal{A}$ which takes as input $\varepsilon \in (0, 1)$, and sample access to two unknown distributions $P \in \mathcal{D}_1, Q \in \mathcal{D}_2$. The algorithm $\mathcal{A}$ returns as output a value $\gamma \in [0, 1]$ such that, with probability[†] at least $2/3$:*

$$\gamma - \varepsilon \leqslant d_{\mathrm{TV}}(P, Q) \leqslant \gamma + \varepsilon.$$

If $\mathcal{D}_1 = \mathcal{D}_2 = \mathcal{D}$, then we refer to such an algorithm as a distance approximation algorithm for $\mathcal{D}$.

Equivalence of distance approximation and tolerant testing: Designing distance approximation algorithms is essentially equivalent to designing tolerant testing algorithms. Indeed, Parnas et al. [PRR06] observed that existence of a distance approximation with sample/time complexity $F(\varepsilon, n)$ for two families of distributions implies a tolerant testing algorithm with complexity $F\left(\frac{\varepsilon_2 - \varepsilon_1}{2}, n\right)$; and conversely, existence of a tolerant testing algorithm with sample/time complexity $F(\varepsilon_2 - \varepsilon_1, n)$ implies an algorithm for distance approximation with sample/time complexity $O(\log(1/\varepsilon) \log\log(1/\varepsilon)) \cdot F(2\varepsilon, n)$. Thus, henceforth we use "distance approximation" and "tolerant testing" interchangeably.

In this work, we design the first computational and sample efficient distance approximation algorithms (equivalently tolerant testing algorithms) for a variety of structured high-dimensional distributions: Bayesian networks, Ising Models, multivariate Gaussians, and interventional distributions arising from causal Bayesian networks. Our results advance the state-of-the-art in the following way:

1. Our algorithm for testing distributions over Bayes nets extends prior work [DP17, CDKS17]. In particular, in [DP17], Daskalakis and Pan presented an algorithm for *non-tolerant* closeness

[†]The success probability can be amplified to $1 - \delta$ by taking the median of $O(\log \delta^{-1})$ independent repetitions of the algorithm with success probability $2/3$.

testing of two Bayes net distributions P and Q over the same known graph [‡]. We present tolerant closeness testing algorithm for two Bayes net distributions P and Q over *two different graphs* that asymptotically matches the sample and time complexity of their algorithm.

2. We design efficient tolerant testers for Ising models. Our first algorithm approximates the distance between any two ferromagnetic Ising models. Our second algorithm approximates the distance between any Ising model and the uniform distribution. Previously proposed testing algorithms for Ising models by [DDK19] do not achieve non-trivial tolerance.

3. Given access to $\text{poly}(n)$ samples from two multivariate Gaussians over $\mathbb{R}^n$, it is a folklore that one can approximate the distance between them. However, that algorithm is not computationally efficient. We design the first efficient algorithm to approximate distance between two multivariate Gaussians, to the best of our knowledge.

4. Given observations from two causal models P and Q described by two Bayesian networks on the same variable set, we give an efficient algorithm to approximate the distance between the interventional distributions obtained by fixing a particular variable. Celebrated work of Tian and Pearl [TP02a, Tia02] gave identifiability conditions. However efficient distance approximation algorithms with finite sample guarantees were non-existent prior to our work.

All our algorithms are based on a common framework. To approximate the distance between $P \in \mathcal{D}_1$ and $Q \in \mathcal{D}_2$, we first *learn* the model parameters for $\hat{P} \in \mathcal{D}_1$ and $\hat{Q} \in \mathcal{D}_2$ that are guaranteed to be close to P and Q respectively. It remains to compute $d_{\mathrm{TV}}(\hat{P}, \hat{Q})$. This is a computationally hard problem in general, but we use the fact that for $\mathcal{D}_1, \mathcal{D}_2$ of interest, we can efficiently approximate the mass functions for $\hat{P}$ and $\hat{Q}$ from their parameters. At this point, we invoke an estimator that approximates $d_{\mathrm{TV}}(\hat{P}, \hat{Q})$ using samples from P and the approximate mass functions for $\hat{P}$ and $\hat{Q}$.

A salient strength of our framework is its conceptual simplicity. In fact we believe that the conceptual simplicity allowed us to apply the framework to a variety of situations leading to algorithms that are potentially amenable to practical implementations. As a first step, we restricted our focus to the above mentioned classical models to capture probabilistic distribution. A natural extension of this work would be to apply our techniques for rigorous verification and testing of neural network models such as Generative Adversarial Networks (GANs) wherein a discriminator is inherently tasked with performing *closeness-testing* for the given data distribution and distribution arising from the generator [LKFO18].

2 Previous work

Prior work most related to our work is in the area of distribution testing. The topic of distribution testing is rooted in statistical hypothesis testing and goes back to Pearson's chi-squared test in 1900. In theoretical computers science, distribution testing research is relatively new and focuses on designing hypothesis testers with optimal sample complexity. Goldreich and Ron [GR11] investigated uniformity testing (distinguishing whether an input distribution P is uniform over its support or ε-far from uniform in total variation distance) and designed a tester with sample complexity $O(m/\varepsilon^4)$ (where m is the size of the sample space). Paninski [Pan08] showed that $\Theta(\sqrt{m}/\varepsilon^2)$ samples are necessary for uniformity testing, and gave an optimal tester when $\varepsilon > m^{-1/4}$. Batu et al. [BFR+13] initiated the investigation of identity (goodness-of-fit) testing and closeness (two-sample) testing and gave testers with sample complexity $\tilde{O}(\sqrt{m}/\varepsilon^6)$ and $\tilde{O}(m^{2/3}\text{poly}(1/\varepsilon))$ respectively. Optimal bounds for these testing problems were obtained in Valiant and Valiant [VV17] ($\Theta(\sqrt{m}/\varepsilon^2)$) and Chan et al. [CDVV14] ($\Theta(\max(m^{2/3}\varepsilon^{-4/3}, \sqrt{m}\varepsilon^{-2}))$) respectively. Tolerant versions of these testing problems have very different sample complexity. In particular, Valiant and Valiant [VV11b, VV10] showed that tolerant uniformity, identity, and closeness testing with respect to the total variation distance have a sample complexity of $\Theta(m/\log m)$. Since the seminal papers of Goldreich and Ron and Batu et al., distribution testing grew into a very active research topic and a wide range of properties of distributions have been studied under this paradigm. This research led to sample-optimal testers for many distribution properties. We refer the reader to the surveys [Can15, Rub12] and references therein for more details and results on the topic.

[‡]They also present non-tolerant testers for the case when the underlying graph is unknown.

When the sample space is a high-dimensional space (such as $\{0,1\}^n$)), the testers designed for general distributions require exponential number of samples ($2^{\Omega(n)}$) if the sample space is $\{0,1\}^n$ for a constant ε). Thus structural assumptions are to be made to design efficient ($\mathrm{poly}(n, 1/\varepsilon)$) and practical testers for many of the testing problems. The study of testing high-dimensional distributions with structural restrictions was initiated only very recently. The work that is most closely related to our work appears in [DDK19, CDKS17, DP17, ABDK18] (these works also give good expositions to other prior work on this topic). These papers consider distributions coming from graphical models including Ising models and Bayes nets. In Daskalakis et al. [DDK19], the authors consider distributions that are drawn from an Ising model and show that identity testing and *independence testing* (testing whether an unknown distribution is close to a product distribution) can be done with $\mathrm{poly}(n, 1/\varepsilon)$ samples where n is the number nodes in the graph associated with the Ising model. In Canonne et al. [CDKS17] and Daskalakis et al. [DP17], the authors consider identity testing and closeness testing for distributions given by Bayes networks of bounded in-degree. Specifically, they design algorithms with sample complexity $\tilde{O}(2^{3(d+1)/4}n/\varepsilon^2)$ that test closeness of distributions over the same Bayes net with n nodes and in-degree d. They also show that $\Theta(\sqrt{n}/\varepsilon^2)$ and $\Theta(\max(\sqrt{n}/\varepsilon^2, n^{3/4}/\varepsilon))$ samples are necessary and sufficient for identity testing and closeness testing respectively of pairs of product distributions (Bayes net with empty graph). Finally, in Acharya et al.[ABDK18], the authors investigate testing problems on *causal Bayesian networks* as defined by Pearl [Pea09] and design efficient $(\mathrm{poly}(n, 1/\varepsilon))$ testing algorithms for certain identity and closeness testing problems for them. All these papers consider designing non-tolerant testers and leave open the problem of designing efficient testers that are tolerant for high-dimensional distributions which is the main focus in this paper.

Our main technical result builds on the work of Canonne and Rubinfeld [CR14]. They consider a *dual access model* for testing distributions. In this model, in addition to independent samples, the testing algorithm has also access to an evaluation oracle that gives probability of any item in the sample space. They establish that having access to the evaluation oracle leads to testing algorithms with sample complexity independent of the size of the sample space. Indeed, in order to design testing algorithms, they give an algorithm to additively estimate the total variation distance between two unknown distributions in the dual access model. Our distance estimation algorithm is a direct extension of this algorithm. *Conditional sampling model* has been another related model of interest recently [CFGM16, CRS14, CM19].

Novelty of our work: We would like to emphasize that the core conceptual and novel contribution of our work is the establishment of a connection between testing in the dual access model (and in the conditional sampling model) to testing and distance approximation in the standard sampling model. These two models have been investigated separately. Here we use the former results to derive several new efficient tolerant testing algorithms in the standard model for high dimensional distributions, thus extending the state-of-the-art in this area. In this regard, we extend [CR14] to derive Algorithm 1, which in our view is intended to be simple and flexible. We consider the simplicity of Algorithm 1 a core strength of our work.

Comparison with [CR14]: Technically, [CR14] assumes a perfect access to the probability mass functions of the two distributions. Instead we work with approximate access to p.m.f.s, the approximation being parameterized by β and γ. In our opinion, this generalization (in Appendix A) does not follow trivially. The usage of approximation has allowed us to obtain results for several high dimensional distributions that do not follow directly from [CR14]. For example, let us consider the Ising model. In this case, given samples from two ferromagnetic Ising models P and Q, we approximately learn the model parameters [KM17] and estimate the partition functions [JS93], to evaluate the p.m.f.s approximately. The later result takes parameters of a ferromagnetic Ising model as input and returns a (randomized) PTIME $(1 \pm \gamma)$-multiplicative approximation of its partition function, and therefore we obtain a PTIME algorithm. In contrast, since the computation of the partition function given a fully known ferromagnetic Ising model is known to be $\#\mathsf{P}$-complete [JS93] (see Theorem 15 of their paper) and as the algorithm given in [CR14] does not allow for multiplicative errors, directly applying it would lead to an algorithm with $\mathsf{P}^{\#\mathsf{P}}$ complexity. The approximation parameter β was used for designing a distance approximation algorithm for all four classes considered in this paper.

