arXiv:2002.05378v2 [cs.DS] 14 Feb 2020

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

Arnab Bhattacharyya[†]
arnabb@nus.edu.sg

Sutanu Gayen[‡]
sutanugayen@gmail.com

Kuldeep S. Meel[§]
meel@comp.nus.edu.sg

N. V. Vinodchandran[¶]
vinod@unl.edu

February 17, 2020

Abstract

We design efficient distance approximation algorithms for several classes of structured high-dimensional distributions. Specifically, we show algorithms for the following problems:

- 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 $\mathrm{poly}(n, \varepsilon)$ 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 $\mathrm{poly}(n, \varepsilon)$ 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 $\mathrm{poly}(n, \varepsilon)$ 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 $\mathrm{poly}(n, \varepsilon)$ samples, 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

Our results are the first efficient distance approximation algorithms for these well-studied problems. They are derived using a simple and general connection to distribution learning algorithms. The distance approximation algorithms imply new efficient algorithms for *tolerant* testing of closeness of the above-mentioned structured high-dimensional distributions.

*Author names are in alphabetical order

[†]National University of Singapore. Supported in part by Start-up Grant WBS R252000A33133 and an Amazon Research Award.

[‡]National University of Singapore. Supported in part by AB's Start-up Grant WBS R252000A33133.

[§]National University of Singapore.

[¶]University of Nebraska, Lincoln. Research mostly conducted while visiting National University of Singapore.

1 Introduction

A fundamental challenge in statistics and computer science is to devise hypothesis tests that use a small number of samples. A classic problem of this type is *identity testing* (or, *goodness-of-fit testing*): given samples from an unknown distribution P over a domain $\mathcal{S}$, does P equal a specific reference distribution Q? A sequence of works [Pan08, BFR+13, VV14, 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. A related problem is *closeness testing* (or, *two-sample testing*): given samples from two unknown distributions P and Q over $\mathcal{S}$, does $P = Q$? Here, it is known that $\Theta(|\mathcal{S}|^{2/3}\varepsilon^{-4/3} + |\mathcal{S}|^{1/2}\varepsilon^{-2})$ 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 $d_{\mathrm{TV}}(P, Q) \leqslant \varepsilon_1$ and $d_{\mathrm{TV}}(P, Q) > \varepsilon_2$ where ε_1 and ε_2 are user-supplied parameters.

In this work, we design sample- and time-efficient tolerant identity and closeness testers for natural classes of distributions over Σ^n. More precisely, we focus on *distance approximation* algorithms:

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 $\mathcal{D}$.

Remark 1.2. *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 $2/3$.*

The distance approximation problem and the tolerant testing problem are equivalent in the setting we consider. A distance approximation algorithm for $(\mathcal{D}_1, \mathcal{D}_2)$ immediately gives a tolerant closeness testing algorithm for two input distributions $P \in \mathcal{D}_1$ and $Q \in \mathcal{D}_2$ with the same

asymptotic sample and time complexity bounds. Also a tolerant closeness testing algorithm for distributions in $\mathcal{D}_1$ and $\mathcal{D}_2$ gives a distance approximation algorithm for $(\mathcal{D}_1, \mathcal{D}_2)$, although with slightly worse sample and time complexity bounds (resulting from a binary search approach). Indeed this connection was explored in the property testing setting in [PRR06] which established a general translation result. Thus, in the rest of this paper we will focus on the distance approximation problem and the results translate to appropriate tolerant testing problems. The bounds on the sample and time complexity will be phrased in terms of the description lengths of $\mathcal{D}_1$ and $\mathcal{D}_2$.

2 New Results

We design new sample and time efficient distance approximation algorithms for several well-studied families of high-dimensional distributions given sample access. We accomplish this by prescribing a general strategy for designing distance approximation algorithms. In particular, we first design an algorithm to approximate the distance between a pairs of distributions. However, this algorithm needs both sample access and an approximate evaluation oracle. We crucially observe that a learning algorithm that outputs a representation of the unknown distribution given sample access, can often efficiently simulate the approximation oracle. Thus the final algorithm only needs sample access. This general strategy coupled with appropriate learning algorithms, leads to a number of new distance approximation algorithms (and hence new tolerant testers) for well-studied families of high-dimensional probability distributions.

