Testing Bayesian Networks

Clément L. Canonne   CCANONNE@CS.COLUMBIA.EDU
Columbia University

Ilias Diakonikolas   DIAKONIK@USC.EDU
University of Southern California

Daniel M. Kane   DAKANE@CS.UCSD.EDU
University of California, San Diego

Alistair Stewart   STEWART.AL@GMAIL.COM
University of Southern California

Abstract

This work initiates a systematic investigation of testing *high-dimensional* structured distributions by focusing on testing *Bayesian networks* – the prototypical family of directed graphical models. A Bayesian network is defined by a directed acyclic graph, where we associate a random variable with each node. The value at any particular node is conditionally independent of all the other non-descendant nodes once its parents are fixed. Specifically, we study the properties of identity testing and closeness testing of Bayesian networks. Our main contribution is the first non-trivial efficient testing algorithms for these problems and corresponding information-theoretic lower bounds. For a wide range of parameter settings, our testing algorithms have sample complexity *sublinear* in the dimension and are sample-optimal, up to constant factors.

Keywords: distribution testing, property testing, Bayesian networks, graphical models

1. Introduction

1.1. Background

Distribution testing has its roots in statistical hypothesis testing Neyman and Pearson (1933); Lehmann and Romano (2005) and was initiated in Goldreich and Ron (2000); Batu et al. (2000). The paradigmatic problem in this area is the following: given sample access to an arbitrary distribution P over a domain of size N, determine whether P has some global property or is "far" from any distribution having the property. A natural way to solve this problem would be to learn the distribution in question to good accuracy, and then check if the corresponding hypothesis is close to one with the desired property. However, this testing-via-learning approach requires $\Omega(N)$ samples and is typically suboptimal. The main goal in this area is to obtain *sample-optimal* testers – ideally, testers that draw $o(N)$ samples from the underlying distribution. During the past two decades, a wide range of properties have been studied, and we now have sample-optimal testers for many of these properties Paninski (2008); Chan et al. (2014c); Valiant and Valiant (2014); Diakonikolas and Kane (2016); Diakonikolas et al. (2016a).

We remark that even for the simplest properties, e.g., identity testing, at least $\Omega(\sqrt{N})$ many samples are required for arbitrary distributions over N atoms. While this is an improvement over the $\Omega(N)$ samples required to learn the distribution, a sample upper bound of $O(\sqrt{N})$ is still

impractical if N is very large. For example, suppose that the unknown distribution is supported on $\{0, 1\}^n$. For this high-dimensional setting, a sample complexity bound of $\Theta(2^{n/2})$ quickly becomes prohibitive, when the dimension increases. Notably, the aforementioned $\Omega(\sqrt{N})$ sample lower bound characterizes worst-case instances, which in many cases are unlikely to arise in real-world data. This observation motivates the study of testing *structured* distribution families, where significantly improved testers may be possible. Hence, the following natural question arises: *Can we exploit the structure of the data to perform the desired testing task more efficiently?*

A natural formalization of this question involves viewing the data as samples from a *probabilistic model* – a model that we believe represents the random process generating the samples. The usual assumption is that there exists a known family of probabilistic models – describing a set of probability distributions – and that the data are random samples drawn from an unknown distribution in the family. In this context, the distribution testing problem is the following: Let $\mathcal{C}$ be a family of probabilistic models. The *testing algorithm* has access to independent samples from an unknown $P \in \mathcal{C}$, and its goal is to output "yes" if P has some property $\mathcal{P}$, and output "no" if the total variation distance, $d_{\mathrm{TV}}(P, Q) \stackrel{\text{def}}{=} (1/2)\|P - Q\|_1$, where $\|\cdot\|_1$ denotes the L_1-norm, is at least ϵ to *every* $Q \in \mathcal{C}$ that has property $\mathcal{P}$. The sample complexity of this structured testing problem depends on the underlying family $\mathcal{C}$, and we are interested in obtaining efficient algorithms that are *sample optimal* for $\mathcal{C}$.

More than a decade ago, Batu, Kumar, and Rubinfeld Batu et al. (2004) considered a specific instantiation of this broad question – testing the equivalence between two unknown discrete monotone distributions – and obtained a tester whose sample complexity is poly-logarithmic in the domain size. A recent sequence of works Daskalakis et al. (2013b); Diakonikolas et al. (2015a,b) developed a framework to obtain sample-optimal estimators for testing the identity of structured distributions *over total orders* (e.g., univariate multi-modal or log-concave distributions). The main lesson of these works is that, under reasonable structural assumptions, the sample complexity of testing may dramatically improve – becoming sub-logarithmic or even independent of the support size. Moreover, in all studied cases, one obtains testers with *sub-learning* sample complexity.

1.2. This Work: Testing High-Dimensional Structured Distributions

This paper initiates a systematic investigation of testing properties of *high-dimensional* structured distributions. One of the most general formalisms to succinctly represent such distributions is provided by probabilistic graphical models Wainwright and Jordan (2008); Koller and Friedman (2009). Graphical models compactly encode joint probability distributions in high dimensions. Formally, a graphical model is a graph where we associate a random variable with each node. The key property is that the edge-structure of the graph determines the dependence relation between the nodes.

The general problem of inference in graphical models is of fundamental importance and arises in many applications across several scientific disciplines, see Wainwright and Jordan (2008) and references therein. In particular, the task of learning graphical models has been extensively studied Neapolitan (2003); Daly et al. (2011). A range of information-theoretic and algorithmic results have been developed during the past five decades in various settings, see, e.g., Chow and Liu (1968); Dasgupta (1997); Friedman and Yakhini (1996); Friedman et al. (1997, 2000); Cheng et al. (2002); Chickering (2002); Margaritis (2003); Abbeel et al. (2006); Wainwright et al. (2006); Anandkumar et al. (2012); Santhanam and Wainwright (2012); Loh and Wainwright (2012); Diakonikolas et al.

(2016c) for a few references. In contrast, the general question of *testing* graphical models has received less attention. We propose the following broad set of questions:

Question 1.1 Let $\mathcal{C}$ be a family of high-dimensional graphical models and $\mathcal{P}$ be a property of $\mathcal{C}$. What is the *sample complexity* of testing whether an unknown $P \in \mathcal{C}$ has property $\mathcal{P}$? Can we develop testers for $\mathcal{P}$ with *sub-learning* sample complexity? Can we design *sample-optimal* and *computationally efficient* testers?

We believe that Question 1.1 points to a fundamental research direction that warrants study for its own sake. Moreover, as we explain in the following paragraphs, such estimation tasks arise directly in various practical applications across the data sciences, where sample efficiency is of critical importance. Hence, improved estimators for these tasks may have implications for the analysis of datasets in these areas.

For concreteness, Question 1.1 refers to a single unknown distribution that we have sample access to. We are also naturally interested in the broader setting of testing properties for *collections* of distributions in $\mathcal{C}$. Before we proceed to describe our contributions, a few comments are in order: As previously mentioned, for all global properties of interest (e.g., identity, independence, etc.), the sample complexity of testing the property is bounded from above by the sample complexity of learning an arbitrary distribution from $\mathcal{C}$. Hence, the overarching goal is to obtain testers that use fewer samples than are required to actually learn the model – or to prove that this is impossible. On a related note, in the well-studied setting of testing arbitrary discrete distributions, the main challenge has been to devise sample-optimal testers; the algorithmic aspects are typically straightforward. This is no longer the case in the high-dimensional setting, where the combinatorial structure of the underlying model may pose non-trivial algorithmic challenges.

In this work, we start this line of inquiry by focusing on testing *Bayesian networks* Pearl (1988) (*Bayes nets* or BN for brevity), the prototypical family of *directed* graphical models. Bayesian networks are used for modeling beliefs in many fields including robotics, computer vision, computational biology, natural language processing, and medicine Jensen and Nielsen (2007); Koller and Friedman (2009). Formally, a Bayesian network is defined by a directed acyclic graph (DAG) $\mathcal{S} = (V, E)$, where we associate a random variable with each node. Moreover, the value at any particular node is conditionally independent of all the other non-descendant nodes once its parents are fixed. Hence, for a fixed topology, it suffices to specify the conditional distribution for each node for each configuration of values for its parents.

The main problems that we study in this setting are the related tasks of testing *identity* and *closeness*: In identity testing, we are given samples from an unknown Bayes net P and we want to distinguish between the case that it is equal to versus significantly different from an explicitly given Bayes net Q. In closeness testing, we want to test whether two unknown Bayes nets P, Q are identical versus significantly different. We believe that our techniques can be naturally adapted to test other related properties (e.g., independence), but we have not pursued this direction in the current paper. A related testing problem that we consider is that of *structure testing*: given samples from an unknown Bayes net P, we want to test whether it can be represented with a given graph structure $\mathcal{S}$ or is far from any Bayes net with this structure.

In the prior work on testing *unstructured* discrete distributions, the natural complexity measure was the domain size of the unknown distributions. For the case of Bayes nets, the natural complexity measures are the number of variables (nodes of the DAG) – denoted by n – the maximum in-degree of the DAG – denoted by d – and the alphabet size of the discrete distributions on the nodes. To

avoid clutter in the expressions, we focus on the natural setting that the random variables associated with each node are Bernoulli's, i.e., the domain of the underlying distributions is $\{0, 1\}^n$. (As we will point out, our bounds straightforwardly extend to the case of general alphabets with a necessary polynomial dependence on the alphabet size.)

We note that Bayes nets are a universal representation scheme: *Any* distribution over $\{0, 1\}^n$ can be presented as a BN, if the maximum in-degree d of the graph is unbounded. (Indeed, for $d = n - 1$, one can encode all distributions over $\{0, 1\}^n$.) In fact, as we will see, the sample complexity of testing scales exponentially with d. Therefore, an upper bound on the maximum in-degree is *necessary* to obtain non-trivial upper bounds. Indeed, the most interesting regime is the settting where the number of nodes n is large and the degree d is small. In applications of interest, this assumption will be automatically satisfied. In fact, as we explain in the following subsection, in many relevant applications the maximum in-degree is either 1 (i.e., the graph is a tree) or bounded by a small constant.

1.3. Related Work

We partition the related work intro three groups corresponding to research efforts by different communities.

Computer Science. A large body of work in computer science has focused on designing statistically and computationally efficient algorithms for learning structured distributions in both low and high dimensions Dasgupta (1999); Freund and Mansour (1999); Arora and Kannan (2001); Vempala and Wang (2002); Cryan et al. (2002); Mossel and Roch (2005); Moitra and Valiant (2010); Belkin and Sinha (2010); Daskalakis et al. (2012a,b); Chan et al. (2013); Daskalakis et al. (2013a); Chan et al. (2014a,b); Hardt and Price (2015); Acharya et al. (2015b); De et al. (2015); Daskalakis et al. (2016a); Diakonikolas et al. (2015c, 2016b). On the other hand, the vast majority of the literature in distribution property testing during the past two decades focused on arbitrary discrete distributions, where the main complexity measure was the domain size. See Batu et al. (2000, 2001); Batu (2001); Batu et al. (2002, 2004); Paninski (2008); Valiant and Valiant (2011); Daskalakis et al. (2013b); Acharya et al. (2011); Levi et al. (2011); Indyk et al. (2012); Chan et al. (2014c); Valiant and Valiant (2014); Acharya et al. (2015a); Canonne et al. (2016); Diakonikolas and Kane (2016) for a sample of works, or Rubinfeld (2012); Canonne (2015) for surveys.

A line of work Batu et al. (2004); Daskalakis et al. (2013b); Diakonikolas et al. (2015a,b) studied properties of one-dimensional structured distribution families under various "shape restrictions" on the underlying density. In the high-dimensional setting, Rubinfeld and Servedio Rubinfeld and Servedio (2005) studied the identity testing problem for monotone distributions over $\{0, 1\}^n$. It was shown in Rubinfeld and Servedio (2005) that $\mathrm{poly}(n)$ samples suffice for the case of uniformity testing, but the more general problems of identity testing and independence testing require $2^{\Omega(n)}$ samples. Subsequently, Adamaszek, Cjumaj, and Sohler Adamaszek et al. (2010) generalized these results to continuous monotone distributions over $[0, 1]^n$. A related, yet distinct, line of work studied the problem of testing *whether* a probability distribution has a certain structure Batu et al. (2004); Bhattacharyya et al. (2011); Acharya et al. (2015a); Canonne et al. (2016). The sample complexity bounds in these works scale exponentially with the dimension. Finally, concurrent work of Daskalakis et al. (2016b) considers the questions of testing identity and independence of Ising models.[1]

---

1. Ising models constitute another type of graphical model, roughly speaking where the underlying graph is undirected.

Statistics. The area of hypothesis testing for high-dimensional models has a long history in statistics and is currently an active topic of study. A sequence of early and recent works, starting with Weiss (1960); Bickel (1969); Liu and Singh (1993), has studied the problem of testing the equivalence between two nonparametric high-dimensional distributions in the asymptotic regime. In the parametric setting, Hotelling's T-squared statistic Hotelling (1931) is the classical test for the equivalence of two high-dimensional Gaussians (with known and identical covariance). However, Hotelling's test has the serious defect that it fails when the sample size is smaller than the dimension of the data Bai and Saranadasa (1996). Recent work has obtained testers that, under a high-dimensional Gaussian model (with known covariance), succeed in the sub-linear regime for testing identity Srivastava and Du (2008) and closeness Chen and Qin (2010). A number of more recent works study properties of covariance matrices Cai and Ma (2013), regression Javanmard and Montanari (2014), and linear independence testing Ramdas et al. (2016).