3 Main Result

We first formalize the connection between learning and distance approximation, and then we give our main algorithm for distance approximation. In the next section, we detail the implications for several well-studied families of structured high-dimensional probability distributions.

Given a family of distributions $\mathcal{D}$, a learning algorithm for $\mathcal{D}$ is an algorithm $\mathcal{L}$ that on input $\varepsilon \in (0,1)$ and sample access to a distribution $P \in \mathcal{D}$, returns the description of a distribution $\hat{P}$ such that with probability at least $2/3$, $d_{\mathrm{TV}}(P, \hat{P}) \leqslant \varepsilon$.

Our framework for distance approximation needs to (approximately) evaluate the mass function $\hat{P}(x) \coloneqq \mathbf{Pr}_{X \sim \hat{P}}[X = x]$ for any $x \in \Sigma^n$. More precisely, we require EVAL *approximators*:

Definition 3.1. *Let P be a distribution over a finite set U. A function $E_P : U \to [0,1]$ is a (β, γ)-*EVAL approximator *for P if there exists a distribution $\hat{P}$ over U such that*

- $d_{\mathrm{TV}}(P, \hat{P}) \leqslant \beta$
- $\forall x \in U$, $(1-\gamma) \cdot \hat{P}(x) \leqslant E_P(x) \leqslant (1+\gamma) \cdot \hat{P}(x)$

In our applications, we first use a learning algorithm to obtain parameters that describe $\hat{P}$, and then we compute (or approximate) $\hat{P}(x)$ efficiently in terms of these parameters.

Example 3.2. Suppose $\mathcal{D}$ is the family of product distributions on $\{0,1\}^n$. That is, any $P \in \mathcal{D}$ can be described in terms of n parameters $p_1, \ldots, p_n$ where each p_i is the probability of the i'th coordinate being 1. It is folklore (see e.g. [ADK15]) that there is a learning algorithm which gets $O(n\varepsilon^{-2})$ samples from P and returns the parameters $\hat{p}_1, \ldots, \hat{p}_n$ of a product distribution $\hat{P}$ satisfying $d_{\mathrm{TV}}(P, \hat{P}) \leqslant \varepsilon$ with probability $2/3$. It is clear that given $\hat{p}_1, \ldots, \hat{p}_n$, we can compute $\hat{P}(x)$ for any $x \in \{0,1\}^n$ in linear time as: $\hat{P}(x) = \prod_{i=1}^{n} (x_i \cdot \hat{p}_i + (1-x_i) \cdot (1-\hat{p}_i))$. Thus, there is an algorithm that takes as input sample access to any product distribution P, has sample and time complexity $O(n\varepsilon^{-2})$, and returns a circuit implementing an $(\varepsilon, 0)$-EVAL approximator for P. Moreover, any call to the circuit returns in $O(n)$ time.

We establish the following link between EVAL approximators and distance approximation, achieved using Algorithm 1. Its proof can be found in Appendix A.

Theorem 3.3. *Suppose we have sample access to distributions P and Q over a finite set. Also, suppose we have access to $(\varepsilon, \varepsilon)$-*EVAL *approximators for P and Q. Then, with probability at least 2/3, $d_{\mathrm{TV}}(P, Q)$ can be approximated to within $O(\varepsilon)$ additive error using $O(\varepsilon^{-2})$ samples from P and $O(\varepsilon^{-2})$ calls to the two* EVAL *approximators.*

Algorithm 1: Distance approximation between P and Q

Input : Sample access to distribution P; oracle access to $(\varepsilon, \varepsilon)$-EVAL approximators $\mathcal{C}_P$ and $\mathcal{C}_Q$ for P and Q respectively.
Output : Approximate value of $d_{\mathrm{TV}}(P, Q)$

for $i = 1, \ldots, t = O(\varepsilon^{-2})$ **do**
  Draw a sample x from P;
  $a \leftarrow \mathcal{C}_P(x)$;
  $b \leftarrow \mathcal{C}_Q(x)$;
  $c_i \leftarrow 1_{a>b} \left(1 - \frac{b}{a}\right)$;
return $\frac{1}{t} \sum_{i=1}^{t} c_i$

Thus, in the context of Example 3.2, the above theorem immediately implies a distance approximation algorithm for product distributions using $O(n\varepsilon^{-2})$ samples and time. Theorem 3.3 extends the work of Canonne and Rubinfeld [CR14] who considered the setting $\beta = \gamma = 0$. We discussed the relation to prior work in Section 2.

Testing, learning, and efficiency: It is natural to ask whether we can design substantially more efficient distance approximation (or tolerant testing) algorithms than the ones that are possible via

learning as we do in this paper. We discuss this from the perspective of both sample complexity as well as time complexity.

It is clear that the sample complexity of distance approximations is at most that of learning: from the learnt distributions we can compute the distance using a brute-force algorithm (not computationally efficient). On the other hand, current known results give evidence that typically it is not possible to substantially improve the dependence on the dimension (n), at least for the following two edge cases. Valiant and Valiant [VV11a] have shown that given samples from an unknown distribution over $[m]$, approximating its distance to the uniform distribution up to a constant additive error with 2/3 probability requires $\Omega(m/\log m)$ samples. In contrast, it is well known that we can learn an unknown distribution within constant error with 2/3 success probability using only $O(m)$ samples. Similarly, in the case of high-dimensional distributions over $\{0,1\}^n$, Canonne et al. [CDKS17] have shown that there exists two *product distributions* whose distance approximation up to a constant error with 2/3 probability requires $\Omega(n/\log n)$ samples, whereas an unknown product distribution can be learnt in constant error with 2/3 probability in $O(n)$ samples. Thus typically sample complexities of learning and distance approximation differ only by a logarithmic factor. However, if one is interested in *non-tolerant* testing, substantial improvements are possible. In particular, for the above problems there are algorithms with $O(\sqrt{m})$ [GR11, Pan08] and $O(\sqrt{n})$ [DP17, CDKS17] sample complexity respectively.

From a time complexity perspective, even if we assume that the learning is perfect, computational efficiency remains a challenge for distance estimation in many high-dimensional settings. Sahai and Vadhan [SV03] have shown that tolerant testing of distributions encoded by Boolean circuits is a problem that is complete for the class SZK (problems admitting statistical zero knowledge interactive proofs). The class SZK contains several hard computational problems including Graph Isomorphism. Kiefer [Kie18] has shown that given two completely specified hidden Markov models, it is #P-hard to additively approximate their distance. Bogdanov et al. [BMV08] have shown that given two completely specified Markov Random Fields with hidden variables, it is impossible to approximate their distance in randomized polynomial time unless NP = RP.

By coupling learning algorithms with the template for distance approximation given by Theorem 3.3, we present a number of scenarios where sample and computational efficient distance approximation algorithms can be designed. We also describe a generic method to efficiently improve the success probability of learning algorithms for the families of distributions admitting a fast distance approximation algorithm, which is presented in Appendix F.

4 Applications

4.1 Bayesian Networks

A standard way to model structured high-dimensional distributions is through *Bayesian networks*. A Bayesian network describes how a collection of random variables can be generated one-at-a-time in a directed fashion, and they have been used to model beliefs in a wide variety of domains (see [JN07, KF09] for many pointers to the literature). Formally, a probability distribution P over n variables $X_1, \ldots, X_n \in \Sigma$ is said to be a *Bayesian network on a directed acyclic graph* G with n nodes if[§] for every $i \in [n]$, X_i is conditionally independent of $X_{\text{non-descendants}(i)}$ given $X_{\text{parents}(i)}$. Equivalently, P admits the factorization:

$$P(x) \coloneqq \mathop{\mathbf{Pr}}_{X\sim P}[X = x] = \prod_{i=1}^{n} \mathop{\mathbf{Pr}}_{X\sim P}[X_i = x_i \mid \forall j \in \text{parents}(i), X_j = x_j] \qquad \text{for all } x \in \Sigma^n \quad (1)$$

For example, product distributions are Bayesian networks on the empty graph.

Invoking our framework of distance approximation via EVAL approximators on Bayesian networks, we obtain the following:

Theorem 4.1. *Suppose G_1 and G_2 are two DAGs on n vertices with in-degree at most d. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be the family of Bayesian networks on G_1 and G_2 respectively. Then, there is a distance approximation algorithm for $(\mathcal{D}_1, \mathcal{D}_2)$ that gets $m = \tilde{O}(|\Sigma|^{d+1} n\varepsilon^{-2})$ samples and runs in $O(mn)$ time.*

[§]We use the notation X_S to denote $\{X_i : i \in S\}$ for a set $S \subseteq [n]$.

Theorem 4.1 extends the works of Daskalakis et al. [DP17] and Canonne et al. [CDKS17] who designed efficient *non-tolerant* identity and closeness testers for Bayesian networks. Their arguments appear to be inadequate to design tolerant testers. In addition, their results for general Bayesian networks were restricted to the case when $G_1 = G_2$. Theorem 4.1 immediately gives efficient *tolerant* identity and closeness testers for Bayesian networks even when $G_1 \neq G_2$. Canonne et al. [CDKS17] obtain better sample complexity but they make certain *balancedness* assumption on each conditional probability distribution. Without such assumptions, the sample complexity of our algorithm is optimal.

Theorem 4.1 relies on a new learning algorithm for Bayesian networks on a known DAG G that may be of independent interest. It uses $\tilde{O}(n\varepsilon^{-2}|\Sigma|^{d+1})$ samples where d is the maximum in-degree. It returns another Bayesian network $\hat{P}$ on G, described in terms of the conditional probability distributions $X_i \mid x_{\text{parents}(i)}$ for all $i \in [n]$ and all settings of $x_{\text{parents}(i)} \in \Sigma^{\deg(i)}$. The sample complexity of the algorithm is nearly optimal. Such a learning algorithm was claimed in the appendix of [CDKS17], but the analysis there appears to be incomplete with no immediate fix [Can20].

4.2 Ising Models

Another widely studied model of high-dimensional distributions is the *Ising model*. It was originally introduced in statistical physics as a way to study spin systems ([Isi25]) but has since emerged as a versatile framework to study other systems with pairwise interactions, e.g., social networks ([MS10]), learning in coordination games ([Ell93]), phylogeny trees in evolution ([Ney71, Far73, Cav78]) and image models for computer vision ([GG86]). Formally, a distribution P over variables $X_1, \ldots, X_n \in \{-1, 1\}$ is an *Ising model* if for all $x \in \{-1, 1\}^n$:

$$P(x) = \frac{\exp\left(\sum_{i \neq j \in [n]} A_{ij} x_i x_j + \theta \sum_{i \in [n]} x_i\right)}{\sum_{z \in \{-1,1\}^n} \exp\left(\sum_{i \neq j \in [n]} A_{ij} z_i z_j + \theta \sum_{i \in [n]} z_i\right)} \tag{2}$$

where $\theta \in \mathbb{R}$ is called the *external field* and A_{ij} are called the *interaction terms*. An Ising model is called *ferromagnetic* if all $A_{ij} \geqslant 0$. The *width* of an Ising model as in (2) is $\max_i \sum_j |A_{ij}| + |\theta|$.