2.1 Distance Approximation from EVAL Approximators

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 promised to be in $\mathcal{D}$, returns the description of a distribution $\hat{P}$ such that with probability at least $2/3$, $d_{\mathrm{TV}}(P, \hat{P}) \leqslant \varepsilon$. It turns out that for many natural distribution families $\mathcal{D}$ over Σ^n, one can easily modify known learning algorithms for $\mathcal{D}$ to efficiently output not just a description of $\hat{P}$ but the value of $\hat{P}(x) \coloneqq \mathbf{Pr}_{X \sim \hat{P}}[X = x]$ for any $x \in \Sigma^n$. More precisely, they yield what we call EVAL *approximators*:

Definition 2.1. *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*

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

Typically, the learning algorithm outputs parameters that describe $\hat{P}$, and then $\hat{P}(x)$ can be computed (or approximated) efficiently in terms of these parameters.

Example 2.2. *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 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.

Theorem 2.3. *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.*

Thus, in the context of Example 2.2, the above theorem immediately implies a distance approximation algorithm for product distributions using $O(n\varepsilon^{-2})$ samples and time. Theorem 2.3 extends the work of Canonne and Rubinfeld [CR14] who considered the setting $\beta = \gamma = 0$. We discuss the relation to prior work in Section 2.7.

2.2 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:

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

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 2.4. *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 $O(|\Sigma|^{d+1} mn)$ time.*

We design a learning algorithm for Bayesian networks on a known DAG G that 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 $x_{\text{parents}(i)} \in \Sigma^{\deg(i)}$. Given these conditional probability distributions, we can easily obtain $\hat{P}(x)$ for any x, and hence, an $(\varepsilon, 0)$-EVAL approximator for P, by using (1). Theorem 2.4 then follows from Theorem 2.3.

Theorem 2.4 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

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

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 2.4 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.

2.3 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 2.5. *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 2.5 follows from Theorem 2.3.

Daskalakis et al. [DDK19] studied independent 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 2.5 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, whether ferromagnetic or not. Below, U is the uniform distribution over $\{-1, 1\}^n$.

Theorem 2.6. *There is an algorithm which, given independent samples from an unknown Ising model P over $\{-1,1\}^n$ with width at most d, takes $m = O(e^{O(d)}\varepsilon^{-4}n^8\log(n/\varepsilon)+\varepsilon^{-7}\log^3\frac{1}{\varepsilon})$ samples, $O(mn^2+\varepsilon^{-7}n^2\log^3\frac{1}{\varepsilon})$ 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 2.6 again 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, but now, we can efficiently estimate the ratio $P(x)/P(y)$ between any two $x,y\in\{-1,1\}^n$. At this point, we invoke the uniformity tester shown by Canonne et al. [CRS15] that uses samples from the input distribution as well as pairwise conditional samples (the so-called PCOND oracle model).

2.4 Multivariate Gaussians

Theorem 2.3 applies also when Σ 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 that there is 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 $\Sigma\in\mathbb{R}^{n\times n}$, the distribution $N(\mu,\Sigma)$ has the density function:

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

Invoking our framework on multivariate gaussians, we obtain:

Theorem 2.7. *Let $\mathcal{D}$ be the family of multivariate gaussian distributions, $\{N(\mu,\Sigma):\mu\in\mathbb{R}^n,\Sigma\in\mathbb{R}^{n\times n},\Sigma\succ 0\}$. 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 that for any $P=N(\mu,\Sigma)$, the empirical mean $\hat{\mu}$ and empirical covariance $\hat{\Sigma}$ obtained from $O(n^2\varepsilon^{-2})$ samples from P determines a gaussian $\hat{P}=N(\hat{\mu},\hat{\Sigma})$ satisfying $d_{\mathrm{TV}}(P,\hat{P})\leqslant\varepsilon$ with probability at least $3/4$. To get an EVAL approximator, we need evaluations of $N(\hat{\mu},\hat{\Sigma};x)$ for any x as in (3). Since $\det(\hat{\Sigma})$ is computable in time $O(n^\omega)$, Theorem 2.7 follows from Theorem 2.3.