Applications. The problems of testing identity and closeness of Bayesian networks arise in a number of applications where sample efficiency is critical Friedman et al. (2000); Gonen et al. (2003); Sobel and Kushnir (2003); Almudevar (2010); Nguyen et al. (2011); Rahmatallah et al. (2014); Städler and Mukherjee (2015); Yin et al. (2015). In bioinformatics applications (e.g., gene set analysis), each sample corresponds to an experiment that may be costly or ethically questionable Yin et al. (2015). Specifically, Yin et al. (2015) emphasizes the need of making accurate inferences on *tree structured* Bayesian networks, using an extremely small sample size – significantly smaller than the number of variables (nodes). Almudevar (2010) studies the problem of testing closeness between two unknown Bayesian network models in the context of a biology application, where Bayes nets are used to model gene expression data. The motivation in Almudevar (2010) comes from the need to compare network models for a common set of genes under varying phenotypes, which can be formulated as the problem of testing closeness between two unknown Bayes nets. As argued in Almudevar (2010), due to the small sample size available, it is not feasible to directly learn each BN separately.

Basic Notation and Definitions. Consider a directed acyclic graph (DAG), $\mathcal{S}$, with n vertices that are topologically sorted, i.e., labelled from the set $[n] \stackrel{\text{def}}{=} \{1, 2, \ldots, n\}$ so that all directed edges of $\mathcal{S}$ point from vertices with smaller label to vertices with larger label. A probability distribution P over $\{0, 1\}^n$ is defined to be a *Bayesian network* (or Bayes net) with dependency graph $\mathcal{S}$ if for each $i \in [n]$, we have that $\Pr_{X \sim P}[X_i = 1 \mid X_1, \ldots, X_{i-1}]$ depends only on the values X_j, where j is a parent of i in $\mathcal{S}$. Such a distribution P can be specified by its *conditional probability table*, i.e., the vector of conditional probabilities of $X_i = 1$ conditioned on every possible combination of values to the coordinates of X at the parents of i.

To formalize the above description, we use the following terminology. We will denote by $\text{Parents}(i)$ the set of parents of node i in $\mathcal{S}$. For a vector $X = (X_1, \ldots, X_n)$ and a subset $A \subseteq [n]$, we use X_A to denote the vector $(X_i)_{i \in A}$. We can now give the following definition:

Definition 1 *Let S be the set $\{(i, a) : i \in [n], a \in \{0, 1\}^{|\text{Parents}(i)|}\}$ and $m = |S|$. For $(i, a) \in S$, the* parental configuration $\Pi_{i,a}$ *is defined to be the event that $X_{\text{Parents}(i)} = a$. Once $\mathcal{S}$ is fixed, we may associate to a Bayesian network P the conditional probability table $p \in [0, 1]^S$ given by $p_{i,a} = \Pr_{X \sim P}[X_i = 1 \mid \Pi_{i,a}]$, for $(i, a) \in S$. We note that the distribution P is determined by p.*

We will frequently index p as a vector. That is, we will use the notation p_k, for $1 \le k \le m$, and the associated events Π_k, where each k stands for an $(i, a) \in S$ lexicographically ordered.

2. Testing Identity of Product Distributions

Before considering the general case of testing properties of Bayesian nets and discussing our results in Section 3, we first provide in this section both a sample-optimal efficient algorithm and a matching information-theoretic lower bound for the related, but simpler question of testing identity of *product distributions* over $\{0, 1\}^n$. Our results for this setting can be viewed as discrete analogues of testing identity and closeness of high-dimensional spherical Gaussians, that have been studied in the statistics literature Hotelling (1931); Bai and Saranadasa (1996); Srivastava and Du (2008); Chen and Qin (2010). We note that the Gaussian setting is simpler since the total variation distance can be bounded by the Euclidean distance between the mean vectors, instead of the chi-squared distance.

The structure of this section is as follows: In Section 2.1, we give an identity testing algorithm for n-dimensional binary product distributions with sample complexity $O(\sqrt{n}/\epsilon^2)$. In Section 2.2, we show that this sample bound is information-theoretically optimal.

2.1. Identity Testing Algorithm

As mentioned above, here we are concerned with the problem of testing the identity of an unknown product P with mean vector p against an explicit product distribution Q with mean vector q. Our tester relies on a statistic providing an unbiased estimator of $\sum_i (p_i - q_i)^2/(q_i(1 - q_i))$. Essentially, every draw from P gives us an independent sample from each of the coordinate random variables. In order to relate our tester more easily to the analogous testers for unstructured distributions over finite domains, we consider $\mathrm{Poi}(m)$ samples from each of these coordinate distributions. From there, we construct a random variable Z that provides an unbiased estimator of our chi-squared statistic, and a careful analysis of the variance of Z shows that with $O(\sqrt{n}/\epsilon^2)$ samples we can distinguish between $P = Q$ and P being ϵ-far from Q; leading to the following theorem:

Theorem 2 *There exists a computationally efficient algorithm which, given an explicit product distribution Q (via its mean vector), and sample access to an unknown product distribution P over $\{0, 1\}^n$, has the following guarantees: For any $\epsilon > 0$, the algorithm takes $O\left(\sqrt{n}/\epsilon^2\right)$ samples from P, and distinguishes with probability $2/3$ between the cases that $P = Q$ versus $\|P - Q\|_1 > \epsilon$.*

Proof Let $Q = Q_1 \otimes \cdots \otimes Q_n$ be a known product distribution over $\{0, 1\}^n$ with mean vector q, and $P = P_1 \otimes \cdots \otimes P_n$ be an unknown product distribution on $\{0, 1\}^n$ with unknown mean vector p. The goal is to distinguish, given independent samples from P, between $P = Q$, and $\|P - Q\|_1 > \epsilon$.

Let $0 < \gamma < 1/2$. We say that a product distribution P over $\{0, 1\}^n$ is γ-*balanced* if its mean vector p satisfies $p_i \in [\gamma, 1 - \gamma]$ for all $i \in [n]$. To prove Theorem 2, we can assume without loss of generality that P, Q are γ_0-balanced for $\gamma_0 \stackrel{\text{def}}{=} \frac{\epsilon}{16n}$. Indeed, given sample access to a product distribution P, we can simulate access to the γ_0-balanced product distribution P' by re-randomizing independently each coordinate with probability $2\gamma_0$, choosing it then to be uniform in $\{0, 1\}$. The resulting product distribution P' is γ_0-balanced, and satisfies $\|P - P'\|_1 \leq n \cdot \gamma_0 \leq \frac{\epsilon}{4}$. Therefore, to test the identity of a product distribution P against a product distribution Q with parameter ϵ, it is sufficient to test the identity of the γ_0-balanced product distributions P', Q' (with parameter $\frac{\epsilon}{2}$).

Preprocessing. We also note that by flipping the coordinates i such that $q_i > 1/2$, we can assume that $q_i \in [\gamma_0, 1/2]$ for all $i \in [n]$. This can be done without loss of generality, as q is explicitly given. For any i such that $q_i > \frac{1}{2}$, we replace q_i by $1 - q_i$ and work with the corresponding distribution Q'

Input Error tolerance ϵ, dimension n, balancedness parameter $\gamma \geq \frac{\epsilon}{16n}$, mean vector $q = (q_1, \ldots, q_n) \in [\gamma, 1/2]^n$ of an explicit product distribution Q over $\{0,1\}^n$, and sampling access to a product distribution P over $\{0,1\}^n$.

- Set $\tau \leftarrow \frac{1}{4}\epsilon^2$, $m \leftarrow \left\lceil \frac{2716\sqrt{n}}{\epsilon^2} \right\rceil$.

- Draw $M_1, \ldots, M_n \sim \mathrm{Poi}(m)$ independently, and let $M \leftarrow \max_{i \in [n]} M_i$.

If $M > 2m$ set $W = \tau m^2$

Else Take M samples $X^{(1)}, \ldots, X^{(M)}$ from P, and define

$$W = \sum_{i=1}^n \frac{(W_i - mq_i)^2 - W_i}{q_i(1-q_i)}$$

where $W_i \leftarrow \sum_{j=1}^{M_i} X_i^{(j)}$ for $i \in [n]$.

If $W \geq \tau m^2$ return reject.
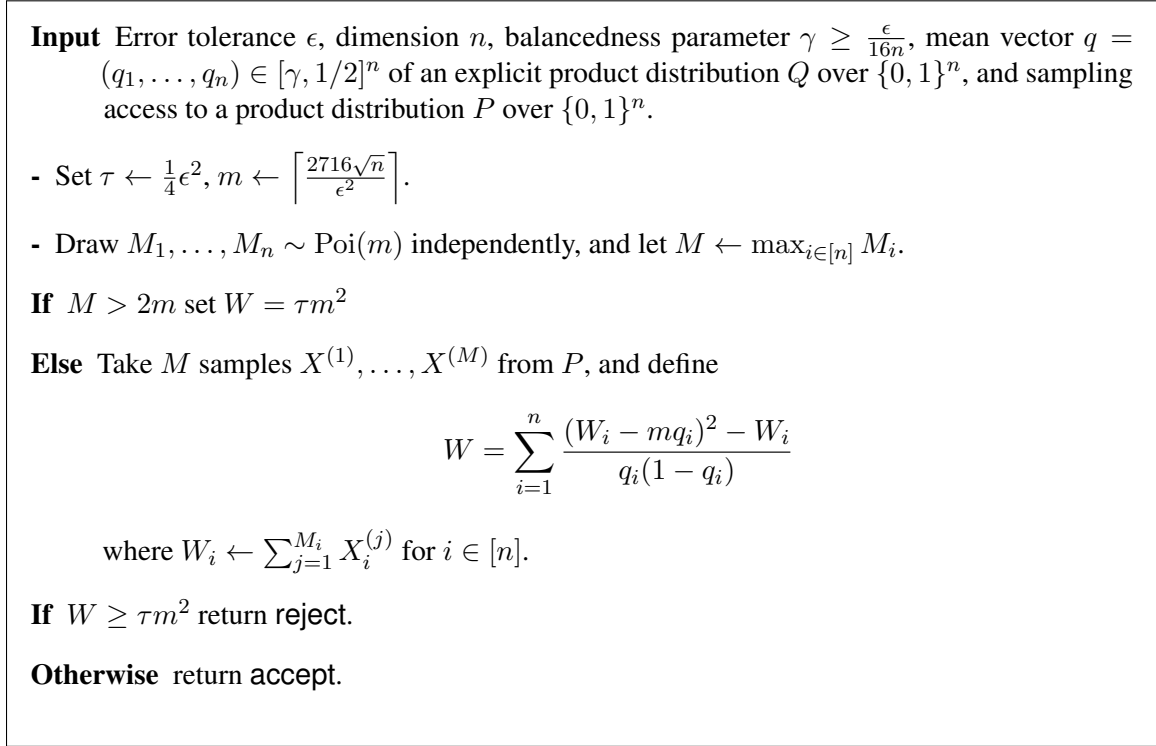
Otherwise return accept.

Figure 1: Identity testing: unknown product distribution P against given product distribution Q.

instead. By flipping the i-th bit of all samples we receive from P, it only remains to test identity of the resulting distribution P' to Q', as all distances are preserved.

Proof of Correctness. Let $m \geq 2716\frac{\sqrt{n}}{\epsilon^2}$, and let $M_1, \ldots, M_n$ be i.i.d. $\mathrm{Poi}(m)$ random variables. We set $M = \max_{i \in [n]} M_i$ and note that $M \leq 2m$ with probability $1 - e^{-\Omega(m)}$ (by a union bound). We condition hereafter on $M \leq 2m$ (our tester will reject otherwise) and take M samples $X^{(1)}, \ldots, X^{(M)}$ drawn from P. We define the following statistic:

$$W = \sum_{i=1}^n \frac{(W_i - mq_i)^2 - W_i}{q_i(1-q_i)} \,,$$

where we write $W_i \overset{\mathrm{def}}{=} \sum_{j=1}^{M_i} X_i^{(j)}$ for all $i \in [n]$. We note that the W_i's are independent, as P is a product distribution and the M_i's are independent. The pseudocode for our algorithm is given in Figure 1. Our identity tester is reminiscent of the "chi-squared type" testers that have been designed for the unstructured univariate discrete setting Chan et al. (2014c); Diakonikolas et al. (2015a); Acharya et al. (2015a).

We start with a simple formula for the expected value of our statistic:

Lemma 3 $\mathbb{E}[W] = m^2 \sum_{i=1}^n \frac{(p_i - q_i)^2}{q_i(1-q_i)}$.

Proof Since $W_i \sim \mathrm{Poi}(mp_i)$ for all i, we can write

$$\mathbb{E}\big[(W_i - mq_i)^2\big] = \mathbb{E}\big[W_i^2\big] - 2mq_i\mathbb{E}[W_i] + m^2q_i^2 = mp_i + m^2(p_i - q_i)^2 \,,$$

and therefore $\mathbb{E}[W] = \sum_{i=1}^{n} \frac{\mathbb{E}[(W_i - mq_i)^2] - \mathbb{E}[W_i]}{q_i(1-q_i)} = m^2 \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i(1-q_i)}$. ∎

As a corollary we obtain:

Claim 4 *If $P = Q$ then $\mathbb{E}[W] = 0$. Moreover, whenever $\|P - Q\|_1 > \epsilon$ we have $\mathbb{E}[W] > \frac{1}{2}m^2\epsilon^2$.*

Proof The first part is immediate from the expression of $\mathbb{E}[W]$. The second follows from Corollary 27, as $m^2\|P - Q\|_1^2 \le 2m^2 \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i(1-q_i)} = 2\mathbb{E}[W]$. ∎

We now proceed to bound from above the variance of our statistic. The completeness case is quite simple:

Claim 5 *If $P = Q$, then $\mathrm{Var}[W] \le 8m^2 n$.*

Proof Suppose that $P = Q$, i.e., $p = q$. By independence, we get $\mathrm{Var}[W] = \sum_{i=1}^{n} \frac{\mathrm{Var}[(W_i - mq_i)^2 - W_i]}{q_i^2(1-q_i)^2}$. Then, using the fact that $\mathbb{E}\big[(W_i - mq_i)^2 - W_i\big] = 0$, we obtain that

$$\mathrm{Var}[(W_i - mq_i)^2 - W_i] = \mathbb{E}\big[((W_i - mq_i)^2 - W_i)^2\big] = 2m^2 q_i^2,$$

where the last equality follows from standard computations involving the moments of a Poisson random variable. From there, recalling that $q_i \in (0, 1/2]$ for all $i \in [n]$, we obtain $\mathrm{Var}[W] = 2m^2 \sum_{i=1}^{n} \frac{1}{(1-q_i)^2} \le 8m^2 n$. ∎

For the soundness case, the following lemma bounds the variance of our statistic from above. We note that the upper bound depends on the balancedness parameter γ.