Invoking our framework on Ising models, we obtain:

Theorem 4.2. *Let $\mathcal{D}$ be the family of ferromagnetic Ising models having width at most d. Then, there is a distance approximation algorithm for $\mathcal{D}$ with sample complexity $m = e^{O(d)}\varepsilon^{-4}n^8 \log(\frac{n}{\varepsilon})$ and runtime $O(mn^2 + \varepsilon^{-2}n^{17}\log n)$.*

We use the parameter learning algorithm by Klivans and Meka [KM17] that learns the parameters $\hat{\theta}, \hat{A}_{ij}$ of another Ising model $\hat{P}$ such that $\hat{P}(x)$ is a $(1 \pm \varepsilon)$ approximation of $P(x)$ for every x. This results holds for any Ising model, ferromagnetic or not. But in order to get an EVAL approximator, we need to compute $\hat{P}(x)$ from $\hat{\theta}, \hat{A}_{ij}$. In general, the partition function (i.e., the sum in the denominator of Equation (2)) may be #P-hard to compute, but for ferromagnetic Ising models, Jerrum and Sinclair [JS93] gave a PTAS for this problem. Thus, we obtain an $(\varepsilon, \varepsilon)$-EVAL approximator for ferromagnetic Ising models that runs in polynomial time, and then Theorem 4.2 follows from Theorem 3.3.

Daskalakis et al. [DDK19] studied independence testing and identity testing for Ising models and design *non-tolerent* testers. Their sample and time complexity have polynomial dependence on the width instead of exponential (as in our case), but their algorithms seem to be inherently non-tolerant. In contrast, our distance approximation algorithm leads to a tolerant closeness-testing algorithm for ferromagnetic Ising models. Also, Theorem 4.2 offers a template for distance approximation algorithms whenever the partition function can be approximated efficiently. In particular, Sinclair et al [SST14] showed a PTAS for computing the partition function of anti-ferromagnetic Ising models in certain parameter regimes.

We also show that we can efficiently approximate the distance to uniformity for any Ising model.

Theorem 4.3. *There is an algorithm which, given independent samples from an unknown Ising model P over $\{-1, 1\}^n$ with width at most d, takes $m = O(e^{O(d)}\varepsilon^{-4}n^8 \log(n/\varepsilon))$ samples, $O(mn^2)$ time and returns a value e such that $|e - d_{\mathrm{TV}}(P, U)| \leqslant \varepsilon$ with probability at least 7/12, where U is the uniform distribution over $\{-1, 1\}^n$.*

The proof of Theorem 4.3 proceeds by learning the parameters $\hat{\theta}, \hat{A}$ of an Ising model $\hat{P}$ that is a multiplicative approximation fo P. As we mentioned earlier, computing the partition function is in general hard. However, we can efficiently estimate the ratio $P(x)/P(y)$ for any two $x, y \in \{-1, 1\}^n$. At this point, we invoke the uniformity tester by Narayanan [Nar21] that uses samples from the input distribution as well as pairwise conditional samples (in the PCOND oracle model).

4.3 Multivariate Gaussians

Theorem 3.3 applies also when the sample space is not finite, e.g., the reals. Then, in the definition of the (β, γ)-EVAL approximator E_P for a distribution P, we require a distribution $\hat{P}$ such that $d_{\mathrm{TV}}(P, \hat{P}) \leqslant \beta$ and E_P is a $(1 \pm \gamma)$-approximation of the *probability density function* of $\hat{P}$ at any x.

The most prominent instance in which we can apply our framework in this setting is for the class of multivariate gaussians, again another widely used model for high-dimensional distributions used throughout the natural and social sciences (see, e.g., [MDLW18]). There are two main reasons for their ubiquity. Firstly, because of the central limit theorem, any physical quantity that is a population average is approximately distributed as a gaussian. Secondly, the gaussian distribution has maximum entropy among all real-valued distributions with a particular mean and covariance; therefore, a gaussian model places the least restrictions beyond the first and second moments of the distribution.

For $\mu \in \mathbb{R}^n$ and positive definite $\mathbb{\Sigma} \in \mathbb{R}^{n \times n}$, the distribution $N(\mu, \mathbb{\Sigma})$ has the density function:

$$N(\mu, \mathbb{\Sigma}; x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\mathbb{\Sigma})}} \exp\left(-\frac{1}{2}(x-\mu)^\top \mathbb{\Sigma}^{-1}(x-\mu)\right) \tag{3}$$

Invoking our framework on multivariate gaussians, we obtain:

Theorem 4.4. *Let $\mathcal{D}$ be the family of multivariate gaussian distributions, $\{N(\mu, \mathbb{\Sigma}) : \mu \in \mathbb{R}^n, \mathbb{\Sigma} \in \mathbb{R}^{n \times n}, \mathbb{\Sigma} \succ 0\}$. Then, there is a distance approximation algorithm for $\mathcal{D}$ with sample complexity $O(n^2\varepsilon^{-2})$ and runtime $O(n^\omega \varepsilon^{-2})$ (where $\omega > 2$ is the matrix multiplication constant).*

It is folklore (see [ABDH+17] for a proof) that for any $P = N(\mu, \mathbb{\Sigma})$, the empirical mean $\hat{\mu}$ and empirical covariance $\hat{\mathbb{\Sigma}}$ obtained from $O(n^2\varepsilon^{-2})$ samples from P determines a gaussian $\hat{P} = N(\hat{\mu}, \hat{\mathbb{\Sigma}})$ satisfying $d_{\mathrm{TV}}(P, \hat{P}) \leqslant \varepsilon$ with probability at least 3/4. To get an EVAL approximator, we need evaluations of $N(\hat{\mu}, \hat{\mathbb{\Sigma}}; x)$ for any x as in (3). Since $\det(\hat{\mathbb{\Sigma}})$ is computable in time $O(n^\omega)$, Theorem 4.4 follows from Theorem 3.3.

This result is interesting because there is no closed-form expression known for the total variation distance between two gaussians of specified mean and covariance. Devroye et al. [DMR18] give expressions for lower- and upper-bounding the total variation distance that are a constant multiplicative factor away from each other. On the other hand, our approach yields a polynomial time randomized algorithm that, given $\mu_1, \mathbb{\Sigma}_1, \mu_2, \mathbb{\Sigma}_2$, approximates the total variation distance $d_{\mathrm{TV}}(N(\mu_1, \mathbb{\Sigma}_1), N(\mu_2, \mathbb{\Sigma}_2))$ upto $\pm\varepsilon$ additive error.

Corollary 4.5. *For any two vectors $\mu_1, \mu_2 \in \mathbb{R}^n$ and two positive-definite matrices $\mathbb{\Sigma}_1, \mathbb{\Sigma}_2 \in \mathbb{R}^{n \times n}$, $d_{\mathrm{TV}}(N(\mu_1, \mathbb{\Sigma}_1), N(\mu_1, \mathbb{\Sigma}_1))$ can be estimated up to an additive ε error in $O(n^3\varepsilon^{-2})$ time.*

Proof. We again invoke Algorithm 2. Since the parameters are already provided, we can readily obtain $(0, 0)$-EVAL approximators for $N(\mu_1, \mathbb{\Sigma}_1)$ and $N(\mu_2, \mathbb{\Sigma}_2)$. For Algorithm 2, we also need sample access to one of the two distributions. It is well known that if $v \sim N(0, I)$ and $\mathbb{\Sigma} = LL^\top$, then $Lv + \mu \sim N(\mu, \mathbb{\Sigma})$; the matrix L can be obtained in $O(n^3)$ time using a Cholesky decomposition. Hence, each sample from $N(\mu_1, \mathbb{\Sigma}_1)$ costs $O(n^3)$ time, so that the entire algorithm runs in $O(n^3\varepsilon^{-2})$ time. □

4.4 Interventional Distributions in Causal Models

A *causal model* for a system of random variables describes not only how the variables are correlated but also how they would change if they were to be externally set to prescribed values. To formalize this, we can use the language of *causal Bayesian networks* due to Pearl [Pea09]. A causal Bayesian network is a Bayesian network with an extra *modularity* assumption: for each node i in the network, the dependence of X_i on $X_{\mathrm{parents}(i)}$ is an autonomous mechanism that does not change even if other parts of the network are changed.

Suppose $\mathcal{P}$ is a causal Bayesian network over variables $X_1, \ldots, X_n$ on a directed acyclic graph G with nodes labeled $\{1, \ldots, n\}$. The nodes in G are partitioned into two sets: *observable* V and *hidden* U. A sample from the observational distribution P yields the values of variables $X_V = \{X_i :\in V\}$. The modularity assumption allows us to define the result of *interventions* on causal Bayesian networks. An intervention is specified by a subset $S \subseteq V$ and an assignment $s \in \Sigma^{|S|}$. In the resulting interventional distribution, the variables in S are fixed to s, while the variables X_i for $i \notin S$ are sampled in topological order as it would have been in the original Bayesian network, according to the conditional probability distribution $X_i \mid X_{\mathrm{parents}(i)}$, where $X_{\mathrm{parents}(i)}$ consist of either variables previously sampled in the topological order or variables in S set by the intervention. Finally, the variables in U are marginalized out. The resulting distribution on X_V is denoted P_s.

The question of inferring the interventional distribution from samples is a fundamental one. We focus on *atomic interventions*, i.e., where the intervention is on a single node $A \in V$. In this case, Tian and Pearl [TP02a, Tia02] exactly characterized the graphs G for which any causal Bayesian network $\mathcal{P}$ on G and for any assignment $a \in \Sigma$ to X_A, the interventional distribution P_a is *identifiable* from the observational distribution P on X_V. For identification to be computationally effective, it is also natural to require certain *strong positivity* condition on P. We show that we can efficiently estimate the distances between interventional distributions of causal Bayesian networks whenever the identifiability and strong positivity conditions are met. See Appendix E for necessary definitions.

Theorem 4.6 (Informal)**.** *Suppose $\mathcal{P}, \mathcal{Q}$ are two unknown causal Bayesian networks on two known graphs G_1 and G_2 on a common observable set V containing a special node A and having bounded in-degree and c-component size. Suppose G_1 and G_2 both satisfy the identifiability condition, and the observational distributions P and Q satisfy the strong positivity condition. Then there is an algorithm which for any $a \in \Sigma$ and parameter $\varepsilon \in (0, 1)$ returns a value e such that $|e - d_{\mathrm{TV}}(P_a, Q_a)| \leqslant \varepsilon$ with probability at least 2/3 using $\mathrm{poly}(|\Sigma|, n, \varepsilon^{-1})$ samples from the observational distributions P and Q and running in time $\mathrm{poly}(|\Sigma|, n, \varepsilon^{-1})$.*

We again use the framework of EVAL approximators to prove the theorem, but there is a complication: we do not get samples from the distributions P_a and Q_a, but only from P and Q. We build on a recent work ([BGK+20]) that shows how to efficiently learn and sample from interventional distributions of atomic interventions using observational samples.