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 (see Corollary 6.3) yields a polynomial time randomized algorithm that, given $\mu_1,\Sigma_1,\mu_2,\Sigma_2$, approximates the total variation distance $d_{\mathrm{TV}}(N(\mu_1,\Sigma_1),N(\mu_2,\Sigma_2))$ upto $\pm\varepsilon$ additive error.

2.5 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 be more formal, 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 $X_{\text{parents}(i)}$ 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 $X_i \mid X_{\text{parents}(i)}$, where $X_{\text{parents}(i)}$ 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 such that for 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 a *strong positivity* condition on P, defined in Section 7. 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:

Theorem 2.8 (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, assuming the identifiability and strong positivity conditions.

Theorem 2.8 solves a very natural problem. To concoct a somewhat realistic example, 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

[†]That is, there exists a well-defined function mapping P to P_a but which may not be computationally effective.

for all individuals, and we further assume 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 2.8 to test whether the effect of the mutation is approximately the same for Africans and Europeans.

2.6 Improving Success of Learning Algorithms Using Distance Estimation

Finally we give a link between efficient distance approximation algorithms and boosting the success probability of learning algorithms. 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 $\mathcal{D}$. We prescribe a method to combine the two algorithms $\mathcal{A}$ and $\mathcal{B}$ to learn an unknown distribution from $\mathcal{D}$ with probability at least $(1-\delta)$. To the best of our knowledge, this connection has not been stated explicitly in the literature. The proof of the following theorem is given in Section 8.

Theorem 2.9. *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 $1-\delta$.*

To achieve the above result we repeat $\mathcal{A}$ independently $R = O(\log\frac{1}{\delta})$ times which guarantees at least $2R/3$ successful repetitions from Chernoff's bound except δ probability, which we condition on. Sucessful repetitions must produce distributions which are pairwise 2ε close by triangle inequality. We approximate the pairwise distances between all pairs of repetitions up to an additive ε and then find out a repetition whose learnt distribution $\hat{P}$ has the most number of other repetitions within 3ε distance. The later number must be at least $2R/3 - 1$, guaranteeing $\hat{P}$ must have a successful repetition within 3ε distance. Thus $\hat{P}$ must be at most 4ε close to P from triangle inequality.

2.7 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 $\tilde{O}(m^{2/3}\mathrm{poly}(1/\varepsilon))$ respectively. Optimal bounds for these testing problems were obtained in Valiant and Valiant [VV14] ($\Theta(\sqrt{m}/\varepsilon^2)$) 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 [VV11, 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.

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 (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 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 (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 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.

Another access model considered in the literature for which such domain independent results are obtained is the *conditional sampling model* introduced independently in Chakraborty et al. [CFGM16] and Canonne et al. [CRS14]. In this model, the tester has access to a conditional sampling oracle that given a subset S of the sample space outputs a random sample from the unknown distribution *conditioned on* S. The conditional sampling model lends itself to algorithms for testing uniformity and testing identity to a known distribution with sample complexity $\tilde{O}(1/\varepsilon^2)$. Building on Chakraborty et al. [CFGM16], Chakraborty and Meel [CM19] proposed a tolerant testing algorithm with sample complexity independent of domain size for testing unifor-

Algorithm 1: 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;
  $\alpha \leftarrow \mathcal{C}_P(x)$;
  $\beta \leftarrow \mathcal{C}_Q(x)$;
  $c_i \leftarrow 1_{\alpha > \beta} \left(1 - \frac{\beta}{\alpha}\right)$;
return $\frac{1}{t} \sum_{i=1}^{t} c_i$

mity of a sampler that takes in a Boolean formula φ as input and the sampler's output generates a distribution over the witnesses of φ.

3 Distance Approximation Algorithm

In this section, we prove Theorem 2.3 which underlies all the other results in this work. In fact, we show the following theorem that is more detailed.

Theorem 3.1. *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 3.1 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 3.1. 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.

Proof. We invoke Algorithm 1. 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 $\frac{2\gamma}{1-\gamma} + \beta + \varepsilon$.

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

$$= \mathop{\mathbf{E}}_{x\sim\hat{P}}\left[1_{\hat{P}(x)>\hat{Q}(x)}\left(1-\frac{\hat{Q}(x)}{\hat{P}(x)}\right)\right]$$

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.