Lemma 6 *We have that $\mathrm{Var}[W] \le 16nm^2 + \left(\frac{32}{\gamma} + 16\sqrt{2n}m\right)\mathbb{E}[W] + \frac{32}{\sqrt{\gamma}}\mathbb{E}[W]^{3/2}$.*

Proof For general p, q, we have that

$$\mathrm{Var}[(W_i - mq_i)^2 - W_i] = \mathbb{E}\big[((W_i - mq_i)^2 - W_i)^2\big] - m^4(p_i - q_i)^4 = 2m^2 p_i^2 + 4m^3 p_i(p_i - q_i)^2 \,,$$

where as before the last equality follows from standard computations involving the moments of a Poisson random variable. This leads to

$$\mathrm{Var}[W] = 2m^2 \sum_{i=1}^{n} \frac{p_i^2}{q_i^2(1-q_i)^2} + 4m^3 \sum_{i=1}^{n} \frac{p_i(p_i - q_i)^2}{q_i^2(1-q_i)^2} \le 8m^2 \sum_{i=1}^{n} \frac{p_i^2}{q_i^2} + 16m^3 \sum_{i=1}^{n} \frac{p_i(p_i - q_i)^2}{q_i^2}.$$

We handle the two terms separately, in a fashion similar to (Acharya et al., 2015a, Lemma 2). For the first term, we can write:

$$
\begin{aligned}
\sum_{i=1}^{n} \frac{p_i^2}{q_i^2} &= \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i^2} + \sum_{i=1}^{n} \frac{2p_i q_i - q_i^2}{q_i^2} = \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i^2} + \sum_{i=1}^{n} \frac{2q_i(p_i - q_i) + q_i^2}{q_i^2} \\
&= n + \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i^2} + \sum_{i=1}^{n} \frac{2(p_i - q_i)}{q_i} = n + \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i^2} + \sum_{i=1}^{n} \frac{2(p_i - q_i)}{q_i} \\
&\underset{\text{(AM-GM)}}{\le} n + \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i^2} + \sum_{i=1}^{n} \left(1 + \frac{(p_i - q_i)^2}{q_i^2}\right) = 2n + 2\sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i^2} \\
&\le 2n + \frac{2}{\gamma} \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i} \le 2n + \frac{4}{m^2\gamma}\mathbb{E}[W] \,.
\end{aligned}
$$

We bound the second term from above as follows:

$$\sum_{i=1}^{n} \frac{p_i(p_i - q_i)^2}{q_i^2} \leq \sum_{i=1}^{n} \frac{p_i}{q_i} \cdot \frac{p_i(p_i - q_i)^2}{q_i} \leq \sqrt{\sum_{i=1}^{n} \frac{p_i^2}{q_i^2}} \sqrt{\sum_{i=1}^{n} \frac{(p_i - q_i)^4}{q_i^2}} \qquad \text{(Cauchy–Schwarz)}$$

$$\leq \left(\sqrt{2n} + \frac{2}{m\sqrt{\gamma}} \sqrt{\mathbb{E}[W]} \right) \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i} \qquad \text{(monotonicity of } \ell_p\text{-norms)}$$

$$= \frac{1}{m^2} \left(\sqrt{2n} + \frac{2}{m\sqrt{\gamma}} \sqrt{\mathbb{E}[W]} \right) \cdot \mathbb{E}[W] \,.$$

Overall, we obtain

$$\mathrm{Var}[W] \leq 16nm^2 + \frac{32}{\gamma} \mathbb{E}[W] + 16m \left(\sqrt{2n} + \frac{2}{m\sqrt{\gamma}} \sqrt{\mathbb{E}[W]} \right) \cdot \mathbb{E}[W]$$

$$= 16nm^2 + \left(\frac{32}{\gamma} + 16\sqrt{2n}m \right) \mathbb{E}[W] + \frac{32}{\sqrt{\gamma}} \mathbb{E}[W]^{3/2} \,.$$

∎

We are now ready to prove correctness.

Lemma 7 *Set $\tau \stackrel{\text{def}}{=} \frac{\epsilon^2}{4}$. Then we have the following:*

- *If $\|P - Q\|_1 = 0$, then $\Pr\big[W \geq \tau m^2 \big] \leq \frac{1}{3}$.*

- *If $\|P - Q\|_1 > \epsilon$, then $\Pr\big[W < \tau m^2 \big] \leq \frac{1}{3}$.*

Proof We start with the soundness case, i.e., assuming $\|P - Q\|_1 > \epsilon$. In this case, Claim 4 implies $\mathbb{E}[W] > 2\tau m^2$. Since $\gamma \geq \frac{\epsilon}{16n}$ and for $m \geq \frac{16}{\epsilon}\sqrt{2n}$, Lemma 6 implies that

$$\mathrm{Var}[W] \leq 16nm^2 + 32\sqrt{2n}m\mathbb{E}[W] + 32 \cdot 4\sqrt{\frac{n}{\epsilon}}\mathbb{E}[W]^{3/2} \,.$$

By Chebyshev's inequality, we have that

$$\Pr\big[W < \tau m^2 \big] \leq \Pr\left[\mathbb{E}[W] - W > \frac{1}{2}\mathbb{E}[W] \right] \leq \frac{4\,\mathrm{Var}[W]}{\mathbb{E}[W]^2} \leq \frac{64nm^2}{\mathbb{E}[W]^2} + \frac{128\sqrt{2n}m}{\mathbb{E}[W]} + \frac{512\sqrt{\frac{n}{\epsilon}}}{\mathbb{E}[W]^{1/2}}$$

$$\leq \frac{4 \cdot 64n}{m^2\epsilon^4} + \frac{2 \cdot 128\sqrt{2n}}{m\epsilon^2} + \frac{4\sqrt{2} \cdot 128\sqrt{n}}{m\epsilon^{3/2}} \leq 128 \left(\frac{2}{C^2} + \frac{5\sqrt{2}}{C} \right) \,,$$

which is at most $1/3$ as long as $C \geq 2716$, that is $m \geq 2716\frac{\sqrt{n}}{\epsilon^2}$.

Turning to the completeness, we suppose $\|P - Q\|_1 = 0$. Then, again by Chebyshev's inequality and Claim 5 we have that

$$\Pr\big[W \geq \tau m^2 \big] = \Pr\big[W \geq \mathbb{E}[W] + \tau m^2 \big] \leq \frac{\mathrm{Var}[W]}{\tau^2 m^4} \leq \frac{128n}{\epsilon^4 m^2} \,,$$

which is no more than $1/3$ as long as $m \geq 8\sqrt{6}\frac{\sqrt{n}}{\epsilon^2}$. ∎

∎

Remark 8 We observe that the aforementioned analysis – specifically Claim 4 and Lemma 7 – can be adapted to provide some tolerance guarantees in the completeness case, that is it implies a tester that distinguishes between $\|P - Q\|_1 \leq \epsilon'$ and $\|P - Q\|_1 > \epsilon$, where $\epsilon' = O(\epsilon^2)$. This extension, however, requires the assumption that Q be balanced: indeed, the exact dependence between ϵ' and ϵ^2 will depend on this balancedness parameter, leading to a tradeoff between tolerance and balancedness. Further, as shown in Section 5.3, this tradeoff is in fact necessary, as tolerant testing of arbitrary product distributions requires $\Omega(n/\log n)$ samples.

2.2. Sample Complexity Lower Bound for Identity Testing

In this section, we prove our matching information-theoretic lower bound for identity testing. In Theorem 9, we give a lower bound for uniformity testing of a product distribution, while Theorem 13 shows a quantitatively similar lower bound for identity testing against the product distribution with mean vector $q = (1/n, \ldots, 1/n)$. To establish these lower bounds, we use the information-theoretic technique from Diakonikolas and Kane (2016): Given a candidate hard instance, we proceed by bounding from above the mutual information between appropriate random variables. More specifically, we construct an appropriate family of hard instances (distributions) and show that a set of k samples taken from a distribution from the chosen family has small shared information with whether or not the distributions are the same.

Theorem 9 *There exists an absolute constant $\epsilon_0 > 0$ such that, for any $0 < \epsilon \leq \epsilon_0$, the following holds: Any algorithm that has sample access to an unknown product distribution P over $\{0,1\}^n$ and distinguishes between the cases that $P = U$ and $\|P - U\|_1 > \epsilon$ with probability $2/3$ requires $\Omega(\sqrt{n}/\epsilon^2)$ samples.*

Proof As previously mentioned, we first define two distributions over product distributions $\mathcal{Y}, \mathcal{N}$:

- $\mathcal{Y}$ is the distribution that puts probability mass 1 on the uniform distribution, $U = \mathrm{Bern}(1/2)^{\otimes n}$;

- $\mathcal{N}$ is the uniform distribution over the set

$$\left\{ \bigotimes_{j=1}^{n} \mathrm{Bern}\left(\frac{1}{2} + (-1)^{b_j} \frac{\epsilon}{\sqrt{n}} \right) \ : \ (b_1, \ldots, b_n) \in \{0,1\}^n \right\} .$$

Lemma 10 $\mathcal{N}$ *is supported on distributions that are $\Omega(\epsilon)$-far from U.*

Proof The proof, deferred to Appendix B, proceeds by considering directly the quantity $\|P - U\|_1$ in order to obtain a lower bound, where $P \overset{\mathrm{def}}{=} \bigotimes_{j=1}^{n} \mathrm{Bern}\left(\frac{1}{2} + \frac{\epsilon}{\sqrt{n}} \right)$; specifically, by focusing on the contribution to the distance from the points in the "middle layers" of the Boolean hypercube. (We note that an argument relying on the more convenient properties of the Hellinger distance with regard to product distributions, while much simpler, would only give a lower bound of $\Omega(\epsilon^2)$ – losing a quadratic factor.) ∎

We will make a further simplification, namely that instead of drawing k samples from $P = P_1 \otimes \cdots \otimes P_n$, the algorithm is given k_i samples from each P_i, where $k_1, \ldots, k_n$ are independent

$\mathrm{Poi}(k)$ random variables. This does not affect the lower bound, as this implies a lower bound on algorithms taking $k^* \overset{\text{def}}{=} \max(k_1, \ldots, k_n)$ samples from P (where the k_i's are as above), and $k^* \geq \frac{k}{2}$ with probability $1 - 2^{-\Omega(n)}$. We now consider the following process: letting $X \sim \mathrm{Bern}(1/2)$ be a uniformly random bit, we choose a distribution P over $\{0, 1\}^n$ by

- Drawing $P \sim \mathcal{Y}$ if $X = 0$, and;

- Drawing $P \sim \mathcal{N}$ if $X = 1$;

- Drawing $k_1, \ldots, k_n \sim \mathrm{Poi}(k)$, and returning k_1 samples from $P_1, \ldots, k_n$ samples from P_n.

For $i \in [n]$, we write N_i for the number of 1's among the k_i samples drawn from P_i, and let $N = (N_1, \ldots, N_n) \in \mathbb{N}^n$. We will invoke this standard fact, as stated in Diakonikolas and Kane (2016):

Fact 11 *Let X be a uniform random bit and Y a random variable taking value in some set $\mathcal{S}$. If there exists a function $f \colon \mathcal{S} \to \{0, 1\}$ such that $\Pr[\, f(Y) = X \,] \geq 0.51$, then $I\,(X; Y) = \Omega(1)$.*

Proof By Fano's inequality, letting $q = \Pr[\, f(Y) \neq X \,]$, we have $h(q) = h(q) + q \log(|\{0, 1\}| - 1) \geq H\,(X \mid Y)$. This implies $I\,(X; Y) = H\,(X) - H\,(X \mid Y) = 1 - H\,(X \mid Y) \geq 1 - h(q) \geq 1 - h(0.49) \geq 2 \cdot 10^{-4}$. $\blacksquare$

The next step is then to bound from above $I\,(X; N)$, in order to conclude that it will be $o(1)$ unless k is taken big enough and invoke Fact 11. By the foregoing discussion and the relaxation on the k_i's, we have that the conditioned on X the N_i are independent (with $N_i \sim \mathrm{Poi}(k p_i)$). Recall now that if X, Y_1, Y_2 are random variables such that Y_1 and Y_2 are independent conditioned on X, by the chain rule we have that

$$H\,((Y_1, Y_2) \mid X) = H\,(Y_1 \mid X) + H\,(Y_2 \mid X, Y_1) = H\,(Y_1 \mid X) + H\,(Y_2 \mid X) \ ,$$

where the second equality follows from conditional independence, and therefore

$$\begin{aligned} I\,(X; (Y_1, Y_2)) &= H\,((Y_1, Y_2)) - H\,((Y_1, Y_2) \mid X) \\ &= H\,(Y_1) + H\,(Y_1 \mid Y_2) - (H\,(Y_1 \mid X) + H\,(Y_2 \mid X)) \\ &\leq H\,(Y_1) + H\,(Y_1) - (H\,(Y_1 \mid X) + H\,(Y_2 \mid X)) \\ &= (H\,(Y_1) - H\,(Y_1 \mid X)) + (H\,(Y_2) - H\,(Y_2 \mid X)) \\ &= I\,(X; Y_1) + I\,(X; Y_2). \end{aligned}$$

This implies that

$$I\,(X; N) \leq \sum_{i=1}^{n} I\,(X; N_i) \ , \tag{1}$$

so that it suffices to bound each $I\,(X; N_i)$ separately.