Theorem 4.6 solves a natural problem. Suppose a biologist wants to compare how a particular point mutation affects the activity of other genes for Africans and for Europeans. Because of ethical reasons, she cannot conduct randomized controlled trials by actively inducing the mutation, but she can draw random samples from the two populations. It is reasonable to assume that the graph structure of the regulatory network is the same for all individuals and that the causal graph over the genes of interest is known (or can be learned through other methods). Also, suppose that the gene expression levels can be discretized. She can then, in principle, use the algorithm proposed in Theorem 4.6 to test whether the effect of the mutation is approximately the same for Africans and Europeans.

4.5 Tightness of Our Bounds

In this paper our focus was mainly establishing upper bounds. We note that the $\Omega(\frac{n}{\log n})$ lower bound from [CDKS17] mentioned earlier for tolerant testing of product distributions, implies the same lower bound for tolerant testing of Bayes nets and atomic interventional distributions. For the Ising model, currently we do not have a lower bound in general.

Broader Impact

This work presents basic algorithms for approximating distances between two high dimensional distributions. While the results are theoretical in nature and do not present any immediate societal consequences, the algorithms have potential to impact practice in the long term.

Acknowledgement We thank the anonymous reviewers of NeurIPS20 for their valuable suggestions for improving our paper. This work was supported in part by National Research Foundation Singapore under its NRF Fellowship Programme [NRF-NRFFAI1-2019-0002, NRF-NRFFAI1-2019-0004] and AI Singapore Programme [AISG-RP-2018-005], NUS ODPRT Grant [R-252-000-685-13], and NUS ODPRT Grant [R-252-000-A33133]. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. The work of Bhattacharyya was additionally supported by an Amazon Research Award. The work of Vinodchandran was supported in part by the US National Science Foundation under grants NSF CCF-184908 and NSF HDR:TRIPODS-1934884. All opinions are of the authors and do not reflect the view of sponsors.

References

[ABDH+17] Hassan Ashtiani, Shai Ben-David, Nick Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Near-optimal sample complexity bounds for robust learning of gaussians mixtures via compression schemes, 2017.

[ABDK18] Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9469–9481, Red Hook, NY, USA, 2018. Curran Associates Inc.

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 3591–3599, Cambridge, MA, USA, 2015. MIT Press.

[BFR+13] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White. Testing closeness of discrete distributions. *Journal of the ACM (JACM)*, 60(1):4, 2013.

[BGK+20] Arnab Bhattacharyya, Sutanu Gayen, Saravanan Kandasamy, Ashwin Maran, and N. V. Vinodchandran. Efficiently learning and sampling interventional distributions from observations. *arXiv preprint*, 2020.

[BMV08] Andrej Bogdanov, Elchanan Mossel, and Salil P. Vadhan. The complexity of distinguishing markov random fields. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques, 11th International Workshop, APPROX 2008, and 12th International Workshop, RANDOM 2008, Boston, MA, USA, August 25-27, 2008. Proceedings*, volume 5171, pages 331–342. Springer, 2008.

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

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

[Cav78] James A Cavender. Taxonomy with confidence. *Mathematical biosciences*, 40(3-4):271–280, 1978.

[CDKS17] Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing bayesian networks. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 370–448, 2017.

[CDVV14] Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1193–1203. SIAM, 2014.

[CFGM16] Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the power of conditional samples in distribution testing. *SIAM J. Comput.*, 45(4):1261–1296, 2016.

[CM19] Sourav Chakraborty and Kuldeep S. Meel. On testing of uniform samplers. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 1 2019.

[CR14] Clément L. Canonne and Ronitt Rubinfeld. Testing probability distributions underlying aggregated data. In *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I*, pages 283–295, 2014.

[CRS14] Clément L. Canonne, Dana Ron, and Rocco A. Servedio. Testing equivalence between distributions using conditional samples. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pages 1174–1192, 2014.

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

[DMR18] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

[DMR+20] Luc Devroye, Abbas Mehrabian, Tommy Reddad, et al. The minimax learning rates of normal and ising undirected graphical models. *Electronic Journal of Statistics*, 14(1):2338–2361, 2020.

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

[Ell93] Glenn Ellison. Learning, local interaction, and coordination. *Econometrica: Journal of the Econometric Society*, pages 1047–1071, 1993.

[Far73] James S Farris. A probability model for inferring evolutionary trees. *Systematic Biology*, 22(3):250–256, 1973.

[Gác18] Pétér Gács, Feb 2018. From László Lovász's lecture notes. http://www.cs.bu.edu/faculty/gacs/courses/cs530/lectures/exact-Gauss.pdf.

[GG86] Stuart Geman and Christine Graffigne. Markov random field image models and their applications to computer vision. In *Proceedings of the international congress of mathematicians*, volume 1, page 2. Berkeley, CA, 1986.

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

[Isi25] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift f́ur Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

[JN07] Finn V. Jensen and Thomas D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer Publishing Company, Incorporated, 2nd edition, 2007.

[JS93] Mark Jerrum and Alistair Sinclair. Polynomial-time approximation algorithms for the ising model. *SIAM Journal on computing*, 22(5):1087–1116, 1993.

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

[Kie18] Stefan Kiefer. On computing the total variation distance of hidden markov models. In Ioannis Chatzigiannakis, Christos Kaklamanis, Dániel Marx, and Donald Sannella, editors, *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018, July 9-13, 2018, Prague, Czech Republic*, volume 107 of *LIPIcs*, pages 130:1–130:13, 2018.

[KM17] Adam Klivans and Raghu Meka. Learning graphical models using multiplicative weights. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 343–354. IEEE, 2017.

[KOPS15] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pages 1066–1100, Paris, France, 03–06 Jul 2015.

[LKFO18] Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. Pacgan: The power of two samples in generative adversarial networks. In *Advances in neural information processing systems*, pages 1498–1507, 2018.

[MDLW18] Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright. *Handbook of Graphical Models*. CRC Press, 2018.

[MS10] Andrea Montanari and Amin Saberi. The spread of innovations in social networks. *Proceedings of the National Academy of Sciences*, 107(47):20196–20201, 2010.

[Mur12] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[Nar21] Shyam Narayanan. Tolerant distribution testing in the conditional sampling model. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA*, 2021.

[Ney71] Jerzy Neyman. Molecular studies of evolution: a source of novel statistical problems. In Shanti S. Gupta and James Yackel, editors, *Statistical Decision Theory and Related Topics*, pages 1 – 27. Academic Press, 1971.

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[Pea09] Judea Pearl. *Causality*. Cambridge university press, 2009.

[PRR06] Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *J. Comput. Syst. Sci.*, 72(6):1012–1042, 2006.

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

[Sri19] Piyush Srivastava, Nov 2019. Personal communication.

[SSS16] Sanjit A Seshia, Dorsa Sadigh, and S Shankar Sastry. Towards verified artificial intelligence. *arXiv preprint arXiv:1606.08514*, 2016.

[SST14] Alistair Sinclair, Piyush Srivastava, and Marc Thurley. Approximation algorithms for two-state anti-ferromagnetic spin systems on bounded degree graphs. *Journal of Statistical Physics*, 155(4):666–686, 2014.

[SV03] Amit Sahai and Salil P. Vadhan. A complete problem for statistical zero knowledge. *J. ACM*, 50(2):196–249, 2003.

[Tia02] Jin Tian. *Studies in causal reasoning and learning*. University of California, Los Angeles, 2002.

[TP02a] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence*, pages 567–573, 2002.

[TP02b] Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, UAI'02, page 519–527, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[VP90] Thomas Verma and Judea Pearl. Causal networks: Semantics and expressiveness. In Ross D. Shachter, Tod S. Levitt, Laveen N. Kanal, and John F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, volume 9 of *Machine Intelligence and Pattern Recognition*, pages 69 – 76. North-Holland, 1990.

[VV10] Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:183, 2010.

[VV11a] Gregory Valiant and Paul Valiant. Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new clts. In Lance Fortnow and Salil P. Vadhan, editors, *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pages 685–694. ACM, 2011.

[VV11b] Gregory Valiant and Paul Valiant. The power of linear estimators. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 403–412. IEEE, 2011.

[VV17] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM J. Comput.*, 46(1):429–455, 2017.

[Wil65] J.H. Wilkinson. *The Algebraic Eigenvalue Problem*. Clarendon Press, 1965.

A Distance Approximation Algorithm

In this section, we prove Theorem 3.3 which underlies all the other results in this work. In fact, we show the following theorem that is more detailed.

Theorem A.1. *Suppose we have sample access to distributions P and Q over a finite set. Also, suppose we can make calls to two circuits $\mathcal{C}_P$ and $\mathcal{C}_Q$ which implement (β, γ)-EVAL approximators for P and Q respectively. Let T be the maximum running time for any call to $\mathcal{C}_P$ or $\mathcal{C}_Q$.*

Then for any $\varepsilon, \delta > 0$, $d_{\mathrm{TV}}(P, Q)$ can be approximated up to an additive error $\frac{2\gamma}{1-\gamma} + 3\beta + \varepsilon$ with probability at least $1 - \delta$, using $O(\varepsilon^{-2} \log \delta^{-1})$ samples from P and $O(\varepsilon^{-2} \log \delta^{-1} \cdot T)$ runtime.

Note that the EVAL approximators in Theorem A.1 must return rational numbers with bounded denominators as they are implemented by circuits with bounded running time. The exact model of computation for the circuits does not matter so much, so we omit its discussion.

We now turn to the proof of Theorem A.1. As mentioned in the Introduction, if $\mathcal{C}_P$ and $\mathcal{C}_Q$ were $(0, 0)$-EVAL approximators, the result already appears in [CR14]. The proof below analyzes how having nonzero β and γ affects the error bound.

Algorithm 2: Distance approximation

Input : Sample access to distribution P; oracle access to circuits $\mathcal{C}_P$ and $\mathcal{C}_Q$.
Output : Approximate value of $d_{\mathrm{TV}}(P, Q)$

for $i = 1, \ldots, t = O(\varepsilon^{-2} \log \delta^{-1})$ **do**
  Draw a sample x from P;
  $a \leftarrow \mathcal{C}_P(x)$;
  $b \leftarrow \mathcal{C}_Q(x)$;
  $c_i \leftarrow 1_{a>b} \left(1 - \frac{b}{a}\right)$;
return $\frac{1}{t} \sum_{i=1}^{t} c_i$

Proof. We invoke Algorithm 2. Notice that the algorithm only requires sample access to one of the two distributions but to both of the EVAL approximators. Let $\hat{P}$ be the distribution β-close to P which is approximated by the output of $\mathcal{C}_P$; similarly define $\hat{Q}$.