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

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 Chernoff 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. □

Theorem 3.1 can be extended to the case that P and Q are distributions over $\mathbb{R}^n$ with infinite support. We change Definition 2.1 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 3.1 and Algorithm 1 continue to hold as stated. In the proof, we merely have to replace the summations with the appropriate integrals.

4 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 4.1. *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:*

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

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. Such a learning algorithm was claimed in the appendix of [CDKS17] but the analysis there appears to be incomplete [Can20]. We show the following proper learning algorithm for Bayesian networks that uses the optimal sample complexity.

Theorem 4.2. *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 $O(|\Sigma|^{d+1} mn)$ time.*

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

Theorem 2.4. *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 $O(|\Sigma|^{d+1} mn)$ time.*

Proof. Given samples from P_1 and P_2 we first learn them as $\hat{P}_1$ and $\hat{P}_2$ using Theorem 4.2 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 3.1 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. □

Our distance estimation algorithm has optimal dependence on n and ε from the following non-tolerant identity testing lower bound of Daskalakis et al.

Theorem 4.3 ([DDK19])**.** *Given sample access to two unknown Bayesian network distributions P_1 and P_2 over $\{0,1\}^n$ on a common known graph, testing $P = Q$ versus $d_{\mathrm{TV}}(P, Q) \geqslant \varepsilon$ with probability $\geqslant 2/3$ requires $\Omega(n\varepsilon^{-2})$ samples.*

It remains to prove Theorem 4.2.

4.1 Learning Bayesian networks

In this section, we prove a strengthened version of Theorem 4.2 that holds for any desired error probability δ.

Theorem 4.4. *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 $O(|\Sigma|^{d+1} mn)$ 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 2 for the binary alphabet case ($\Sigma = \{0, 1\}$) and reduce the general case to the binary case afterwards.