Lemma 12 *Fix any $i \in [n]$, and let X, N_i be as above. Then $I\,(X; N_i) = O(k^2 \epsilon^4 / n^2)$.*

Proof The proof of this rather technical result can be found in Appendix B, and broadly proceeds as follows. The first step is to upper bound $I(X; N_i)$ by a more manageable quantity, $\sum_{a=0}^{\infty} \Pr[N_i = a]\left(1 - \frac{\Pr[N_i=a \mid X=1]}{\Pr[N_i=a \mid X=0]}\right)^2$. After this, giving an upper bound on each summand can be done by performing a Taylor series expansion (in $\epsilon/\sqrt{n}$), relying on the expression of the moment-generating function of the Poisson distribution to obtain cancellations of many low-order terms. ∎

This lemma, along with Eq. (1), gives the desired result, that is

$$I(X; N) \leq \sum_{i=1}^{n} O\left(\frac{\epsilon^4 k^2}{n^2}\right) = O\left(\frac{\epsilon^4 k^2}{n}\right), \tag{2}$$

which is $o(1)$ unless $k = \Omega(\sqrt{n}/\epsilon^2)$. ∎

Theorem 13 *There exists an absolute constant $\epsilon_0 > 0$ such that, for any $\epsilon \in (0, \epsilon_0)$, distinguishing $P = P^*$ and $\|P - P^*\|_1 > \epsilon$ with probability $2/3$ requires $\Omega(\sqrt{n}/\epsilon^2)$ samples, where $P^* \stackrel{\text{def}}{=} \mathrm{Bern}(1/n)^{\otimes n}$.*

Proof The proof will follow the same outline as that of Theorem 9 first defining two distributions over product distributions $\mathcal{Y}, \mathcal{N}$:

- $\mathcal{Y}$ is the distribution that puts probability mass 1 on P^*;

- $\mathcal{N}$ is the uniform distribution over the set

$$\left\{ \bigotimes_{j=1}^{n} \mathrm{Bern}\left(\frac{1}{n}\left(1 + (-1)^{b_j}\epsilon\right)\right) \; : \; (b_1, \ldots, b_n) \in \{0, 1\}^n \right\}.$$

Lemma 14 *With probability $1 - 2^{-\Omega(n)}$, $\mathcal{N}$ is supported on distributions $\Omega(\epsilon)$-far from P^*.*

Proof [Proof of Lemma 14] As for Lemma 10, using Hellinger distance as a proxy would only result in an $\Omega(\epsilon^2)$ lower bound on the distance, so we will compute it explicitly instead. The proof can be found in Appendix B. ∎

The only ingredient missing to conclude the proof is the analogue of Lemma 12:

Lemma 15 *Suppose $\frac{k\epsilon^2}{n} \leq 1$. Fix any $i \in [n]$, and let X, N_i be as above. Then $I(X; N_i) = O(k^2\epsilon^4/n^2)$.*

Proof The proof is similar as that of (Diakonikolas and Kane, 2016, Lemma 3.3), replacing (their) mn by (our) n. For completeness, we provide an alternative proof in Appendix D. ∎

3. Our Results and Techniques

The structure of this section is as follows: In Section 3.1, we provide the statements of our main results in tandem with a brief explanation of their context and the relations between them. techniques.

3.1. Main Results

The focus of this paper is on the properties of *identity testing* and *closeness testing* of Bayes nets. We give the first non-trivial efficient testing algorithms and matching information-theoretic lower bounds for these problems. For a wide range of parameter settings, our algorithms achieve sub-learning sample complexity and are sample-optimal (up to constant factors).

For concreteness, we consider Bayes nets over Bernoulli random variables. We note that our upper bounds straightforwardly extend to general alphabets with a polynomial dependence on the alphabet size (see Remark 55). Let $\mathcal{BN}_{n,d}$ denote the family of Bernoulli Bayes nets on n variables such that the corresponding DAG has maximum in-degree at most d. For most of our results, we will think of the dimension n as being large and the maximum degree d as being comparably small (say, bounded from above by a constant or at most logarithmic in n).

For the inference problems of learning and testing Bayes nets, there are two versions of the problem: The first version corresponds to the setting where the structure of the graph is fixed (and known a priori to the algorithm). In the second version, both the graph and the parameters are unknown to the algorithm. We note that both versions of the problem are interesting, based on the application. The *unknown structure* setting is clearly at least as hard, and typically includes an algorithm for the *fixed structure* case plus additional algorithmic ingredients.

Before we give the statements of our main testing results, we record a nearly tight bound on the sample complexity of learning $\mathcal{BN}_{n,d}$. This bound will be used as a baseline to compare against our efficient testers:

Fact 16 *The sample complexity of learning $\mathcal{BN}_{n,d}$, within total variation distance ϵ, with confidence probability $9/10$, is: (i) $\widetilde{\Theta}(2^d \cdot n/\epsilon^2)$, for all $d \leq n/2$, in the fixed structure setting, and (ii) $\widetilde{\Theta}(2^d \cdot n/\epsilon^2)$ in the unknown structure setting.*

We give a proof of this fact in Appendix A. Fact 16 characterizes the sample complexity of learning Bayes nets (up to logarithmic factors). We remark that our information-theoretic upper bound for the fixed structure case also yields a simple computationally efficient algorithm. The unknown structure regime is much more challenging computationally. For this setting, we provide a nearly tight information-theoretic upper bound that is non-constructive. (The corresponding algorithm runs in exponential time.) In fact, we note that no sample-optimal computationally efficient algorithm is known for unknown structure Bayes nets.

Our first main result concerns the fixed structure regime. For technical reasons, we focus on Bayes nets that satisfy a natural balancedness condition. Roughly speaking, our balancedness condition ensures that the conditional probabilities are bounded away from 0 and 1, and that each parental configuration happens with some minimum probability. Formally, we have:

Definition 17 *A Bayes net P over $\{0,1\}^n$ with structure $\mathcal{S}$ is called (c, C)-balanced if, for all k, we have that (i) $p_k \in [c, 1-c]$, and (ii) $\Pr_P[\Pi_k] \geq C$.*

Under a mild condition on the balancedness, we give sample-optimal and computationally efficient algorithms for testing identity and closeness of Bayes nets. Specifically, for the problem of identity testing against an explicit distribution, we require that the explicit distribution be balanced (no assumption is needed for the unknown Bayes net). For the problem of closeness testing, we require that one of the two unknown distributions be balanced. We are now ready to state our first main theorem:

Theorem 18 (Testing Identity and Closeness of Fixed–Structure Bayes Nets) *For testing identity and closeness of fixed structure Bayes nets P, Q with n nodes and maximum in-degree d, there is an efficient algorithm that uses $O\left(2^{d/2}\sqrt{n}/\epsilon^2\right)$ samples and, assuming that one of P, Q is (c, C)-balanced with $c = \tilde{\Omega}\left(1/\sqrt{n}\right)$ and $C = \tilde{\Omega}\left(d\epsilon^2/\sqrt{n}\right)$, correctly distinguishes between the cases that $P = Q$ versus $\|P - Q\|_1 > \epsilon$, with probability at least $2/3$. Moreover, this sample size is information-theoretically optimal, up to constant factors, for all $d < n/2$, even for the case of uniformity testing.*

The conceptual message of Theorem 18 is that, for the case of fixed structure, testing is information-theoretically easier than learning. Specifically, our result establishes a quadratic gap between learning and identity testing, reminiscent of the analogous gap in the setting of unstructured discrete distributions. We remark here that the information-theoretic lower bounds of Fact 16 (i) hold even for Bayes nets with constant balancedness.

We now turn our attention to the case of unknown structure. Motivated by Theorem 18, it would be tempting to conjecture that one can obtain testers with sub-learning sample complexity in this setting as well. Our first main result for unknown structure testing is an information-theoretic lower bound, showing that this is not the case. Specifically, even for the most basic case of tree-structured Bays Nets ($d = 1$) with unknown structure, uniformity testing requires $\Omega(n/\epsilon^2)$ samples. It should be noted that our lower bound applies even for Bayes nets with *constant* balancedness. Formally, we have:

Theorem 19 (Sample Lower Bound for Uniformity Testing of Unknown Tree-Structured BN) *Any algorithm that, given sample access to a balanced tree-structured Bayes net P over $\{0, 1\}^n$, distinguishes between the cases $P = U$ and $\|P - U\|_1 > \epsilon$ (where U denotes the uniform distribution over $\{0, 1\}^n$), with probability $2/3$, requires $\Omega(n/\epsilon^2)$ samples from P.*

At the conceptual level, our above lower bound implies that in the unknown topology case – even for the simplest non-trivial case of degree-1 Bayes nets – identity testing is information-theoretically essentially as hard as learning. That is, in some cases, no tester with sub-learning sample complexity exists. We view this fact as an interesting phenomenon that is absent from the previously studied setting of testing unstructured discrete distributions.

Theorem 19 shows that testing Bayes nets can be as hard as learning. However, it is still possible that testing is easier than learning in *most* natural situations. For the sake of intuition, let us examine our aforementioned lower bound more carefully. We note that the difficulty of the problem originates from the fact that the explicit distribution is the uniform distribution, which can be thought of as having any of a large number of possible structures. We claim that this impediment can be circumvented if the explicit distribution satisfies some non-degeneracy conditions. Intuitively, we want these conditions to ensure *robust identifiability* of the structure: that is, that any (unknown) Bayes net sufficiently close to a non-degenerate Bayes net Q must also share the same structure.

For tree structures, there is a very simple non-degeneracy condition. Namely, that for each node, the two conditional probabilities for that node (depending on the value of its parent) are non-trivially far from each other. For Bayes nets of degree more than one, our non-degeneracy condition is somewhat more complicated to state, but the intuition is still simple: By definition, non-equivalent Bayesian network structures satisfy different conditional independence constraints. Our non-degeneracy condition rules out some of these possible new conditional independence constraints, as *far* from being satisfied by the non-degenerate Bayesian network. Let $\gamma > 0$ be a parameter

quantifying non-degeneracy. Under our non-degeneracy condition, we can design a *structure tester* with the following performance guarantee:

Theorem 20 (Structure Testing for Non-Degenerate Bayes Nets) *Let $\mathcal{S}$ be a structure of degree at most d and P be a degree at most d Bayes net over $\{0,1\}^n$ with structure $\mathcal{S}'$ whose underlying undirected graph has no more edges than $\mathcal{S}$. There is an algorithm that uses $O\left((2^d + d\log n)/\gamma^2\right)$ samples from P, runs in time $O\left(n^{d+3}/\gamma^2\right)$, and distinguishes between the following two cases with probability at least $2/3$: (i) P can be expressed as a degree-d Bayes net with structure $\mathcal{S}$ that is γ-non-degenerate; or (ii) P cannot be expressed as a Bayes net with structure $\mathcal{S}$.*

By invoking the structure test of the above theorem, we can reduce the identity testing with unknown structure to the case of known structure, obtaining the following:

Theorem 21 (Testing Identity of Non-Degenerate Unknown Structure Bayes Nets) *There exists an algorithm with the following guarantees. Given the description of a degree-d Bayes net Q over $\{0,1\}^n$, which is (c, C) balanced and γ-non-degenerate for $c = \tilde{\Omega}\left(1/\sqrt{n}\right)$ and $C = \tilde{\Omega}\left(d\epsilon^2/\sqrt{n}\right)$, $\epsilon > 0$, and sample access to a distribution P, promised to be a degree-d Bayes net with no more edges than Q, the algorithm takes $O\left(2^{d/2}\sqrt{n}/\epsilon^2 + (2^d + d\log n)/\gamma^2\right)$ samples from P, runs in time $O\left(n\right)^{d+3}(1/\gamma^2 + 1/\epsilon^2)$, and distinguishes with probability at least $2/3$ between (i) $P = Q$ and (ii) $\|P - Q\|_1 > \epsilon$.*

We remark that we can obtain an analogous result for the problem of testing closeness. See Section 9.

We have shown that, without any assumptions, testing is almost as hard as learning for the case of trees. An interesting question is whether this holds for high degrees as well. We show that for the case of high degree sub-learning sample complexity is possible. We give an identity testing algorithm for degree-d Bayes nets with unknown structure, *without* balancedness or degeneracy assumptions. While the dependence on the number of nodes n of this tester is suboptimal, it does essentially achieve the "right" dependence on the degree d, that is $2^{d/2}$:

Theorem 22 (Sample Complexity Upper Bound of Identity Testing) *Given the description of a degree-d Bayes net Q over $\{0,1\}^n$, $\epsilon > 0$, and sample access to a degree-d Bayes net P, we can distinguish between the cases that $P = Q$ and $\|P - Q\|_1 > \epsilon$, with probability at least $2/3$, using $2^{d/2}\mathrm{poly}(n, 1/\epsilon)$ samples from P.*

(See Theorem 80 for a more detailed statement handling closeness testing as well.) The message of this result is that when the degree d increases, specifically for $d = \Omega(\log n)$, the sample complexity of testing becomes lower than the sample complexity of learning. We also show an analogue of Theorem 22 for closeness testing of two unknown Bayes nets, under the additional assumption that we know the topological ordering of the unknown DAGs.