We have $|d_{\mathrm{TV}}(P, Q) - d_{\mathrm{TV}}(\hat{P}, \hat{Q})| \leqslant d_{\mathrm{TV}}(P, \hat{P}) + d_{\mathrm{TV}}(Q, \hat{Q}) \leqslant 2\beta$ from the triangle inequality. Hence, it is sufficient to approximate $d_{\mathrm{TV}}(\hat{P}, \hat{Q})$ additively up to $\frac{2\gamma}{1-\gamma} + \beta + \varepsilon$.

$$
\begin{aligned}
d_{\mathrm{TV}}(\hat{P}, \hat{Q}) &= \frac{1}{2} \sum_x |\hat{P}(x) - \hat{Q}(x)| \\
&= \sum_{x : \hat{P}(x) > \hat{Q}(x)} (\hat{P}(x) - \hat{Q}(x)) \\
&= \sum_{x : \hat{P}(x) > \hat{Q}(x)} \left(1 - \frac{\hat{Q}(x)}{\hat{P}(x)}\right) \hat{P}(x) && \text{(Since } \hat{P}(x) > 0\text{)} \\
&= \mathop{\mathbf{E}}_{x \sim \hat{P}} \left[1_{\hat{P}(x) > \hat{Q}(x)} \left(1 - \frac{\hat{Q}(x)}{\hat{P}(x)}\right) \right]
\end{aligned}
$$

From the above, if we have complete access (both evaluation and sample) to $\hat{P}$ and $\hat{Q}$, then we can estimate the distance with $O(\frac{1}{\varepsilon^2} \log \frac{1}{\delta})$ samples and evaluations. However as we have only approximate evaluations of $\hat{P}$ and $\hat{Q}$ and samples from the original distribution P, we need some additional arguments. Let E_P and E_Q be the functions implemented by the circuits $\mathcal{C}_P$ and $\mathcal{C}_Q$ respectively.

$$
\begin{aligned}
d_{\mathrm{TV}}(\hat{P},\hat{Q}) &= \sum_x 1_{\hat{P}(x)>\hat{Q}(x)}\left(1-\frac{\hat{Q}(x)}{\hat{P}(x)}\right)\hat{P}(x) \\
&= \underbrace{\sum_x 1_{E_P(x)>E_Q(x)}\left(1-\frac{E_Q(x)}{E_P(x)}\right)\hat{P}(x)}_{A} + \\
&\qquad \underbrace{\sum_x \left[1_{\hat{P}(x)>\hat{Q}(x)}\left(1-\frac{\hat{Q}(x)}{\hat{P}(x)}\right) - 1_{E_P(x)>E_Q(x)}\left(1-\frac{E_Q(x)}{E_P(x)}\right)\right]\hat{P}(x)}_{B}
\end{aligned}
$$

We start with an upper bound for the absolute value of the error term B. We consider the partition of sample space into S_1, S_2 and S_3, where $S_1 = \{x : 1_{\hat{P}(x)>\hat{Q}(x)} = 1_{E_P(x)>E_Q(x)}\}$, $S_2 = \{x : 1_{\hat{P}(x)>\hat{Q}(x)} > 1_{E_P(x)>E_Q(x)}\}$ and $S_3 = \{x : 1_{\hat{P}(x)>\hat{Q}(x)} < 1_{E_P(x)>E_Q(x)}\}$.

$$
\begin{aligned}
|B| &= \left|\sum_x \left[1_{\hat{P}(x)>\hat{Q}(x)}\left(1-\frac{\hat{Q}(x)}{\hat{P}(x)}\right) - 1_{E_P(x)>E_Q(x)}\left(1-\frac{E_Q(x)}{E_P(x)}\right)\right]\hat{P}(x)\right| \\
&\leqslant \sum_x \left|\left[1_{\hat{P}(x)>\hat{Q}(x)}\left(1-\frac{\hat{Q}(x)}{\hat{P}(x)}\right) - 1_{E_P(x)>E_Q(x)}\left(1-\frac{E_Q(x)}{E_P(x)}\right)\right]\hat{P}(x)\right| \\
&= \sum_{x\in S_1} 1_{\hat{P}(x)>\hat{Q}(x)}\left|\frac{\hat{Q}(x)}{\hat{P}(x)} - \frac{E_Q(x)}{E_P(x)}\right|\hat{P}(x) + \sum_{x\in S_2} 1_{\hat{P}(x)>\hat{Q}(x)}\left(1-\frac{\hat{Q}(x)}{\hat{P}(x)}\right)\hat{P}(x) + \\
&\qquad \sum_{x\in S_3} 1_{E_P(x)>E_Q(x)}\left(1-\frac{E_Q(x)}{E_P(x)}\right)\hat{P}(x)
\end{aligned}
$$

For x in S_1 with $\hat{P}(x) > \hat{Q}(x)$, $\frac{(1-\gamma)}{(1+\gamma)}\frac{\hat{Q}(x)}{\hat{P}(x)} \leqslant \frac{E_Q(x)}{E_P(x)} \leqslant \frac{(1+\gamma)}{(1-\gamma)}\frac{\hat{Q}(x)}{\hat{P}(x)}$ so that $\left|\frac{\hat{Q}(x)}{\hat{P}(x)} - \frac{E_Q(x)}{E_P(x)}\right| \leqslant \frac{2\gamma}{1-\gamma}\frac{\hat{Q}(x)}{\hat{P}(x)} < \frac{2\gamma}{1-\gamma}$. For x in S_2, $\hat{P}(x) > \hat{Q}(x)$ implies $E_P(x) \leqslant E_Q(x)$ and hence, $(1-\gamma)\hat{P}(x) \leqslant E_P(x) \leqslant E_Q(x) \leqslant (1+\gamma)\hat{Q}(x)$ so that $\hat{Q}(x)/\hat{P}(x) \geqslant \frac{1-\gamma}{1+\gamma}$. For x in S_3, $E_P(x) > E_Q(x)$ implies $\hat{P}(x) \leqslant \hat{Q}(x)$, and hence, $\frac{E_Q(x)}{E_P(x)} \geqslant \frac{(1-\gamma)\hat{Q}(x)}{(1+\gamma)\hat{P}(x)} \geqslant \frac{1-\gamma}{1+\gamma}$. Therefore:

$$
\begin{aligned}
|B| &\leqslant \sum_{x\in S_1}\frac{2\gamma}{1-\gamma}\hat{P}(x) + \sum_{x\in S_2}\frac{2\gamma}{1+\gamma}\hat{P}(x) + \sum_{x\in S_3}\frac{2\gamma}{1+\gamma}\hat{P}(x) \\
&\leqslant \frac{2\gamma}{1-\gamma}
\end{aligned}
$$

Now consider the term A:

$$
\begin{aligned}
A &= \sum_x 1_{E_P(x)>E_Q(x)}\left(1-\frac{E_Q(x)}{E_P(x)}\right)\hat{P}(x) \\
&= \underbrace{\sum_x 1_{E_P(x)>E_Q(x)}\left(1-\frac{E_Q(x)}{E_P(x)}\right)P(x)}_{C} + \sum_x 1_{E_P(x)>E_Q(x)}\left(1-\frac{E_Q(x)}{E_P(x)}\right)(\hat{P}(x)-P(x)).
\end{aligned}
$$

Note that: $\left|\sum_x 1_{E_P(x)>E_Q(x)}\left(1-\frac{E_Q(x)}{E_P(x)}\right)(\hat{P}(x)-P(x))\right| \leqslant \sum_x |\hat{P}(x)-P(x)| \leqslant \beta$. So, $|d_{\mathrm{TV}}(\hat{P},\hat{Q}) - C| \leqslant \frac{2\gamma}{1-\gamma} + \beta$. We can rewrite C as $\mathbf{E}_{x\sim P}\left[1_{E_P(x)>E_Q(x)}\left(1-\frac{E_Q(x)}{E_P(x)}\right)\right]$. Since $1_{E_P(x)>E_Q(x)}\left(1-\frac{E_Q(x)}{E_P(x)}\right)$ lies in $[0,1]$, by the Hoeffding bound, we can estimate the expectation up to ε additive error with probability at least $(1-\delta)$ by averaging $O(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$ samples from P. $\square$

Theorem A.1 can be extended to the case that P and Q are distributions over $\mathbb{R}^n$ with infinite support. We change Definition 3.1 so that $E_P(x)$ is a $(1 \pm \gamma)$-approximation of $\hat{f}(x)$ where $\hat{f}(x)$ is the probability density function for $\hat{P}$. Then, Theorem A.1 and Algorithm 2 continue to hold as stated. In the proof, we merely have to replace the summations with the appropriate integrals.

B Bayesian networks

First we apply our distance estimation algorithm for tolerant testing of high dimensional distributions coming from bounded in-degree Bayesian networks. Bayesian networks defined below are popular probabilistic graphical models for describing high-dimensional distributions succinctly.

Definition B.1. *A* Bayesian network P *on a directed acyclic graph* G *over the vertex set* $[n]$ *is a joint distribution of the* n *random variables* $(X_1, X_2, \ldots, X_n)$ *over the sample space* Σ^n *such that for every* $i \in [n]$ X_i *is conditionally independent of* $X_{\text{non-descendants}(i)}$ *given* $X_{\text{parents}(i)}$*, where for* $S \subseteq [n]$*,* X_S *is the joint distribution of* $(X_i : i \in S)$*, and parents and non-descendants are defined from* G*.*

P *factorizes as follows:*

$$P(x) \coloneqq \Pr_{X \sim P}[X = x] = \prod_{i=1}^{n} \Pr_{X \sim P}[X_i = x_i \mid \forall j \in \text{parents}(i), X_j = x_j] \qquad \text{for all } x \in \Sigma^n \quad (4)$$

Hence a Bayesian network can be completely described by a set of conditional distributions for every variable X_i*, for every fixing of its parents* $X_{\text{parents}(i)}$*.*

To construct an EVAL approximator for a Bayesian network, we first learn it using an efficient algorithm. We show the following proper learning algorithm for Bayesian networks that uses near-optimal sample complexity [CDKS17].

Theorem B.2. *There is an algorithm that given a parameter* $\varepsilon > 0$ *and sample access to an unknown Bayesian network distribution* P *on a known directed acyclic graph* G *of in-degree at most* d*, returns a Bayesian network* $\hat{P}$ *on* G *such that* $d_{\mathrm{TV}}(P, \hat{P}) \leqslant \varepsilon$ *with probability* $\geqslant 9/10$*. Letting* Σ *denote the range of each variable* X_i*, the algorithm takes* $m = O(|\Sigma|^{d+1} n \log(|\Sigma|^{d+1} n)\varepsilon^{-2})$ *samples and runs in* $O(mn)$ *time.*

This directly gives us a distance estimation algorithm for Bayesian networks.