Algorithm 2: 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} ← the number of samples where i's parents are set to a;
        if N_{i,a} ⩾ t then
            Q(i | a) ← the Laplace corrected empirical distribution at node i in the subset of
              samples where i's parents are set to a;
        else
            Q(i | a) ← uniformly random bit;
```

The *Laplace corrected 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 4.5 ([KOPS15])**.** *Let D be an unknown distribution over k items. Let $\hat{D}$ be the Laplace corrected 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 4.6. *Let P and Q be two Bayesian networks over the same graph G. Then*

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

Lemma 4.7. *For $m = 24n2^d \log(n2^d)/\varepsilon$ and $t = 12\log(n2^d)$, Algorithm 2 satisfies $KL(P,Q) \leqslant 6\varepsilon$ 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.

Consider the event "all heavy (i,a) tuples satisfy $N_{i,a} \geqslant n2^d P[\Pi[i,a]]t/\varepsilon$". It is easy to see from Chernoff and union bounds that this event holds with 19/20 probability. Hence for the rest of the argument, we condition on this event. In this case, all heavy items satisfy $N_{i,a} \geqslant t$.

Now, we see that:

– For any heavy (i,a), by Theorem 4.5,

$$\mathbf{E}[\mathrm{KL}(P(i \mid a), Q(i \mid a))] \leqslant \frac{\varepsilon}{10n2^d \cdot P[\Pi[i,a]]}.$$

– For any light (i, a) that satisfies $N_{i,a} \geqslant t$, it follows from Theorem 4.5 that $\mathbf{E}[\mathrm{KL}(P(i \mid a), Q(i \mid a))] \leqslant 1$.

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

Using Theorem 4.6, we get

$$\mathbf{E}[\mathrm{KL}(P, Q)] \leqslant \sum_{(i,a) \text{ heavy}} P[\Pi[i, a]] \cdot \frac{\varepsilon}{10n2^d \cdot P[\Pi[i, a]]} + \sum_{(i,a) \text{ light}} \frac{\varepsilon}{n2^d} \cdot 1 \leqslant 1.1\varepsilon.$$

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 4.7 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 2.9.

5 Ising Models

In this section, we give a distance approximation algorithm for the class of bounded-width ferromagnetic Ising models. Recall from Section 2.3 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:

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

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 $A_{ij} \geqslant 0$.

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 5.1 (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 5.2. *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 2.5. *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 5.1 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 5.1, 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 5.2 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 3.1 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. □

5.1 Distance to uniformity

Next we give an algorithm for estimating the distance between an unknown Ising model and the uniform distribution over $\{-1,1\}^n$.

Theorem 2.6. *There is an algorithm which, given independent samples from an unknown Ising model P over $\{-1,1\}^n$ with width at most d, takes $m = O(e^{O(d)}\varepsilon^{-4}n^8\log(n/\varepsilon) + \varepsilon^{-7}\log^3\frac{1}{\varepsilon})$ samples, $O(mn^2 + \varepsilon^{-7}n^2\log^3\frac{1}{\varepsilon})$ 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 ising model using Theorem 5.1. 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 5.1. Canonne et al. [CRS15] have given an algorithm for computing distance to uniformity from an unknown distribution using sampling and pairwise conditional sampling (PCOND) access to it using $m_1 = O(\varepsilon^{-19}\log^8\frac{1}{\varepsilon})$ PCOND samples and $m_2 = O(\varepsilon^{-7}\log^3\frac{1}{\varepsilon})$ samples with probability 2/3 up to a $O(\varepsilon)$ additive error. A closer look at their algorithm reveals that all their PCOND accesses are made from a routine called 'COMPARE', whose job is to compute the ratio of probabilities

of two points x and z with probability $1-\delta$ upto $(1\pm\eta)$-factor using conditional samples. In fact it suffices for their algorithm to correctly compute the ratio if it is in $[1/K, K]$, report 'HIGH' if it is in $(K,\infty]$, and 'LOW' if it is in $[0, 1/K)$ for a parameter K. In the case of ising model, assuming success of Theorem 5.1 we can replace the routine 'COMPARE' by computing N_x/N_z using the parameters of the learnt model upto $(1\pm\varepsilon)$ approximation in $O(n^2)$ time with $\delta = 0$. Their algorithm makes $m_3 = O(\varepsilon^{-7}\log^3\frac{1}{\varepsilon})$ calls to 'COMPARE'. Using their choices of various parameters our theorem follows. □

6 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 $\Sigma \in \mathbb{R}^{n\times n}$, the Gaussian distribution $N(\mu,\Sigma)$ has the pdf:

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

We use the following folklore learning result for learning the two Gaussians.

Theorem 6.1. *Let P be an n-dimensional Gaussian distribution. Let $\hat{\mu}\in\mathbb{R}^n$ and $\hat{\Sigma}\in\mathbb{R}^{n\times n}$ 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 $\hat{P} = N(\hat{\mu},\hat{\Sigma})$ satisfies $d_{\mathrm{TV}}(P,\hat{P}) \leqslant \varepsilon$.*

We are now ready to prove Theorem 2.7 restated below.