3.2. Organization

This paper is organized as follows: In Section 4, we give the necessary definitions and tools we will require. Section 5 gives our matching upper and lower bounds for closeness testing of product distributions (recall that Section 2 was concerned with the simpler question of *identity* testing of product distributions). Section 6 then provides an overview of what is required to generalize these

techniques from product distributions to Bayes nets, discussing at a high-level the following sections. In Section 7 we study the identity testing for Bayes nets with known structure: We give an identity tester that works under a mild balancedness condition on the explicit Bayes net distribution, and also show that the sample complexity of our algorithm is optimal, up to constant factors. In Section 8, we study the identity testing for unknown structure Bayes nets: We start by proving a sample complexity lower bound showing that, for the unknown structure regime, uniformity testing is information-theoretically as hard as learning – even for the case of trees. We then show that this lower bound can be circumvented under a natural non-degeneracy condition on the explicit Bayes net distribution. Specifically, we give an identity tester with sub-learning sample complexity for all low-degree non-degenerate Bayes nets. Our identity tester for unknown structure non-degenerate Bayes nets relies on a novel structure tester that may be of interest in its own right. Section 9 studies the corresponding closeness testing problems for both known and unknown structure Bayes nets. Finally, in Section 10 we consider the case of high-degree Bayes nets and obtain testers for identity and closeness of unknown-structure Bayes nets. Our testers in this section have optimal (and sub-learning) sample complexity as a function of the maximum in-degree d and polynomial dependence in the dimension n.

3.3. Concurrent and Independent Work

Contemporaneous work by Daskalakis and Pan (2016) studies the identity testing problem for Bayes nets with *the same known* graph structure. Using different arguments, they obtain a tester with sample complexity $\tilde{O}(2^{(3/4)d} \cdot n/\epsilon^2)$ and running time $O_\epsilon(n^{d+1})$ for this problem. This sample bound is comparable to that of our Theorem 80 (that works without assumptions on the parameters), having the right dependence on $n, 1/\epsilon$ (as follows from our Theorem 19 and Fact 16) and a sub-optimal dependence on the degree d. As previously mentioned, a sample complexity of $\Omega(n/\epsilon^2)$ is relevant for high-degree Bayes nets. For the case of low-degree (which is the main focus of our paper), one can straightforwardly obtain the same sample bound just by learning the distribution (Fact 16). Daskalakis and Pan (2016) also obtain an $O(n^{1/2}/\epsilon^2)$ upper bound for testing identity against a known product, matching our Theorem 2. (This sample bound is optimal by our Theorem 9.)

4. Preliminaries

In this section, we record the basic definitions and technical tools we shall use throughout this paper. **Basic Notation and Definitions.** The L_1-distance between two discrete probability distributions P, Q supported on a set A is defined as $\|P - Q\|_1 = \sum_{x \in A} |P(x) - Q(x)|$. Our arguments will make essential use of related distance measures, specifically the KL-divergence, defined as $D(P\|Q) = \sum_{x \in A} P(x) \log \frac{P(x)}{Q(x)}$, and the Hellinger distance, defined as $d_{\mathrm{H}}(P, Q) = (1/\sqrt{2}) \cdot \sqrt{\sum_{x \in A} (\sqrt{P(x)} - \sqrt{Q(x)})^2}$.

We write $\log$ and $\ln$ for the binary and natural logarithms, respectively, and by $H(X)$ the (Shannon) entropy of a discrete random variable X (as well as, by extension, $H(P)$ for the entropy of a discrete distribution P). We denote by $I(X; Y)$ the mutual information between two random variables X and Y, defined as $I(X; Y) = \sum_{x,y} \Pr[(X, Y) = (x, y)] \log \frac{\Pr[(X,Y)=(x,y)]}{\Pr[X=x]\Pr[Y=y]}$. For a probability distribution P, we write $X \sim P$ to indicate that X is distributed according to P. For probability distributions P, Q, we will use $P \otimes Q$ to denote the product distribution with marginals P and Q.

Identity and Closeness Testing.   We now formally define the testing problems that we study.

Definition 23 (Identity testing) *An* identity testing algorithm of distributions belonging to a class $\mathcal{C}$ *is a randomized algorithm which satisfies the following. Given a parameter $0 < \epsilon < 1$ and the explicit description of a reference distribution $Q \in \mathcal{C}$, as well as access to independent samples from an unknown distribution $P \in \mathcal{C}$, the algorithm outputs either* accept *or* reject *such that the following holds:*

- *(Completeness) if $P = Q$, then the algorithm outputs* accept *with probability at least $2/3$;*

- *(Soundness) if $\|P - Q\|_1 \geq \epsilon$, then the algorithm outputs* reject *with probability at least $2/3$.*

Note that by the above definition the algorithm is allowed to answer arbitrarily if neither the completeness nor the soundness cases hold. The closeness testing problem is similar, except that now both P, Q are unknown and are only available through independent samples.

Definition 24 (Closeness testing) *A* closeness testing algorithm of distributions belonging to a class $\mathcal{C}$ *is a randomized algorithm which satisfies the following. Given a parameter $0 < \epsilon < 1$ and access to independent samples from two unknown distributions $P, Q \in \mathcal{C}$, the algorithm outputs either* accept *or* reject *such that the following holds:*

- *(Completeness) if $P = Q$, then the algorithm outputs* accept *with probability at least $2/3$;*

- *(Soundness) if $\|P - Q\|_1 \geq \epsilon$, then the algorithm outputs* reject *with probability at least $2/3$.*

Finally, we also consider a third related question, that of *structure testing*:

Definition 25 (Structure testing) *Let $\mathcal{C}$ be a family of Bayes nets. A* structure testing algorithm of Bayes nets belonging to $\mathcal{C}$ *is a randomized algorithm which satisfies the following. Given a parameter $0 < \epsilon < 1$ and the explicit description of a DAG $\mathcal{S}$, as well as access to independent samples from an unknown $P \in \mathcal{C}$, the algorithm outputs either* accept *or* reject *such that the following holds:*

- *(Completeness) if P can be expressed as a Bayes net with structure $\mathcal{S}$, then the algorithm outputs* accept *with probability at least $2/3$;*

- *(Soundness) if $\|P - Q\|_1 > \epsilon$ for every $Q \in \mathcal{C}$ with structure $\mathcal{S}$, then the algorithm outputs* reject *with probability at least $2/3$.*

In all cases the two relevant complexity measures are the *sample complexity, i.e., the number of samples drawn by the algorithm, and the* time complexity of the algorithm. The golden standard is to achieve sample complexity that is information-theoretically optimal and time-complexity linear in the sample complexity.

In this work, the family $\mathcal{C}$ will correspond to the family of Bayes nets over $\{0, 1\}^n$, where we will impose an upper bound d on the maximum in-degree of each node. For $d = 0$, i.e., when the underlying graph has no edges, we obtain the family of product distributions over $\{0, 1\}^n$.

Relations between Distances. We will require a number of inequalities relating the L_1-distance, the KL-divergence, and the Hellinger distance between distributions. We state a number of inequalities relating these quantities that we will use extensively in our arguments. The simple proofs are deferred to Appendix C.

Recall that a binary product distribution is a distribution over $\{0, 1\}^n$ whose coordinates are independent; and that such a distribution is determined by its mean vector. We have the following:

Lemma 26 *Let P, Q be binary product distributions with mean vectors $p, q \in (0, 1)^n$. We have that*

$$2 \sum_{i=1}^{n} (p_i - q_i)^2 \leq \mathrm{D}(P\|Q) \leq \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i(1 - q_i)} \ . \tag{3}$$

In particular, if there exists $\alpha > 0$ such that $q \in [\alpha, 1 - \alpha]^n$, we obtain

$$2\|p - q\|_2^2 \leq \mathrm{D}(P\|Q) \leq \frac{1}{\alpha(1 - \alpha)}\|p - q\|_2^2 \ . \tag{4}$$

Recall that for any pair of distributions P, Q, Pinsker's inequality states that $\|P - Q\|_1^2 \leq 2\,\mathrm{D}(P\|Q)$. This directly implies the following:

Corollary 27 *Let P, Q be binary product distributions with mean vectors $p, q \in (0, 1)^n$. We have that*

$$\|P - Q\|_1^2 \leq 2 \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i(1 - q_i)} \ .$$

The following lemma states an incomparable and symmetric upper bound on the L_1-distance, as well as a lower bound.

Lemma 28 *Let P, Q be binary product distributions with mean vectors $p, q \in (0, 1)^n$. Then it holds that*

$$\min\left(c, \|p - q\|_2^4\right) \leq \|P - Q\|_1^2 \leq 8 \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{(p_i + q_i)(2 - p_i - q_i)} \ .$$

for some absolute constant $c \in (0, 1)$. (Moreover, one can take $c = 4(1 - e^{-3/2}) \simeq 3.11$.)

While the above is specific to product distributions, we will require analogous inequalities for Bayes nets. We start with the following simple lemma:

Lemma 29 *Let P and Q be Bayes nets with the same dependency graph. In terms of the conditional probability tables p and q of P and Q, we have:*

$$2 \sum_{k=1}^{m} \Pr_P\left[\Pi_k\right] (p_k - q_k)^2 \leq \mathrm{D}(P\|Q) \leq \sum_{k=1}^{m} \Pr_P\left[\Pi_k\right] \frac{(p_k - q_k)^2}{q_k(1 - q_k)} \ .$$

Finally, we state an alternative bound, expressed with respect to the Hellinger distance between two Bayes nets:

Lemma 30 ((Diakonikolas et al., 2016c, Lemma 4)) *Let P and Q be Bayes nets with the same dependency graph. In terms of the conditional probability tables p and q of P and Q, we have:*

$$\mathrm{d}_{\mathrm{H}}(P, Q)^2 \leq 2 \sum_{k=1}^{m} \sqrt{\Pr_P\left[\Pi_k\right] \Pr_Q\left[\Pi_k\right] \frac{(p_k - q_k)^2}{(p_k + q_k)(2 - p_k - q_k)}} \ .$$

5. Testing Closeness of Product Distributions

Our first set of results involves sample-optimal testers and matching information-theoretic lower bounds for testing closeness of product distributions over $\{0, 1\}^n$ (recall that in Section 2, we settled the related question of identity testing of such product distributions). Specifically, in Section 5.1, we give a closeness testing algorithm for n-dimensional binary product distributions with sample complexity $O(\sqrt{n}/\epsilon^2)$; then, we show in Section 5.2 that this sample bound is information-theoretically optimal.

5.1. Closeness Testing Algorithm

Compared to identity, testing closeness between two unknown product distributions is somewhat more complicated and requires additional ideas. As is the case when comparing unknown discrete distributions on $[n]$, we have the difficulty that we do not know how to scale our approximations to the $(p_i - q_i)^2$ terms. We are forced to end up rescaling using the total number of samples drawn with $x_i = 1$ as a proxy for $1/(q_i)$. This leaves us with a statistic reminiscent of that used in Chan et al. (2014c), which can be shown to work with a related but more subtle analysis. First, in our setting, it is no longer the case that the sum of the q_i's is $O(1)$, and this ends up affecting the analysis, making our sample complexity depend on $n^{3/4}$ instead of $n^{2/3}$ as in the unstructured case. Second, to obtain the optimal sample complexity as a function of both n and ϵ, we need to partition the coordinates into two groups based on the value of their marginals and apply a different statistic to each group. It turns out that the sample complexity of our closeness testing algorithm is $O(\max(n^{1/2}/\epsilon^2, n^{3/4}/\epsilon))$:

Theorem 31 *There exists an efficient algorithm which, given sample access to two unknown product distributions P, Q over $\{0, 1\}^n$, has the following guarantees. For any $\epsilon \in (0, 1)$, the algorithm takes $O\left(\max\left(\sqrt{n}/\epsilon^2, n^{3/4}/\epsilon\right)\right)$ samples from P and Q, and distinguishes with probability $2/3$ between (i) $\|P - Q\|_1 = 0$ and (ii) $\|P - Q\|_1 > \epsilon$.*

The rest of this section is devoted to the proof of the above theorem.

Proof Let P, Q be two product distributions on $\{0, 1\}^n$ with mean vectors $p, q \in [0, 1]^n$. For $S \subseteq [n]$, we denote by P_S and Q_S the product distributions on $\{0, 1\}^{|S|}$ obtained by restricting P and Q to the coordinates in S. Similarly, we write $p_S, q_S \in [0, 1]^{|S|}$ for the vectors obtained by restricting p, q to the coordinates in S, so that P_S has mean vector p_S.

High-level Idea. The basic idea of the algorithm is to divide the coordinates in two bins U, V: one containing the indices where both distributions have marginals very close to 0 (specifically, at most $1/m$, where m is our eventual sample complexity), and one containing the remaining indices, on which at least one of the two distributions is roughly balanced. Since P and Q can only be far from each other if at least one of $\|P_U - Q_U\|_1$, $\|P_V - Q_V\|_1$ is big, we will test separately each case. Specifically, we will apply two different testers: one "χ^2-based tester" (with sample complexity $O\left(\sqrt{n}/\epsilon^2\right)$) to the "heavy bin" U – which relies on the fact that the marginals of P, Q on U are balanced by construction – and one "ℓ_2-tester" (with sample complexity $O\left(n^{3/4}/\epsilon\right)$) to the "light bin" V – relying on the fact that $\|p_V\|_2$, $\|q_V\|_2$ are small. The pseudocode of our algorithm is given in Figure 2.

Sample Complexity. Hereafter, we let

$$m \stackrel{\text{def}}{=} C \max \left(\frac{\sqrt{n}}{\epsilon^2}, \frac{n^{3/4}}{\epsilon} \right) ,$$

for some absolute constant $C > 0$ to be determined in the course of the analysis. We let $M_1, \ldots, M_n$ and $M'_1, \ldots, M'_n$ be i.i.d. $\mathrm{Poi}(m)$ random variables, set $M = \max_{i \in [n]} M_i$, $M' = \max_{i \in [n]} M'_i$; and note that $M, M' \leq 2m$ with probability $1 - e^{-\Omega(m)}$ (by a union bound). We will condition hereafter on the event that $M, M' \leq 2m$ and our tester will reject otherwise.