Theorem 4.1. *Suppose* G_1 *and* G_2 *are two DAGs on* n *vertices with in-degree at most* d*. Let* $\mathcal{D}_1$ *and* $\mathcal{D}_2$ *be the family of Bayesian networks on* G_1 *and* G_2 *respectively. Then, there is a distance approximation algorithm for* $(\mathcal{D}_1, \mathcal{D}_2)$ *that gets* $m = \tilde{O}(|\Sigma|^{d+1} n \varepsilon^{-2})$ *samples and runs in* $O(mn)$ *time.*

Proof. Given samples from P_1 and P_2 we first learn them as $\hat{P}_1$ and $\hat{P}_2$ using Theorem B.2 in d_{TV} distance $\varepsilon/4$. This step costs $m = O(|\Sigma|^{d+1} n \log(|\Sigma|^{d+1} n)\varepsilon^{-2})$ samples and $O(|\Sigma|^{d+1} mn)$ time and succeeds with probability 4/5. $\hat{P}_1$ and $\hat{P}_2$ gives efficient $(\varepsilon/4, 0)$-EVAL approximators from Equation (4). It follows from Theorem A.1 that we can estimate $d_{\mathrm{TV}}(P_1, P_2)$ up to an ε additive error using $O(\varepsilon^{-2})$ additional samples from P_1 except for 1/5 probability. □

Regarding lower bounds, Canonne et al. [CDKS17] have shown a lower bound of $\Omega(n/\log n)$ samples for deciding for two product distributions P and Q over $\{0,1\}^n$, whether $d_{\mathrm{TV}}(P, Q) \leqslant \varepsilon_0$ versus $d_{\mathrm{TV}}(P, Q) \geqslant 2\varepsilon_0$ with probability 2/3 for a constant ε_0. On the other hand, Daskalakis et al. [DDK19] have shown that there exists an unknown Bayes net P over $\{0,1\}^n$ whose underlying graph is unknown but known to be a tree such that deciding $d_{\mathrm{TV}}(P, U) = 0$ versus $d_{\mathrm{TV}}(P, U) \geqslant \varepsilon$ with 2/3 probability requires $\Omega(n\varepsilon^{-2})$ samples, where U is the uniform distribution over $\{0,1\}^n$.

B.1 Learning Bayesian networks

In this section, we prove a strengthened version of Theorem B.2 that holds for any desired error probability δ.

Theorem B.3. *There is an algorithm that given parameters $\varepsilon, \delta > 0$ and sample access to an unknown Bayesian network distribution P on a known directed acyclic graph G of in-degree at most d, returns a Bayesian network Q on G such that $d_{\mathrm{TV}}(P, Q) \leqslant \varepsilon$ with probability $\geqslant (1 - \delta)$. Letting Σ denote the alphabet for each variable X_i, the algorithm takes $m = O(|\Sigma|^{d+1} n \log(|\Sigma|^{d+1} n)\varepsilon^{-2} \log \frac{1}{\delta})$ samples and runs in $O(mn \log^2 \frac{1}{\delta})$ time.*

We actually prove a stronger bound on the distance between P and Q in terms of the KL divergence. The KL divergence between two distributions P and Q is defined as $\mathrm{KL}(P, Q) = \sum_i P(i) \ln \frac{P(i)}{Q(i)}$. From Pinsker's inequality, we have ${d_{\mathrm{TV}}}^2(P, Q) \leqslant 2\mathrm{KL}(P, Q)$. Thus a d_{TV} learning result follows from a KL learning result. We present Algorithm 3 for the binary alphabet case ($\Sigma = \{0, 1\}$) and reduce the general case to the binary case afterwards.

The *add-1 empirical estimator* takes z samples from a distribution over k items and assigns to item i the probability $(z_i + 1)/(z + k)$ where z_i is the number of occurrences of item i in the samples. We will use the following general result for learning a distribution in KL distance.

Theorem B.4 ([KOPS15])**.** *Let D be an unknown distribution over k items. Let $\hat{D}$ be the add-1 empirical distribution of z samples from D. Then for $k \geqslant 2, z \geqslant 1$, $\mathbf{E}[\mathrm{KL}(D, \hat{D})] \leqslant (k-1)/(z+1)$.*

We will use a KL local additivity result for Bayesian networks, a proof of which is given in [CDKS17]. For a Bayesian network P, a vertex i, and a setting a value a of its parents, let $\Pi[i, a]$ denote the event that parents of i take value a, and let $P(i \mid a)$ denote the distribution at vertex i when its parents takes value a.

Theorem B.5. *Let P and Q be two Bayesian networks over the same graph G. Then*

$$\mathrm{KL}(P, Q) = \sum_i \sum_a P[\Pi[i, a]] \cdot \mathrm{KL}(P(i \mid a), Q(i \mid a))$$

Algorithm 3: Fixed-structure Bayesian network learning

Input : Samples from an unknown Bayesian network P over $\{0, 1\}^n$ on a known graph G of in-degree $\leqslant d$, parameters m, t
Output : A Bayesian network Q over G

Get m samples from P;
for *every vertex i* **do**
  for *every fixing a of i's parents* **do**
    $N_{i,a} \leftarrow$ the number of samples where i's parents are set to a;
    if $N_{i,a} \geqslant t$ **then**
      $Q(i \mid a) \leftarrow$ the add-1 empirical distribution at node i in the subset of samples where i's parents are set to a;
    else
      $Q(i \mid a) \leftarrow$ uniformly random bit;

Lemma B.6. *For $m = 24n2^d \log(n2^d)/\varepsilon$ and $t = 12 \log(n2^d)$, Algorithm 3 satisfies $\mathrm{KL}(P, Q) \leqslant 5\varepsilon$ with probability at least 3/4 over the randomness of sampling.*

Proof. Call a tuple (i, a) *heavy* if $P[\Pi[i, a]] \geqslant \frac{\varepsilon}{2^d n}$ and *light* otherwise. Let $N_{i,a}$ denote the number of samples where i's parents are a.

For every heavy (i, a), let $F_{i,a}$ be the event "$N_{i,a} \geqslant n2^d P[\Pi[i, a]]t/\varepsilon$" and $G_{i,a} = \bigwedge_{\substack{(j,b) \text{ heavy} \\ (j,b) \neq (i,a)}} F_{(j,b)}$. Let $F = G_{i,a} \wedge F_{i,a}$. It is easy to see from Chernoff and union bounds that F is true with 19/20 probability. Hence for the rest of the argument, we condition on this event. In this case, all heavy items satisfy $N_{i,a} \geqslant t$.

Then for any random variable X, $\mathbf{E}[X \mid F_{i,a}] = \mathbf{E}[X \mid F]\Pr[G_{i,a} \mid F_{i,a}] + \mathbf{E}[X \mid F_{i,a} \wedge \overline{G_{i,a}}]\Pr[\overline{G_{i,a}} \mid F_{i,a}]$. Hence, $\mathbf{E}[X \mid F] \leqslant \frac{20}{19} \mathbf{E}[X \mid F_{i,a}]$. Similarly, $\mathbf{E}[X \mid F] \leqslant \frac{20}{19} \mathbf{E}[X]$.

Now, we see that:

– For any heavy (i,a), by Theorem B.4,

$$\mathbf{E}[\mathrm{KL}(P(i \mid a), Q(i \mid a)) \mid F_{i,a}] \leqslant \frac{\varepsilon}{12n2^d \cdot P[\Pi[i,a]]}.$$

– Similarly, for any light (i,a) that satisfies $N_{i,a} \geqslant t$, it follows from Theorem B.4 that $\mathbf{E}[\mathrm{KL}(P(i \mid a), Q(i \mid a)) \mid N_{i,a} \geqslant t] \leqslant \frac{1}{12}$.

– Items which do not satisfy $N_{i,a} \geqslant t$ must be light for which $[\mathrm{KL}(P(i \mid a), Q(i \mid a)) \mid N_{i,a} < t] \leqslant p \ln 2p + (1-p)\ln 2(1-p) \leqslant \ln 2$, where $p = P[i=1|a]$, since in that case $Q(i \mid a)$ is the uniformly random bit.

Using Theorem B.5, we get

$$\mathbf{E}[\mathrm{KL}(P,Q) \mid F] \leqslant \frac{20}{19}\left[\sum_{(i,a)\text{ heavy}} P[\Pi[i,a]] \cdot \frac{\varepsilon}{12n2^d \cdot P[\Pi[i,a]]} + \sum_{(i,a)\text{ light}} \frac{\varepsilon}{n2^d} \ln 2\right] \leqslant \varepsilon.$$

The lemma follows from Markov's inequality. □

Now we reduce the case when Σ is not binary to the binary case. We can encode each $\sigma \in \Sigma$ of the Bayesian network as a $\log|\Sigma|$ size boolean string which gives us a Bayesian network of degree $(d+1)\log|\Sigma|$ over $n\log|\Sigma|$ variables. Then we apply Lemma B.6 to get a learning algorithm with $O(\varepsilon)$ error in d_{TV} and 3/4 success probability. Subsequently we repeat $O(\log\frac{1}{\delta})$ times and find out a successful repetition using Theorem F.1.

C Ising Models

In this section, we give a distance approximation algorithm for the class of bounded-width ferromagnetic Ising models. Recall from Section 4.2 that a probability distribution P from this class is over the sample space $\{-1,1\}^n$ and that $P(x)$, the probability of an item $x \in \{-1,1\}^n$, is proportional to the numerator:

$$N(x) = \exp\left(\sum_{i,j} A_{i,j}x_ix_j + \theta\sum_i x_i\right),$$

where $A_{i,j}$s and θ are parameters of the model. The constant of proportionality, also called the *partition function* of the Ising model is $Z = \sum_x N(x)$, which gives $P(x) = N(x)/Z$. The *width* of the Ising model is defined as $\max_i \sum_j |A_{i,j}| + \theta$. In a *ferromagnetic* Ising model, each $A_{i,j} \geqslant 0$.

Given two such Ising models, we give an algorithm for additively estimating their total variation distance. We first learn these two Ising models up to total variation distance $\varepsilon/8$ using the following learning algorithm given by Klivans and Meka [KM17]. In fact, it gives a stronger $(1\pm\varepsilon)$ multiplicative approximation guarantee for every probability value.