Theorem 2.7. *Let $\mathcal{D}$ be the family of multivariate gaussian distributions, $\{N(\mu,\Sigma) : \mu\in\mathbb{R}^n, \Sigma\in\mathbb{R}^{n\times n}, \Sigma\succ 0\}$. 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 6.1 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 3.1, 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 6.2. *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.*

As a by-product of our analysis, we also obtain an efficient randomized algorithm to compute the total deviation distance between two gaussians specified by their parameters.

Corollary 6.3. *For any two vectors $\mu_1,\mu_2\in\mathbb{R}^n$ and two positive-definite matrices $\Sigma_1,\Sigma_2\in\mathbb{R}^{n\times n}$, $d_{\mathrm{TV}}(N(\mu_1,\Sigma_1),N(\mu_1,\Sigma_1))$ can be estimated up to an additive ε error in $O(n^3\varepsilon^{-2})$ time.*

Proof. We again invoke Algorithm 1. Since the parameters are already provided, we can readily obtain $(0,0)$-EVAL approximators for $N(\mu_1,\Sigma_1)$ and $N(\mu_2,\Sigma_2)$. For Algorithm 1, we also need sample access to one of the two distributions. It is well known that if $v\sim N(0,I)$ and $\Sigma = LL^\top$, then $Lv+\mu\sim N(\mu,\Sigma)$; the matrix L can be obtained in $O(n^3)$ time using a Cholesky decomposition. Hence, each sample from $N(\mu_1,\Sigma_1)$ costs $O(n^3)$ time, so that the entire algorithm runs in $O(n^3\varepsilon^{-2})$ time. □

7 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 7.1 (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 $x \in \Sigma^{|V|}$:*

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

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 7.2 (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 $\mathsf{Pa}^+(S) = S \cup \mathsf{Pa}(S)$.

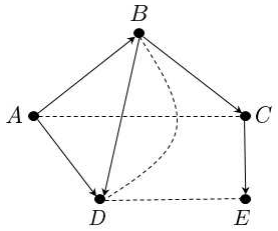

Figure 1: An acyclic directed mixed graph (ADMG) where the bidirected edges are depicted as dashed. The in-degree of the graph is 2. The c-components are $\{A, C\}$ and $\{B, D, E\}$.

Assumption 7.3 (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 7.2) and α-strong positivity (Assumption 7.3) 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 7.4 (EVAL approximator [BGK+20])**.** *For any intervention a to A and parameter $\varepsilon \in (0, 1)$, there is an algorithm that takes $m = \tilde{O}\left(\frac{|\Sigma|^{5kd} n}{\alpha \varepsilon^2}\right)$ samples from P, and in $O(mn|\Sigma|^{2kd})$ time, returns a circuit $E_{P,a}$. With probability at least $2/3$, the circuit $E_{P,a}$ implements an $(\varepsilon, 0)$-EVAL approximator for P_a, and it runs in $O(n)$ time for all inputs.*

Theorem 7.5 (Generator [BGK+20])**.** *For any intervention a to A and parameter $\varepsilon \in (0, 1)$, there is an algorithm that takes $m = \tilde{O}\left(\frac{|\Sigma|^{5kd} n}{\alpha \varepsilon^2}\right)$ samples from P, and in $O(mn|\Sigma|^{2kd})$ time, returns a probabilistic circuit $G_{P,a}$ that generates samples of a distribution $\tilde{P}_a$ satisfying $d_{\mathrm{TV}}(P_a, \tilde{P}_a) \leqslant \varepsilon$. On each call, the circuit takes $O(n|\Sigma|^{2kd}\varepsilon^{-1} \log \delta^{-1})$ time and outputs a sample of $\tilde{P}_a$ with probability at least $1 - \delta$.*

We give a distance approximation algorithm for identifiable atomic interventional distributions using the above two results and Theorem 3.1.

Theorem 7.6 (Formal version of Theorem 2.8)**.** *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 7.2 and Assumption 7.3 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 $m = \tilde{O}\left(\frac{|\Sigma|^{5kd} n}{\alpha \varepsilon^2}\right)$ samples from P and Q, runs in time $\tilde{O}(mn|\Sigma|^{2kd} + n|\Sigma|^{2kd}\varepsilon^{-3})$ 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 7.5 to obtain the generators for distributions $\tilde{P}_a$ and $\tilde{Q}_a$ that are $\varepsilon/10$ close to the two interventional distributions P_a and Q_a respectively in d_{TV}. By triangle inequality, it suffices to estimate $d_{\mathrm{TV}}(\tilde{P}_a, \tilde{Q}_a)$ up to an additive $4\varepsilon/5$ error. Next we invoke

Theorem 7.4 to obtain circuits $E_{P,a}$ and $E_{\mathcal{Q},a}$ that each implement $(\varepsilon/10, 0)$-EVAL approximators for the two interventional distributions P_a and Q_a respectively. Let $\hat{P}_a$ and $\hat{Q}_a$ denote the two distributions that $E_{P,a}$ and $E_{Q,a}$ respectively compute evaluations of. Using the triangle inequality, $d_{\mathrm{TV}}(\tilde{P}_a, \hat{P}_a) \leqslant \varepsilon/5$ and $d_{\mathrm{TV}}(\tilde{Q}_a, \hat{Q}_a) \leqslant \varepsilon/5$. Thus $E_{P,a}$ and $E_{Q,a}$ are $(\varepsilon/5, 0)$-EVAL approximators for $\tilde{P}_a$ and $\tilde{Q}_a$ respectively. From Theorem 3.1, we need $O(\varepsilon^{-2})$ samples from $\tilde{P}_a$ and $O(\varepsilon^{-2})$ calls to $E_{P,a}$ and $E_{Q,a}$ to estimate $d_{\mathrm{TV}}(\tilde{P}_a, \tilde{Q}_a)$ up to an additive $4\varepsilon/5$ error. □

8 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 3: 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 ⩽ i ⩽ R = O(log 1/δ) do
    P_i ← Run A on samples from P to get a learnt distribution;
    count_i ← 0;
for every unordered pair 0 ⩽ i < j ⩽ R do
    d_ij ← Estimate distance between P_i and P_j up to additive error ε using B;
    if d_ij ⩽ 3ε then
        count_i ← count_i + 1;
        count_j ← count_j + 1;
i* = arg max_i count_i;
return P_{i*};
```