Without loss of generality, as in the previous sections, we will assume that $\frac{\epsilon}{16n} \leq p_i, q_i \leq \frac{3}{4}$ for every $i \in [n]$. Indeed, this can be ensured by the simple preprocessing step below.

Preprocessing. Using $O(\log n)$ samples from P and Q, we can ensure without loss of generality that all p_i, q_i are at most $3/4$ (with probability $9/10$). Namely, we estimate every p_i, q_i to an additive $1/64$, and proceed as follows:

- If the estimate of q_i is not within an additive $\pm\frac{1}{32}$ of that of p_i, we output reject and stop;

- If the estimate of p_i is more than $43/64$, mark i as "swapped" and replace X_i by $1 - X_i$ (for P) and Y_i by $1 - Y_i$ (for Q) in all future samples.

Assuming correctness of the estimates (which holds with probability at least $9/10$), if we pass this step then $|p_i - q_i| < \frac{1}{16}$ for all i. Moreover, if i was not swapped, then it means that we had $p_i \leq 43/64 + 1/64 < 3/4$, and therefore $q_i < 43/64 + 1/64 + 1/16 = 3/4$. Now, if we had $q_i > 3/4$, then $p_i > 3/4 - 1/16$ and the estimate of p_i would be more than $3/4 - 1/16 - 1/64 = 43/64$.

Proof of Correctness. For m as above, define $U, V \subseteq [n]$ by $V \stackrel{\text{def}}{=} \left\{ i \in [n] : \max(p_i, q_i) < \frac{1}{m} \right\}$ and $U \stackrel{\text{def}}{=} [n] \setminus V$. We start with the following simple claim:

Claim 32 *Assume $\|P - Q\|_1 > \epsilon$. Then, at least one of the following must hold: (i) $\|p_V - q_V\|_2^2 > \frac{\epsilon^2}{16n}$, or (ii) $\sum_{i \in U} \frac{(p_i - q_i)^2}{p_i + q_i} > \frac{\epsilon^2}{64}$.*

Proof Since $\epsilon < \|P - Q\|_1 \leq \|P_U - Q_U\|_1 + \|P_V - Q_V\|_1$, at least one of the two terms in the RHS must exceed $\frac{\epsilon}{2}$. We now recall that, by Lemma 28, it holds that $\|P_U - Q_U\|_1^2 \leq 8 \sum_{i \in U} \frac{(p_i - q_i)^2}{(p_i + q_i)(2 - p_i - q_i)}$ and from the further assumption that $p_i, q_i \leq \frac{3}{4}$ that $\|P_U - Q_U\|_1^2 \leq 16 \sum_{i \in U} \frac{(p_i - q_i)^2}{p_i + q_i}$.

Using subadditivity and the Cauchy–Schwartz inequality, we also have

$$\|P_V - Q_V\|_1 \leq \sum_{i \in V} \|P_i - Q_i\|_1 = 2 \sum_{i \in V} |p_i - q_i| = 2\|p_V - q_V\|_1$$
$$\leq 2\sqrt{|V|}\|p_V - q_V\|_2 \leq 2\sqrt{n}\|p_V - q_V\|_2 ,$$

from where we derive that $\|p_V - q_V\|_2^2 \geq \frac{1}{4n}\|P_V - Q_V\|_1^2$. This completes the proof. ∎

We now define $U', V' \subseteq [n]$ (our "proxies" for U, V) as follows: Taking m samples from both P and Q, we let V' be the set of indices which were never seen set to one in any sample, and U' be its complement. We have the following:
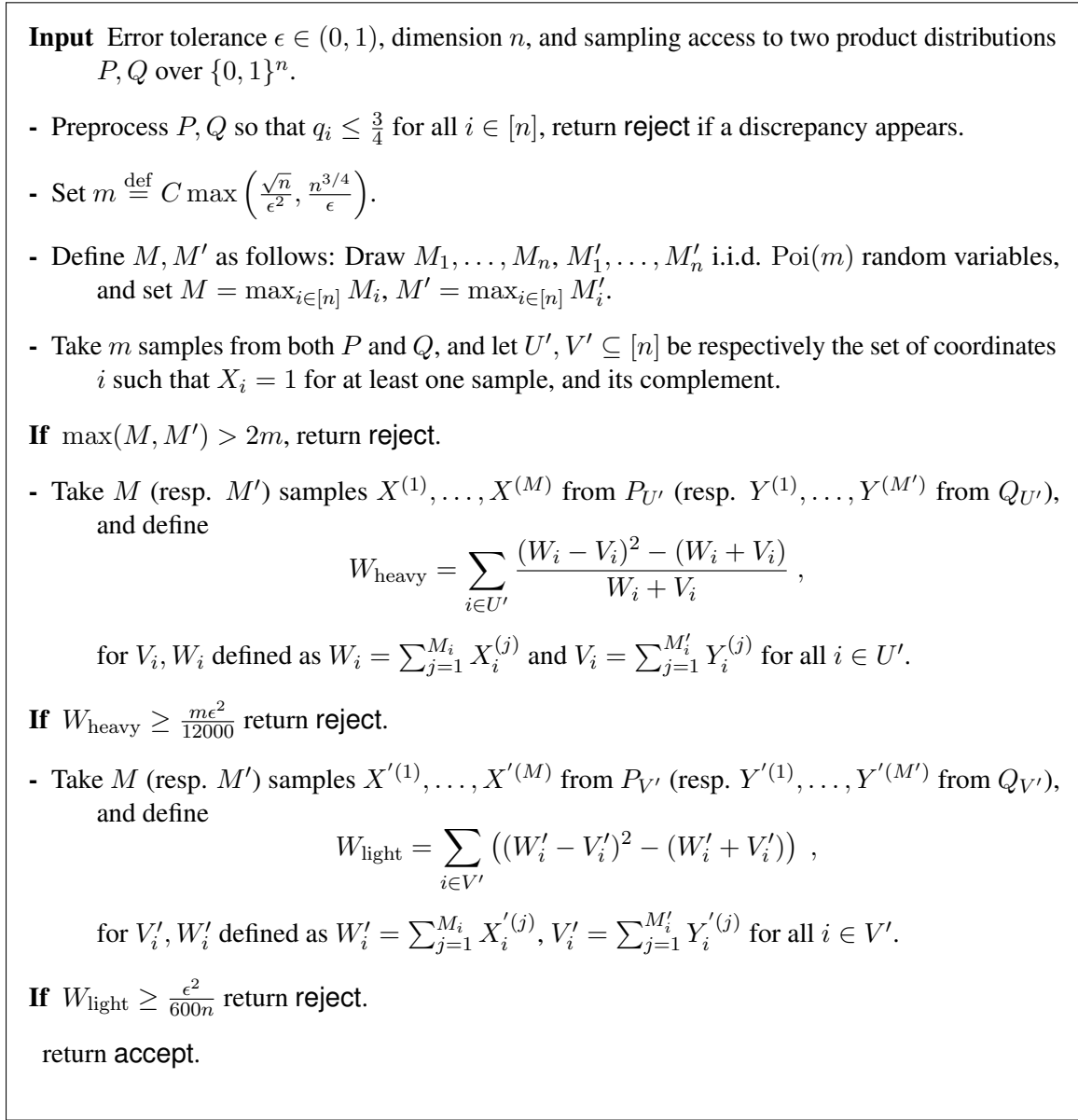

Input Error tolerance $\epsilon \in (0, 1)$, dimension n, and sampling access to two product distributions P, Q over $\{0, 1\}^n$.

- Preprocess P, Q so that $q_i \leq \frac{3}{4}$ for all $i \in [n]$, return reject if a discrepancy appears.

- Set $m \stackrel{\text{def}}{=} C \max \left(\frac{\sqrt{n}}{\epsilon^2}, \frac{n^{3/4}}{\epsilon} \right)$.

- Define M, M' as follows: Draw $M_1, \ldots, M_n, M'_1, \ldots, M'_n$ i.i.d. $\mathrm{Poi}(m)$ random variables, and set $M = \max_{i \in [n]} M_i$, $M' = \max_{i \in [n]} M'_i$.

- Take m samples from both P and Q, and let $U', V' \subseteq [n]$ be respectively the set of coordinates i such that $X_i = 1$ for at least one sample, and its complement.

If $\max(M, M') > 2m$, return reject.

- Take M (resp. M') samples $X^{(1)}, \ldots, X^{(M)}$ from $P_{U'}$ (resp. $Y^{(1)}, \ldots, Y^{(M')}$ from $Q_{U'}$), and define
$$W_{\text{heavy}} = \sum_{i \in U'} \frac{(W_i - V_i)^2 - (W_i + V_i)}{W_i + V_i} \ ,$$
for V_i, W_i defined as $W_i = \sum_{j=1}^{M_i} X_i^{(j)}$ and $V_i = \sum_{j=1}^{M'_i} Y_i^{(j)}$ for all $i \in U'$.

If $W_{\text{heavy}} \geq \frac{m\epsilon^2}{12000}$ return reject.

- Take M (resp. M') samples $X'^{(1)}, \ldots, X'^{(M)}$ from $P_{V'}$ (resp. $Y'^{(1)}, \ldots, Y'^{(M')}$ from $Q_{V'}$), and define
$$W_{\text{light}} = \sum_{i \in V'} \left((W'_i - V'_i)^2 - (W'_i + V'_i) \right) \ ,$$
for V'_i, W'_i defined as $W'_i = \sum_{j=1}^{M_i} X_i'^{(j)}$, $V'_i = \sum_{j=1}^{M'_i} Y_i'^{(j)}$ for all $i \in V'$.

If $W_{\text{light}} \geq \frac{\epsilon^2}{600n}$ return reject.

return accept.

Figure 2: Closeness testing between two unknown product distributions P, Q over $\{0, 1\}^n$.

Claim 33 *Assume $\|P - Q\|_1 > \epsilon$. Then, at least one of the following two cases must hold:*
(i) $\mathbb{E}\left[\|p_{V'} - q_{V'}\|_2^2 \right] > \frac{\epsilon^2}{150n}$, or (ii) $\mathbb{E}\left[\sum_{i \in U' \cap U} \frac{(p_i - q_i)^2}{p_i + q_i} \right] > \frac{\epsilon^2}{128}$.

Proof By definition, any fixed i belongs to V' with probability $(1-p_i)^m(1-q_i)^m$, and so

$$\mathbb{E}\left[\|p_{V'} - q_{V'}\|_2^2\right] = \sum_{i=1}^n (p_i - q_i)^2 \cdot (1-p_i)^m(1-q_i)^m \geq \sum_{i \in V}(p_i-q_i)^2 \cdot (1-p_i)^m(1-q_i)^m$$

$$\geq \left(1 - \frac{1}{m}\right)^{2m} \sum_{i \in V}(p_i-q_i)^2 = \left(1 - \frac{1}{m}\right)^{2m} \|p_V - q_V\|_2^2 \geq \frac{1}{9}\|p_V - q_V\|_2^2 \,,$$

for $m \geq 10$. Similarly,

$$\mathbb{E}\left[\sum_{i \in U' \cap U} \frac{(p_i - q_i)^2}{p_i + q_i}\right] = \sum_{i \in U} \frac{(p_i - q_i)^2}{p_i + q_i} \cdot (1 - (1-p_i)^m(1-q_i)^m)$$

$$\geq \left(1 - \left(1 - \frac{1}{m}\right)^{2m}\right) \sum_{i \in U} \frac{(p_i - q_i)^2}{p_i + q_i} \geq \frac{1}{2} \sum_{i \in U} \frac{(p_i - q_i)^2}{p_i + q_i} \,,$$

and in both cases the proof follows by Claim 32. ∎

We will require the following implication:

Claim 34 *Assume $\|P - Q\|_1 > \epsilon$. Then, at least one of the following must hold with probability at least $4/5$ (over the choice of U', V'): (i) $\|p_{V'} - q_{V'}\|_2^2 > \frac{\epsilon^2}{300n}$, or (ii) $\sum_{i \in U' \cap U} \frac{(p_i - q_i)^2}{p_i + q_i} > \frac{\epsilon^2}{2000}$.*

Proof First, assume that $\|p_V - q_V\|_2^2 > \frac{\epsilon^2}{16n}$, and let V'' denote the random variable $V' \cap V$. By (the proof of) Claim 33, we have $\mathbb{E}\left[\|p_{V''} - q_{V''}\|_2^2\right] \geq \frac{1}{9}\|p_V - q_V\|_2^2 > \frac{\epsilon^2}{150n}$. Writing $m^2\|p_{V''} - q_{V''}\|_2^2 = \sum_{i=1}^n m^2(p_i - q_i)^2 \mathbf{1}_{i \in V''}$ (note that each summand is in $[0, 1]$), we then get by a Chernoff bound that

$$\Pr\left[\|p_{V''} - q_{V''}\|_2^2 < \frac{\epsilon^2}{300n}\right] < e^{-\frac{1}{8}\frac{m^2\epsilon^2}{150n}} < e^{-\frac{C}{1200}\frac{1}{\epsilon^2}} < \frac{1}{5} \,,$$

using our setting of m (for an appropriate choice of the constant $C > 0$).

Suppose now that $\sum_{i \in U} \frac{(p_i - q_i)^2}{p_i + q_i} > \frac{\epsilon^2}{64}$. We divide the proof in two cases.

- Case 1: there exists $i^* \in U$ such that $\frac{(p_i - q_i)^2}{p_i + q_i} > \frac{\epsilon^2}{2000}$. Then $\Pr\left[\sum_{i \in U' \cap U} \frac{(p_i - q_i)^2}{p_i + q_i} > \frac{\epsilon^2}{2000}\right] \geq \Pr[i^* \in U'] \geq 1 - \left(1 - \frac{1}{m}\right)^{2m} > \frac{4}{5}$.