Theorem C.1 (Theorem 7.3 in [KM17])**.** *There is an algorithm which, given independent samples from an unknown Ising model P with width at most d, returns parameters $\hat{A}_{i,j}$ and $\hat{\theta}$ such that the Ising model $\hat{P}$ constructed with the latter parameters satisfies $(1-\varepsilon)P(x) \leqslant \hat{P}(x) \leqslant (1+\varepsilon)P(x)$ for all $x \in \{-1,1\}^n$. This algorithm takes $m = e^{O(d)}\varepsilon^{-4}n^8\log(n/\delta\varepsilon)$ samples, $O(mn^2)$ time and succeeds with probability $1-\delta$.*

However learning the parameters of an Ising model is not enough to efficiently evaluate the probability at arbitrary points. Naively computing the constant of proportionality Z would take 2^n time. For certain classes of Ising models polynomial time algorithms are known which approximates Z up to a $(1\pm\varepsilon)$ approximation factor. In particular we use the following approximation algorithm for ferromagnetic[¶] Ising models due to Jerrum and Sinclair [JS93].

[¶]As pointed out by [Sri19], Jerrum and Sinclair's result (and hence, our result) extends to the *non-uniform external field* setting where there is a θ_i for each i instead of $\theta_1 = \cdots = \theta_n = \theta$, with the restriction that each $\theta_i \geqslant 0$.

Theorem C.2. *There is an algorithm which given the parameters of a ferromagnetic Ising model distribution P, in $O(\varepsilon^{-2}n^{17}\log n)$ time returns a number $\hat{Z}$ such that with probability at least 9/10, $(1-\varepsilon)Z \leqslant \hat{Z} \leqslant (1+\varepsilon)Z$, where Z is the partition function of P.*

Combining the previous two results with our general distance estimation algorithm, we can now obtain our main result for Ising models which we restate below.

Theorem 4.2. *Let $\mathcal{D}$ be the family of ferromagnetic Ising models having width at most d. Then, there is a distance approximation algorithm for $\mathcal{D}$ with sample complexity $m = e^{O(d)}\varepsilon^{-4}n^8\log(\frac{n}{\varepsilon})$ and runtime $O(mn^2 + \varepsilon^{-2}n^{17}\log n)$.*

Proof. We first use Theorem C.1 to get the parameters for a pair of Ising models $\hat{P}$ and $\hat{Q}$ which are, with probability at least $9/10$, pointwise $(1 \pm \varepsilon/8)$ approximations to P and Q. If $\hat{P}$ or $\hat{Q}$ has any negative pairwise interaction term, then we modify them to zero, thus making $\hat{P}$ and $\hat{Q}$ ferromagnetic. We claim that since P and Q are ferromagnetic to start with, this can only improve the approximation factor. The reason is that Klivans and Meka, in their proof of Theorem C.1, show the more general result that for any *log-polynomial distribution*, i.e, any distribution P on $\{-1,1\}^n$ where $P(x) \propto \exp(T(x))$ for a bounded-degree polynomial T, they can obtain a polynomial $\hat{T}$ with the same degree that satisfies a bound on $\|T - \hat{T}\|_1 = \sum_\alpha |T[\alpha] - \hat{T}[\alpha]|$ where $T[\alpha]$ and $\hat{T}[\alpha]$ are the coefficients of the monomial indexed by α. It is clear that if $T[\alpha] \geqslant 0$, changing $\hat{T}[\alpha]$ to $\max(0, \hat{T}[\alpha])$ can only reduce $\|T - \hat{T}\|_1$.

Abusing notation for simplicity, henceforth let $\hat{P}$ and $\hat{Q}$ be the distributions after this modification. Let $N_{\hat{P}}(x)$ and $N_{\hat{Q}}(x)$ be the numerators for $\hat{P}$ and $\hat{Q}$ respectively. Then we apply Theorem C.2 to estimate, with probability $4/5$, the partition functions $\hat{Z}_P$ and $\hat{Z}_Q$ of $\hat{P}$ and $\hat{Q}$ respectively up to a $(1 \pm \varepsilon/8)$ multiplicative factor. Therefore, $E_P(x) = N_{\hat{P}}(x)/\hat{Z}_P$ and $E_Q(x) = N_{\hat{Q}}(x)/\hat{Z}_Q$ are $(\varepsilon/8, \varepsilon/4)$-EVAL approximators for P and Q respectively, where the $\varepsilon/8$-close distributions are $\hat{P}$ and $\hat{Q}$. It follows from Theorem A.1 that conditioned on the above, we can estimate $d_{\mathrm{TV}}(P,Q)$ up to an ε additive error with probability at least $9/10$. □

Remark C.3. Klivans and Meka [KM17] have also given an algorithm for recovering the underlying dependency graph of an n-dimensional ising model using $O(\exp(O(d)/\eta^4)\log(\frac{n}{\eta\rho}))$ samples assuming its width at most d and $\min_{i,j:A_{i,j}\neq 0}|A_{i,j}| \geqslant \eta$. Devroye et al. [DMR+20] have given a minimax-optimal algorithm that given the underlying graph, learns an Ising model in $d_{\mathrm{TV}} \leqslant \varepsilon$ using $O(n^2/\varepsilon^2)$ samples with 9/10 probability. These two results can be daisy-chained to improve the sample complexity of learning an unknown ising model and hence of our distance approximation algorithm. However, as noted in Section 6 of the later paper, this algorithm is not polynomial time and hence we will not get a polynomial time algorithm for distance approximation.

C.1 Distance to uniformity

Next we give an algorithm for estimating the distance between an unknown Ising model and the uniform distribution over $\{-1,1\}^n$. We use the following recent result by Narayanan [Nar21].

Theorem C.4 (Restated from [Nar21])**.** *Let U and D be the uniform distribution and any other distribution over $[N]$ respectively, such that we can sample from D, as well as compute the ratio $D(i)/D(j)$ for any $i \neq j \in [N]$ up to $(1 \pm \varepsilon)$ error for any $0 < \varepsilon < 1$ in unit time. Then $d_{\mathrm{TV}}(D,U)$ can be approximated up to an additive error using $\widetilde{O}(\varepsilon^{-2})$ samples with 2/3 probability.*

Theorem 4.3. *There is an algorithm which, given independent samples from an unknown Ising model P over $\{-1,1\}^n$ with width at most d, takes $m = O(e^{O(d)}\varepsilon^{-4}n^8\log(n/\varepsilon))$ samples, $O(mn^2)$ time and returns a value e such that $|e - d_{\mathrm{TV}}(P,U)| \leqslant \varepsilon$ with probability at least 7/12, where U is the uniform distribution over $\{-1,1\}^n$.*

Proof. We first learn the parameters of the unknown ising model from samples using Theorem C.1. As we noted earlier computing the partition function naively is intractable in general. However computing N_x/N_z, the ratio of the probabilities of two items x, y can be computed in $O(n^2)$ time up to $(1 \pm \varepsilon)$ approximation from Theorem C.1. The result follows from Theorem C.4. □

D Multivariate Gaussians

In this section we give an algorithm for additively estimating the total variation distance between two unknown multidimensional Gaussian distributions. For a mean vector $\mu \in \mathbb{R}^n$ and a positive definite covariance matrix $\mathbb{\Sigma} \in \mathbb{R}^{n \times n}$, the Gaussian distribution $N(\mu, \mathbb{\Sigma})$ has the pdf:

$$N(\mu, \mathbb{\Sigma}; x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\mathbb{\Sigma})}} \exp\left(-\frac{1}{2}(x-\mu)^\top \mathbb{\Sigma}^{-1}(x-\mu)\right) \tag{5}$$

We use the following folklore result (see [ABDH+17] for a proof) for learning the two Gaussians.

Theorem D.1. *Let P be an n-dimensional Gaussian distribution. Let $\hat{\mu} \in \mathbb{R}^n$ and $\hat{\mathbb{\Sigma}} \in \mathbb{R}^{n \times n}$ be the empirical mean and the empirical covariance defined by $O(n^2\varepsilon^{-2})$ samples from P. Then, with probability at least $9/10$, the distribution $\hat{P} = N(\hat{\mu}, \hat{\mathbb{\Sigma}})$ satisfies $d_{\mathrm{TV}}(P, \hat{P}) \leqslant \varepsilon$.*

We are now ready to prove Theorem 4.4 restated below.

Theorem 4.4. *Let $\mathcal{D}$ be the family of multivariate gaussian distributions, $\{N(\mu, \mathbb{\Sigma}) : \mu \in \mathbb{R}^n, \mathbb{\Sigma} \in \mathbb{R}^{n \times n}, \mathbb{\Sigma} \succ 0\}$. Then, there is a distance approximation algorithm for $\mathcal{D}$ with sample complexity $O(n^2\varepsilon^{-2})$ and runtime $O(n^\omega \varepsilon^{-2})$ (where $\omega > 2$ is the matrix multiplication constant).*

Proof. We first apply Theorem D.1 to obtain $\hat{P}$ and $\hat{Q}$ such that each is within $\varepsilon/4$ distance from P and Q respectively. Since we can evaluate the pdf of $\hat{P}$ and $\hat{Q}$ exactly, they serve as $(\varepsilon/4, 0)$ EVAL -approximators for P and Q. Each determinant computation costs $O(n^\omega)$ time. Subsequently from (the continuous analog of) Theorem A.1, using $O(\varepsilon^{-2})$ samples from P and $O(n^\omega\varepsilon^{-2})$ time, we can estimate $d_{\mathrm{TV}}(P, Q)$ up to an additive ε error with probability at least $4/5$. □

Remark D.2. The above time analysis uses the unrealistic real RAM model in which real number computations can be carried out exactly upto infinite precision. However, there are strongly polynomial time algorithms for computing matrix determinant and inverse [Gác18, Wil65], so that even in the more realistic word RAM model, the above algorithm runs in polynomial time.

E Causal Bayesian Networks under Atomic Interventions

We describe Pearl's notion of causality from [Pea09]. Central to his formalism is the notion of an *intervention*. Given a variable set V and a subset $S \subset V$, an intervention $\mathsf{do}(s)$ is the process of fixing the set of variables in S to the values s. If the original distribution on V is P, we denote the *interventional distribution* as P_s, intuitively, the distribution induced on V when an external force sets the variables in S to s.

Another important component of Pearl's formalism is that some variables may be hidden (latent). The hidden variables can neither be observed nor be intervened upon. Let V and U denote the subsets corresponding to observable and hidden variables respectively. Given a directed acyclic graph H on $V \cup U$ and a subset $S \subseteq (V \cup U)$, we use $\Pi_H(S)$ and $\mathsf{Pa}_H(S)$ to denote the set of all parents and observable parents respectively of S, excluding S, in H. When the graph H is clear, we may omit the subscript.