Theorem 2.9. *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 3. 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 and multidimensional Gaussians.

References

[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*, NIPS18, page 94699481, Red Hook, NY, USA, 2018. Curran Associates Inc. 9, 18

[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. 2, 8

[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. 7, 18, 19

[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. 9

[Can20] Clément Canonne, Jan 2020. Personal communication. 13

[Cav78] James A Cavender. Taxonomy with confidence. *Mathematical biosciences*, 40(3-4):271–280, 1978. 5

[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. 4, 5, 9, 13, 14

[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. 2, 8

[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. 9

[CM19] Sourav Chakraborty and Kuldeep S. Meel. On testing of uniform samplers. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, 1 2019. 9

[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. 4, 9, 10

[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. 9

[CRS15] Clément L. Canonne, Dana Ron, and Rocco A. Servedio. Testing probability distributions using conditional samples. *SIAM J. Comput.*, 44(3):540–616, 2015. 6, 16

[DDK19] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing ising models. *IEEE Trans. Information Theory*, 65(11):6829–6852, 2019. 5, 9, 13

[DMR18] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians. *arXiv preprint arXiv:1810.08693*, 2018. 6

[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. 4, 9

[Ell93] Glenn Ellison. Learning, local interaction, and coordination. *Econometrica: Journal of the Econometric Society*, pages 1047–1071, 1993. 5

[Far73] James S Farris. A probability model for inferring evolutionary trees. *Systematic Biology*, 22(3):250–256, 1973. 5

[Gác18] Pétér Gács, Feb 2018. From László Lovász's lecture notes. http://www.cs.bu.edu/faculty/gacs/courses/cs530/lectures/exact-Gauss.pdf. 17

[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. 5

[GR11] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, pages 68–75. Springer, 2011. 8

[Isi25] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift f́ur Physik A Hadrons and Nuclei*, 31(1):253–258, 1925. 5

[JN07] Finn V. Jensen and Thomas D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer Publishing Company, Incorporated, 2nd edition, 2007. 4

[JS93] Mark Jerrum and Alistair Sinclair. Polynomial-time approximation algorithms for the ising model. *SIAM Journal on computing*, 22(5):1087–1116, 1993. 5, 15

[KF09] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. 4

[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. 5, 15

[KOPS15] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In Peter Grnwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1066–1100, Paris, France, 03–06 Jul 2015. PMLR. 14

[MDLW18] Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright. *Handbook of Graphical Models*. CRC Press, 2018. 6

[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. 5

[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. 5

[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. 2, 8

[Pea09] Judea Pearl. *Causality*. Cambridge university press, 2009. 7, 9, 18

[PRR06] Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *J. Comput. Syst. Sci.*, 72(6):1012–1042, 2006. 3

[Rub12] Ronitt Rubinfeld. Taming big probability distributions. *ACM Crossroads*, 19(1):24–28, 2012. 9

[Sri19] Piyush Srivastava, Nov 2019. Personal communication. 15

[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. 5

[Tia02] Jin Tian. *Studies in causal reasoning and learning*. University of California, Los Angeles, 2002. 7

[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. 7, 18

[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*, UAI02, page 519527, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. 18

[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. 18

[VV10] Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:183, 2010. 9

[VV11] 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. 9

[VV14] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, FOCS '14, pages 51–60, Washington, DC, USA, 2014. IEEE Computer Society. 2, 8

[Wil65] J.H. Wilkinson. *The Algebraic Eigenvalue Problem*. Clarendon Press, 1965. 17