- Case 2: $\frac{(p_i - q_i)^2}{p_i + q_i} \leq \frac{\epsilon^2}{2000}$ for all $i \in U$. Then, writing $X_i \stackrel{\text{def}}{=} \frac{2000}{\epsilon^2} \frac{(p_i - q_i)^2}{p_i + q_i} \mathbf{1}_{i \in U' \cap U} \in [0, 1]$ for all $i \in [n]$, we have $\mathbb{E}[\sum_{i=1}^n X_i] \geq \frac{2000}{128}$ by Claim 33, and a multiplicative Chernoff bound ensures that

$$\Pr\left[\sum_{i \in U' \cap U} \frac{(p_i - q_i)^2}{p_i + q_i} < \frac{\epsilon^2}{2000}\right] \leq \Pr\left[\sum_{i=1}^n X_i < 1\right] \leq e^{-\frac{2000}{8 \cdot 128}} < \frac{1}{5} \,,$$

concluding the proof. ∎

Finally, we will need to bound the expected ℓ_2-norm of $p_{V'}$ and $q_{V'}$.

Claim 35 *For U', V' defined as above, we have $\mathbb{E}\left[\|p_{V'}\|_2^2\right], \mathbb{E}\left[\|q_{V'}\|_2^2\right] \leq \frac{n}{m^2}$.*

Proof By symmetry, it is sufficient to bound $\mathbb{E}\left[\|p_{V'}\|_2^2\right]$. We have

$$\mathbb{E}\left[\|p_{V'}^2\|_2\right] = \sum_{i=1}^n p_i^2 \cdot (1-p_i)^m (1-q_i)^m \leq \sum_{i=1}^n p_i^2 \cdot (1-p_i)^m.$$

Studying the auxiliary function $f \colon x \in [0,1] \mapsto x^2(1-x)^m$, we see that it achieves a maximum at $\frac{2}{m+2}$. We can then bound

$$\mathbb{E}\left[\|p_{V'}^2\|_2\right] \leq n \cdot f\left(\frac{2}{m+2}\right) \sim_{m\to\infty} \frac{4n}{e^2 m^2},$$

and so $\mathbb{E}\left[\|p_{V'}^2\|_2\right] \leq \frac{n}{m^2}$ for m large enough (and this actually holds for any $m \geq 1$). ∎

In what follows, we analyze our statistics W_{heavy} and W_{light}, conditioning on U', V'.

Case 1: discrepancy in U'. We assume that Algorithm 2 reached the line where W_{heavy} is computed, and show the following:

Lemma 36 *If $P = Q$, then with probability at least $9/10$ we have $W_{\text{heavy}} \leq \frac{m\epsilon^2}{12000}$. Conversely, if $\sum_{i \in U \cap U'} \frac{(p_i - q_i)^2}{p_i + q_i} > \frac{\epsilon^2}{2000}$, then $W_{\text{heavy}} \geq \frac{m\epsilon^2}{12000}$ with probability at least $9/10$.*

Proof Recall that the W_i's are independent, as P is a product distribution and the M_i's are independent. Similarly for the V_i's. We have:

Claim 37 *If $P = Q$, then $\mathbb{E}[W_{\text{heavy}}] = 0$. Moreover, if $\sum_{i \in U \cap U'} \frac{(p_i - q_i)^2}{p_i + q_i} > \frac{\epsilon^2}{2000}$, then $\mathbb{E}[W_{\text{heavy}}] > \frac{m\epsilon^2}{6000}$.*

Proof Note that $W_i \sim \text{Poi}(mp_i)$ and $V_i \sim \text{Poi}(mq_i)$ for all $i \in U'$. From there, we can compute (as in Chan et al. (2014c))

$$\mathbb{E}\left[\frac{(W_i - V_i)^2 - (W_i + V_i)}{W_i + V_i}\right] = m \frac{(p_i - q_i)^2}{p_i + q_i}\left(1 - \frac{1 - e^{-m(p_i + q_i)}}{m(p_i + q_i)}\right),$$

by first conditioning on $W_i + V_i$. This immediately gives the first part of the claim. As for the second, observing that $1 - \frac{1-e^{-x}}{x} \geq \frac{1}{3}\min(1, x)$ for $x \geq 0$, and that $p_i + q_i \geq \frac{1}{m}$ for all $i \in U$, by definition we get

$$\mathbb{E}[W_{\text{heavy}}] = m \sum_{i \in U'} \frac{(p_i - q_i)^2}{p_i + q_i}\left(1 - \frac{1 - e^{-m(p_i + q_i)}}{m(p_i + q_i)}\right) \geq \frac{1}{3}m \sum_{i \in U \cap U'} \frac{(p_i - q_i)^2}{p_i + q_i} \geq \frac{m\epsilon^2}{6000}.$$
∎

We can now bound the variance of our estimator:

Claim 38 $\mathrm{Var}[W_{\mathrm{heavy}}] \leq 2n + 5m \sum_{i \in U'} \frac{(p_i - q_i)^2}{p_i + q_i} \leq 7n + \frac{3}{5} \mathbb{E}[W_{\mathrm{heavy}}]$. *In particular, if $P = Q$ then* $\mathrm{Var}[W_{\mathrm{heavy}}] \leq 2n$.

Proof The proof of the first inequality is similar to that in (Chan et al., 2014c, Lemma 5), with a difference in the final bound due to the fact that the p_i's and q_i's no longer sum to one. For completeness, we give the proof below.

First, as by independence of the V_i, W_i's we have $\mathrm{Var}[W_{\mathrm{heavy}}] = \sum_{i \in U'} \mathrm{Var}\left[\frac{(W_i - V_i)^2 - (W_i + V_i)}{W_i + V_i}\right]$, it is sufficient to bound each summand individually. In order to do so, we split the variance calculation into two parts: the variance conditioned on $W_i + V_i = j$, and the component of the variance due to the variation in j. Writing for convenience

$$f(W_i, V_i) \stackrel{\mathrm{def}}{=} \frac{(W_i - V_i)^2 - W_i - V_i}{W_i + V_i} \, ,$$

we have that

$$\mathrm{Var}[f(X, Y)] \leq \max_j \left(\mathrm{Var}[f(X, Y) \mid X + Y = j]\right) + \mathrm{Var}[\mathbb{E}[f(X, Y) \mid X + Y = j]] \, .$$

We now bound the first term. Since $(W_i - V_i)^2 = (j - 2V_i)^2$, and V_i is distributed as $\mathrm{Bin}(j, \alpha)$ (where for conciseness we let $\alpha \stackrel{\mathrm{def}}{=} \frac{q_i}{p_i + q_i}$), we can compute the variance of $(j - 2V_i)^2$ from standard expressions for the moments of the Binomial distribution as $\mathrm{Var}[(j - 2V_i)^2] = 16j(j - 1)\alpha(1 - \alpha)\left((j - \frac{3}{2})(1 - 2\alpha)^2 + \frac{1}{2}\right)$. Since $\alpha(1 - \alpha) \leq \frac{1}{4}$ and $j - \frac{3}{2} < j - 1 < j$, this in turn is at most $j^2(2 + 4j(1 - 2\alpha)^2)$. Because the denominator is $W_i + V_i$ which equals j, we must divide this by j^2, make it 0 when $j = 0$, and take its expectation as j is distributed as $\mathrm{Poi}(m(p_i + q_i))$. This leads to

$$\mathrm{Var}[f(W_i, V_i) \mid W_i + V_i = j] \leq 2(1 - e^{-m(p_i + q_i)}) + 4m\frac{(p_i - q_i)^2}{p_i + q_i} \, .$$

We now consider the second component of the variance–the contribution to the variance due to the variation in the sum $W_i + V_i$. Since for fixed j, as noted above, we have V_i distributed as $\mathrm{Bin}(j, \alpha)$, we have

$$\mathbb{E}[(W_i - V_i)^2] = \mathbb{E}[j^2 - 4jV_i + 4V_i^2] = j^2 - 4j^2\alpha + 4(j\alpha - j\alpha^2 + j^2\alpha^2) = j^2(1 - 2\alpha)^2 + 4j\alpha(1 - \alpha) \, .$$

We finally subtract $W_i + V_i = j$ and divide by j to yield $(j - 1)(1 - 2\alpha)^2$, except with a value of 0 when $j = 0$ by definition. However, note that replacing the value at $j = 0$ with 0 can only lower the variance. Since the sum $j = W_i + V_i$ is drawn from a Poisson distribution with parameter $m(p_i + q_i)$, we thus have:

$$\mathrm{Var}\left[\mathbb{E}[f(W_i, V_i) | W_i + V_i = j]\right] \leq m(p_i + q_i)(1 - 2\alpha)^4 \leq m(p_i + q_i)(1 - 2\alpha)^2 = m\frac{(p_i - q_i)^2}{p_i + q_i} \, .$$

Summing the final expressions of the previous two paragraphs yields a bound on the variance of $f(W_i V_i)$ of

$$2(1 - e^{-m(p_i + q_i)}) + 5m\frac{(p_i - q_i)^2}{p_i + q_i} \leq 2 + 5m\frac{(p_i - q_i)^2}{p_i + q_i} \, ,$$

as $1 - e^{-x} \le 1$ for all x. This shows that

$$\mathrm{Var}[W_{\mathrm{heavy}}] \le 2n + 5m \sum_{i \in U'} \frac{(p_i - q_i)^2}{p_i + q_i} = 2n + 5m \sum_{i \in U' \cap U} \frac{(p_i - q_i)^2}{p_i + q_i} + 5m \sum_{i \in U' \cap V} \frac{(p_i - q_i)^2}{p_i + q_i}$$

$$\le 2n + \frac{3}{5}\mathbb{E}[W_{\mathrm{heavy}}] + 5m \sum_{i \in U' \cap V} \frac{(p_i - q_i)^2}{p_i + q_i} ,$$

so it only remains to bound the last term. But by definition, $i \in V$ implies $0 \le p_i, q_i < \frac{1}{m}$, from which

$$5m \sum_{i \in U' \cap V} \frac{(p_i - q_i)^2}{p_i + q_i} \le 5 \sum_{i \in U' \cap V} \frac{|p_i - q_i|}{p_i + q_i} \le 5|U' \cap V| \le 5n .$$

This completes the proof. ∎

With these two claims in hand, we are ready to conclude the proof of Lemma 36. We start with the soundness case, i.e. assuming $\sum_{i \in U \cap U'} \frac{(p_i - q_i)^2}{p_i + q_i} > \frac{\epsilon^2}{2000}$. Then, by Chebyshev's inequality and Claim 37 we have that

$$\Pr\left[W_{\mathrm{heavy}} < \frac{m\epsilon^2}{12000} \right] \le \Pr\left[\mathbb{E}[W_{\mathrm{heavy}}] - W_{\mathrm{heavy}} > \frac{1}{2}\mathbb{E}[W_{\mathrm{heavy}}] \right] \le \frac{4 \, \mathrm{Var}[W_{\mathrm{heavy}}]}{\mathbb{E}[W_{\mathrm{heavy}}]^2}$$

$$\le \frac{28n}{\mathbb{E}[W_{\mathrm{heavy}}]^2} + \frac{12}{5\mathbb{E}[W_{\mathrm{heavy}}]} \qquad \text{(by Claim 38)}$$

$$\le \frac{9 \cdot 2000^2 \cdot 28n}{m^2 \epsilon^4} + \frac{36 \cdot 2000}{5\epsilon^2 m} = O\left(\frac{n}{m^2\epsilon^4} + \frac{1}{\epsilon^2 m} \right) . \qquad (5)$$

We want to bound this quantity by $1/10$, for which it suffices to have $m > C\frac{\sqrt{n}}{\epsilon^2}$ for an appropriate choice of the absolute constant $C > 0$ in our setting of m.

Turning to the completeness, assume that $\|P - Q\|_1 = 0$. Then, by Chebyshev's inequality, and invoking Claim 38 we have:

$$\Pr\left[W \ge \frac{m\epsilon^2}{12000} \right] = \Pr\left[W \ge \mathbb{E}[W] + \frac{m\epsilon^2}{12000} \right] \le \frac{36 \cdot 2000^2 \, \mathrm{Var}[W]}{\epsilon^4 m^2} = O\left(\frac{n}{\epsilon^4 m^2} \right) ,$$

which is no more than $1/10$ for the same choice of m.