Definition E.1 (Causal Bayesian Network)**.** *A* (semi-Markovian) causal Bayesian network (CBN) *on variables $X_1, \ldots, X_n$ is a collection of interventional distributions defined by a tuple $\langle V, U, G, \{\mathbf{Pr}[X_i \mid x_{\Pi(i)}] : i \in V, x_{\Pi(i)} \in \Sigma^{|\Pi(i)|}\}, \mathbf{Pr}[X_U]\}\rangle$, where (i) G is a directed acyclic graph on $V \cup U = [n]$, (ii) $\mathbf{Pr}[X_i \mid x_{\Pi(i)}]$ is the conditional probability distribution of X_i given that its parents $X_{\Pi(i)}$ take the values $x_{\Pi(i)}$, and (iii) $\mathbf{Pr}[X_U]$ is the distribution of the hidden variables $\{X_i : i \in U\}$.*

A CBN $\mathcal{P} = \langle V, U, G, \{\mathbf{Pr}[X_i \mid x_{\Pi(i)}] : i \in V, x_{\Pi(i)} \in \Sigma^{|\Pi(i)|}\}, \mathbf{Pr}[X_U]\rangle$ *defines a unique interventional distribution P_s for every subset $S \subseteq V$ (including $S = \emptyset$) and assignment $s \in \Sigma^{|S|}$, as follows. For all $x \in \Sigma^{|V|}$:*

$$P_s(x) = \begin{cases} \sum_u \prod_{i \in V \setminus S} \mathbf{Pr}[x_i \mid x_{\pi(i)}] \cdot \mathbf{Pr}[X_U = u] & \text{if } x \text{ is consistent with } s \\ 0 & \text{otherwise.} \end{cases}$$

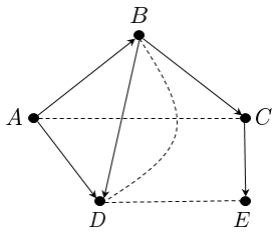

Figure 1: An acyclic directed mixed graph (ADMG) where the bidirected edges are depicted as dashed. The in-degree of the graph is 2. The c-components are $\{A, C\}$ and $\{B, D, E\}$.

We use P to denote the observational distribution ($S = \emptyset$). G is said to be the causal graph *corresponding to the* CBN $\mathcal{P}$.

It is standard in the causality literature [TP02b, VP90, ABDK18] to assume that each variable in U is a source node with exactly two children from V, since there is a known algorithm [TP02b, VP90] which converts a general causal graph into such graphs. Given such a causal graph, we remove every source node Z from G and put a *bidirected* edge between its two observable children X_1 and X_2. We end up with an Acyclic Directed Mixed Graph (ADMG) graph G, having vertex set V and having edge set $E^{\rightarrow} \cup E^{\leftrightarrow}$ where $E^{\rightarrow}$ are the directed edges and $E^{\leftrightarrow}$ are the bidirected edges. The *in-degree* of G is the maximum number of directed edges coming into any vertex in V. A *c-component* refers to any maximal subset of V which is interconnected by bidirected edges. Then V gets partitioned into c-components: $S_1, S_2, \ldots, S_\ell$. Figure 1 shows an example.

Throughout this section, we focus on *atomic* interventions, i.e. interventions on a single variable. Let $A \in V$ correspond to this variable. Without loss of generality, suppose $A \in S_1$. Tian and Pearl [TP02a] showed that in an ADMG G as above, P_a can be completely determined from P for all $a \in \Sigma$ iff the following condition holds.

Assumption E.2 (Identifiability wrt A)**.** *There does not exist a path of bidirected edges between A and any child of A. Equivalently, no child of A belongs to S_1.*

Recently algorithms and sample complexity bounds for learning and sampling from identifiable atomic interventional distributions were given in [BGK+20] under the following additional assumption. For $S \subseteq V$, let $\mathsf{Pa}^+(S) = S \cup \mathsf{Pa}(S)$.

Assumption E.3 (α-strong positivity wrt A)**.** *Suppose A lies in the c-component S_1, and let $Z = \mathsf{Pa}^+(S_1)$. For every assignment z to Z, $P(Z = z) > \alpha$.*

We state the two main results of [BGK+20], which given sampling access to the observational distribution P of an unknown causal Bayesian network on a known ADMG return an $(\varepsilon, 0)$-EVAL approximator and an approximate generator for P_a. For the two results below, suppose the CBN $\mathcal{P}$ satisfies identifiablity (Assumption E.2) and α-strong positivity (Assumption E.3) with respect to a variable $A \in V$. Let d denote the maximum in-degree of the graph G and k denote the size of its largest c-component.

Theorem E.4 (EVAL approximator and Sampler [BGK+20])**.** *For any intervention a to A and parameter $\varepsilon \in (0, 1)$, there is an algorithm that takes $m = \tilde{O}\left(\frac{|\Sigma|^{2kd} n}{\alpha^k \varepsilon^2} \log \frac{1}{\delta}\right)$ samples from P, and in $O(mn \log^2 \frac{1}{\delta})$ time, returns a distribution $\hat{P}_a$ such that $d_{\mathrm{TV}}(P_a, \hat{P}_a) \leqslant \varepsilon$ with probability at least $(1 - \delta)$ and returns a circuit $E_{P,a}$ such that:*

- *Evaluation: Given an assignment x to the nodes, $E_{P,a}$ outputs $\hat{P}_a(x)$ exactly in $O(n)$ time.*
- *Generation: Obtaiing an independent sample from $\hat{P}_a$ takes $O(n)$ time.*

We give a distance approximation algorithm for identifiable atomic interventional distributions using the above result and Theorem A.1.

Theorem E.5 (Formal version of Theorem 4.6)**.** *Suppose $\mathcal{P}, \mathcal{Q}$ are two unknown CBN's on two known ADMGs G_1 and G_2 on a common observable set V both satisfying Assumption E.2 and*

Assumption E.3 wrt a special vertex A. Let d denote the maximum in-degree, and k denote the size of the largest c-component of G_1 and G_2.

Then there is an algorithm which for any $a \in \Sigma$ and parameter $\varepsilon \in (0,1)$, takes $m = \tilde{O}\left(\frac{|\Sigma|^{2kd}n}{\alpha^k\varepsilon^2}\log\frac{1}{\delta}\right)$ samples from P and Q, runs in time $\tilde{O}(mn\log^2\frac{1}{\delta})$ and returns a value e such that $|e - d_{\mathrm{TV}}(P_a, Q_a)| \leqslant \varepsilon$ with probability at least 2/3.

Proof. We first invoke Theorem E.4 to learn the two distributions as $\hat{P}_a$ and $\hat{Q}_a$with the distance parameter ε, which serve as $(\varepsilon, 0)$-EVAL approximators for P_a and Q_a respectively. Once learnt, no further samples are needed from P_a and Q_a. $\hat{P}_a$ can be sampled in $O(n)$ time from Theorem E.4. The result follows from Theorem A.1. □

F Improving Success of Learning Algorithms Using Distance Estimation

In this section we give a general algorithm for improving the success probability of learning certain families of distributions. Specifically, let $\mathcal{D}$ be a family of distributions for which we have a learning algorithm $\mathcal{A}$ in d_{TV} distance ε that succeeds with probability 3/4. Suppose there is also a distance approximation algorithm $\mathcal{B}$ for $\mathcal{D}$. The algorithm presented below, which uses $\mathcal{A}$ and $\mathcal{B}$, learns an unknown distribution from $\mathcal{D}$ with probability at least $(1-\delta)$.

Algorithm 4: High probability distribution learning

Data: Samples from an unknown distribution P

Result: A distribution $\hat{P}$ such that $d_{\mathrm{TV}}(P, \hat{P}) \leqslant \varepsilon$ with probability $1-\delta$

for $0 \leqslant i \leqslant R = O(\log\frac{1}{\delta})$ **do**
  $P_i \leftarrow$ Run $\mathcal{A}$ on samples from P to get a learnt distribution;
  $count_i \leftarrow 0$;
for *every unordered pair* $0 \leqslant i < j \leqslant R$ **do**
  $d_{ij} \leftarrow$ Estimate distance between P_i and P_j up to additive error ε using $\mathcal{B}$;
  if $d_{ij} \leqslant 3\varepsilon$ **then**
    $count_i \leftarrow count_i + 1$;
    $count_j \leftarrow count_j + 1$;
$i^* = \arg\max_i count_i$;
return P_{i^*};

Theorem F.1. *Let $\mathcal{D}$ be a family of distributions. Suppose there is a learning algorithm $\mathcal{A}$ which for any $P \in \mathcal{D}$ takes $m_{\mathcal{A}}(\varepsilon)$ samples from P and in time $t_{\mathcal{A}}(\varepsilon)$ outputs a distribution P_1 such that $d_{\mathrm{TV}}(P, P_1) \leqslant \varepsilon$ with probability at least 3/4. Suppose there is a distance approximation algorithm $\mathcal{B}$ for $\mathcal{D}$ that given any two completely specified distributions P_1 and P_2 estimates $d_{\mathrm{TV}}(P_1, P_2)$ up to an additive error ε in $t_{\mathcal{B}}(\varepsilon, \delta)$ time with probability at least $(1-\delta)$. Then there is an algorithm that uses $\mathcal{A}$ and $\mathcal{B}$ as subroutines, takes $O(m_{\mathcal{A}}(\varepsilon/4)\log\frac{1}{\delta})$ samples from P, runs in $O(t_{\mathcal{A}}(\varepsilon/4)\log\frac{1}{\delta} + t_{\mathcal{B}}(\varepsilon/4, \frac{\delta}{210000\log^2\frac{2}{\delta}})\log^2\frac{1}{\delta})$ time and returns a distribution $\hat{P}$ such that $d_{\mathrm{TV}}(P, \hat{P}) \leqslant \varepsilon$ with probability at least $1-\delta$.*

Proof. The boosting algorithm is given in Algorithm 4. We take $R = 324\log\frac{2}{\delta}$ repetitions of $\mathcal{A}$ to get the distributions P_is. From Chernoff's bound at least $2R/3$ distributions (successful) satisfy $d_{\mathrm{TV}}(P_i, P) \leqslant \varepsilon$ with probability at least $1-\delta/2$, which we condition on henceforth. These successful distributions have pairwise distance at most 2ε. Conditioned on the $\binom{R}{2}$ calls to $\mathcal{B}$ succeeding, the pairwise distances between the successful distributions are at most 3ε. Hence every successful i has its count value at least $2R/3 - 1$. This means i^*, which has the maximum count value $(\geqslant 2R/3 - 1)$ must intersect at least one successful i' such that $d_{\mathrm{TV}}(P_{i^*}, P_{i'}) \leqslant 3\varepsilon$. By triangle inequality we get $d_{\mathrm{TV}}(P_{i^*}, P) \leqslant 4\varepsilon$.

It suffices for each call to $\mathcal{B}$ succeed with probability at least $\frac{\delta}{2R^2}$. □

Assuming black-box access to $\mathcal{A}$, $O(m_{\mathcal{A}}\log\frac{1}{\delta})$ samples are needed in the worst case to learn with $1-\delta$ probability, since otherwise all the $o(\log\frac{1}{\delta})$ repetitions may fail. We can apply the above

algorithm to improve the success probability of learning Bayesian networks on a given graph with small indegree.