Case 2: discrepancy in V'. We now assume that Algorithm 2 reached the line where W_{light} is computed, and show the following:

Lemma 39 *If $P = Q$, then with probability at least $9/10$ we have $W_{\mathrm{light}} \le \frac{\epsilon^2}{600n}$. Conversely, if $\|p_{V'} - q_{V'}\|_2^2 > \frac{\epsilon^2}{300n}$, then $W_{\mathrm{light}} \ge \frac{\epsilon^2}{600n}$ with probability at least $9/10$.*

Proof We condition on $\|p_V'\|_2^2, \|q_V'\|_2^2 \le \frac{20n}{m^2}$, which by Claim 35, a union bound, and Markov's inequality happens with probability at least $19/20$. The analysis is similar to (Chan et al., 2014c, Section 3), observing that the $(V_i')_{i \in V'}, (W_i')_{i \in V'}$'s are mutually independent Poisson random variables, V_i' (resp. W_i') having mean mp_i (resp. mq_i). Namely, following their analysis, the statistic W_{light} is an unbiased estimator for $m^2\|p_{V'} - q_{V'}\|_2^2$ with variance

$$\mathrm{Var}[W_{\mathrm{light}}] \le 8m^3\sqrt{b}\|p_{V'} - q_{V'}\|_2^2 + 8m^2 b ,$$

25

where $b \stackrel{\text{def}}{=} \frac{20n}{m^2}$ is our upper bound on $\|p'_V\|_2^2, \|q'_V\|_2^2$. From there, setting $\epsilon' \stackrel{\text{def}}{=} \frac{\epsilon}{\sqrt{n}}$ and applying Chebyshev's inequality, we get that there exists an absolute constant $C' > 0$ such the completeness and soundness guarantees from the lemma holds with probability at least $19/20$, provided that $m > C' \frac{\sqrt{b}}{\epsilon'^2}$, i.e.,

$$m > C' \frac{n}{\epsilon^2} \cdot \frac{\sqrt{20n}}{m^2} = \sqrt{20} C' \frac{n^{3/2}}{m\epsilon^2} \ .$$

Solving for m shows that choosing $m \geq C \frac{n^{3/4}}{\epsilon}$ for some absolute constant $C > 0$ is enough. A union bound then allows us to conclude the proof of the lemma, guaranteeing correctness with probability at least $1 - \frac{1}{20} - \frac{1}{20} = \frac{9}{10}$. ∎

∎

∎

5.2. Sample Complexity Lower Bound for Closeness Testing

In this section, we prove a matching information-theoretic lower bound for testing closeness of two unknown arbitrary product distributions; showing that, perhaps surprisingly, the sample complexity of Theorem 31 is in fact optimal, up to constant factors:

Theorem 40 *There exists an absolute constant $\epsilon_0 > 0$ such that, for any $0 < \epsilon \leq \epsilon_0$, the following holds: Any algorithm that has sample access to two unknown* product *distribution P, Q over $\{0,1\}^n$ and distinguishes between the cases that $P = Q$ and $\|P - Q\|_1 > \epsilon$ requires $\Omega(\max(\sqrt{n}/\epsilon^2, n^{3/4}/\epsilon))$ samples.*

Before delving into the proof, we give some intuition for the $n^{3/4}$ term of the lower bound lower bound. Recall that the hard family of instances for distinguishing discrete distributions over $[n]$ had (a) many "light" bins (domain elements) of probability mass approximately $1/n$, where either $p_i = q_i$ on each bin or $p_i = q_i(1 \pm \epsilon)$ in each bin, and (b) a number of "heavy" bins where $p_i = q_i \approx 1/k$ (where k was the number of samples taken). The goal of the heavy bins was to "add noise" and hide the signal from the light bins. In the case of discrete distributions over $[n]$, we could only have k such heavy bins. In the case of product distributions, there is no such restriction, and we can have $n/2$ of them in our hard instance. The added noise leads to an increased sample complexity of testing closeness in the high-dimensional setting.

Proof The first part of the lower bound, $\Omega(\sqrt{n}/\epsilon^2)$, follows from Theorem 9; we focus here on the second term, $\Omega(n^{3/4}/\epsilon)$, and consequently assume hereafter that $\sqrt{n}/\epsilon^2 < n^{3/4}/\epsilon$. Let $k \geq 1$ be fixed, and suppose we have a tester that takes $k = o(n^{3/4}/\epsilon)$ samples: we will show that it can only be correct with vanishing probability. We will again follow the information-theoretic framework of Diakonikolas and Kane (2016) for proving distribution testing lower bounds, first defining two distributions over pairs of product distributions $\mathcal{Y}, \mathcal{N}$:

- $\mathcal{Y}$: for every $i \in [n]$, independently choose (p_i, q_i) to be either $p_i = q_i = \frac{1}{k}$ with probability $1/2$, and $p_i = q_i = \frac{1}{n}$ otherwise; and set $P \stackrel{\text{def}}{=} \bigotimes_{j=1}^n \text{Bern}(p_i)$, $Q \stackrel{\text{def}}{=} \bigotimes_{j=1}^n \text{Bern}(q_i)$.

- $\mathcal{N}$: for every $i \in [n]$, independently choose (p_i, q_i) to be either $p_i = q_i = \frac{1}{k}$ with probability $1/2$, and $(\frac{1+\epsilon}{n}, \frac{1-\epsilon}{n})$ or $(\frac{1-\epsilon}{n}, \frac{1+\epsilon}{n})$ uniformly at random otherwise; and set $P \overset{\text{def}}{=} \bigotimes_{j=1}^{n} \text{Bern}(p_i)$, $Q \overset{\text{def}}{=} \bigotimes_{j=1}^{n} \text{Bern}(q_i)$.

Note that in both $\mathcal{Y}$ and $\mathcal{N}$, with overwhelming probability the pairs (P, Q) have roughly $n/2$ marginals with (equal) parameter $1/k$, and roughly $n/2$ marginals with parameter $\Theta(1)/n$.

Lemma 41 *With probability $1 - 2^{-\Omega(n)}$, a uniformly chosen pair $(P, Q) \sim \mathcal{N}$ satisfies $\|P - Q\|_1 = \Omega(\epsilon)$.*

Proof Similar to that of Lemma 14. ∎

We will as before make a further simplification, namely that instead of drawing k samples from $P = P_1 \otimes \cdots \otimes P_n$ and $Q = Q_1 \otimes \cdots \otimes Q_n$, the algorithm is given k_i samples from each P_i (resp. k_i' from Q_i), where $k_1, \ldots, k_n, k_1', \ldots, k_n'$ are independent $\text{Poi}(k)$ random variables. We now consider the following process: letting $X \sim \text{Bern}(1/2)$ be a uniformly random bit, we choose a pair of distributions (P, Q) (both P and Q being probability distributions over $\{0, 1\}^n$) by

- Drawing $(P, Q) \sim \mathcal{Y}$ if $X = 0$, and;

- Drawing $(P, Q) \sim \mathcal{N}$ if $X = 1$;

- Drawing $k_1, k_1', \ldots, k_n, k_n' \sim \text{Poi}(k)$, and returning k_i samples from each P_i and k_i' samples from each Q_i

For $i \in [n]$, we let N_i and M_i denote respectively the number of 1's among the k_i samples drawn from P_i and k_i' samples drawn from Q_i, and write $N = (N_1, \ldots, N_n) \in \mathbb{N}^n$ (and $M \in \mathbb{N}^n$ for Q). The next step is then to upperbound $I(X; (N, M))$, in order to conclude that it will be $o(1)$ unless k is taken big enough and invoke Fact 11. By the foregoing discussion and the relaxation on the k_i's, we have that the conditioned on X the N_i's (and M_i's) are independent (with $N_i \sim \text{Poi}(kp_i)$ and $M_i \sim \text{Poi}(kq_i)$). This implies that

$$I(X; (N, M)) \leq \sum_{i=1}^{n} I(X; (N_i, M_i)) \tag{6}$$

so that it suffices to bound each $I(X; (N_i, M_i))$ separately.

Lemma 42 *Fix any $i \in [n]$, and let X, N_i, M_i be as above. Then $I(X; (N_i, M_i)) = O(k^4 \epsilon^4 / n^4)$.*

Proof By symmetry it is enough to consider only the case of $i = 1$, so that we let $(A, B) = (N_1, M_1)$.

Since $A \sim \text{Poi}(kp_1)$ and $B \sim \text{Poi}(kq_1)$ with $(p_1, q_1) = (1/k, 1/k)$ or $(p_1, q_1) = (1/n, 1/n)$ uniformly if $X = 0$, and

$$(p_1, q_1) = \begin{cases} (\frac{1}{k}, \frac{1}{k}) & \text{w.p. } \frac{1}{2} \\ (\frac{1+\epsilon}{n}, \frac{1-\epsilon}{n}) & \text{w.p. } \frac{1}{4} \\ (\frac{1-\epsilon}{n}, \frac{1+\epsilon}{n}) & \text{w.p. } \frac{1}{4} \end{cases}$$

if $X = 1$, a computation similar as that of (Diakonikolas and Kane, 2016, Proposition 3.8) yields that, for any $i, j \in \mathbb{N}$

$$\Pr[(A, B) = (i, j) \mid X = 0] = \frac{1}{2i!j!}\left(e^{-2k/k}\left(\frac{k}{k}\right)^{i+j} + e^{-2k/n}\left(\frac{k}{n}\right)^{i+j}\right)$$

$$= \frac{1}{2i!j!}\left(e^{-2} + e^{-2k/n}\left(\frac{k}{n}\right)^{i+j}\right)$$

$$\Pr[(A, B) = (i, j) \mid X = 1] = \frac{1}{2i!j!}\left(e^{-2k/k}\left(\frac{k}{k}\right)^{i+j}\right.$$

$$\left. + e^{-2k/n}\left(\frac{k}{n}\right)^{i+j}\left(\frac{(1+\epsilon)^i(1-\epsilon)^j + (1-\epsilon)^i(1+\epsilon)^j}{2}\right)\right)$$

$$= \frac{1}{2i!j!}\left(e^{-2} + e^{-2\frac{k}{n}}\left(\frac{k}{n}\right)^{i+j}\left(\frac{(1+\epsilon)^i(1-\epsilon)^j + (1-\epsilon)^i(1+\epsilon)^j}{2}\right)\right).$$

Note in particular that for $0 \leq i + j \leq 1$, this implies that $\Pr[(A, B) = (i, j) \mid X = 0] = \Pr[(A, B) = (i, j) \mid X = 1]$. From the above, we obtain

$$I(X; (A, B)) = O(1) \cdot \sum_{i,j \geq 0} \frac{(\Pr[(A, B) = (i, j) \mid X = 0] - \Pr[(A, B) = (i, j) \mid X = 1])^2}{\Pr[(A, B) = (i, j) \mid X = 0] + \Pr[(A, B) = (i, j) \mid X = 1]}$$

$$= O(1) \cdot \sum_{i+j \geq 2} \frac{(\Pr[(A, B) = (i, j) \mid X = 0] - \Pr[(A, B) = (i, j) \mid X = 1])^2}{\Pr[(A, B) = (i, j) \mid X = 0] + \Pr[(A, B) = (i, j) \mid X = 1]}$$

$$= O(1) \cdot \sum_{i+j \geq 2} e^{-\frac{4k}{n}} \frac{\left(\frac{k}{n}\right)^{2(i+j)}}{2i!j!} \frac{(1 - \frac{1}{2}((1+\epsilon)^i(1-\epsilon)^j + (1-\epsilon)^i(1+\epsilon)^j))^2}{2e^{-2} + o(1)}$$

$$= O\left((k\epsilon/n)^4\right)$$

where the second-to-last inequality holds for $k = o(n)$. (Which is the case, as $\sqrt{n}/\epsilon^2 < n^{3/4}/\epsilon$ implies that $n^{3/4}/\epsilon < n$, and we assumed $k = o(n^{3/4}/\epsilon)$.) ∎

This lemma, along with Eq. (6), immediately implies the result:

$$I(X; (N, M)) \leq \sum_{i=1}^{n} O\left(\left(\frac{k\epsilon}{n}\right)^4\right) = O\left(\frac{k^4\epsilon^4}{n^3}\right) \tag{7}$$

which is $o(1)$ unless $k = \Omega(n^{3/4}/\epsilon)$.

∎

5.3. Ruling Out Tolerant Testing Without Balancedness

In this section, we show that any *tolerant* identity testing algorithm for product distributions must have sample complexity near-linear in n if the explicitly given distribution is very biased.

Theorem 43 *There exists an absolute constant $\epsilon_0 < 1$ such that the following holds. Any algorithm that, given a parameter $\epsilon \in (0, \epsilon_0]$ and sample access to product distributions P, Q over $\{0, 1\}^n$, distinguishes between $\|P - Q\|_1 < \epsilon/2$ and $\|P - Q\|_1 > \epsilon$ with probability at least $2/3$ requires $\Omega(n/\log n)$ samples. Moreover, the lower bound still holds in the case where Q is known, and provided as an explicit parameter.*

Proof The basic idea will be to reduce to the case of tolerant testing of two arbitrary distributions p and q over $[n]$. In order to do this, we define the following function from distributions of one type to distributions of the other:

If p is a distribution over $[n]$, define $F_\delta(p)$ to be the distribution over $\{0, 1\}^n$ obtained by taking $\mathrm{Poi}(\delta)$ samples from p and returning the vector x where $x_i = 1$ if and only if i was one of these samples drawn. Note that, because of the Poissonization, $F_\delta(p)$ is a product distribution. We have the following simple claim:

Claim 44 *For any $\delta \in (0, 1]$ and distributions p, q on $[n]$, $d_{\mathrm{TV}}(F_\delta(p), F_\delta(q)) = (\delta + O(\delta^2))d_{\mathrm{TV}}(p, q)$.*

Proof In one direction, we can take correlated samples from $F_\delta(p)$ and $F_\delta(q)$ by sampling a from $\mathrm{Poi}(\delta)$ and then taking a samples from each of p and q, using these to generate our samples from $F_\delta(p), F_\delta(q)$. For fixed a, the variation distance between $F_\delta(p)$ and $F_\delta(q)$ conditioned on that value of a is clearly at most $a d_{\mathrm{TV}}(p, q)$. Therefore, $d_{\mathrm{TV}}(F_\delta(p), F_\delta(q)) \leq \mathbb{E}[a] d_{\mathrm{TV}}(p, q) = \delta d_{\mathrm{TV}}(p, q)$.

In the other direction, note that $F_\delta(p)$ and $F_\delta(q)$ each have probability $\delta + O(\delta^2)$ of returning a vector of weight 1. This is because $\mathrm{Poi}(\delta) = 1$ with probability $\delta e^{-\delta} = \delta + O(\delta^2)$ and since $\mathrm{Poi}(\delta) > 1$ with probability $O(\delta^2)$. Let $G(p)$ and $G(q)$ denote the distributions $F_\delta(p)$ and $F_\delta(q)$ conditioned on returning a vector of weight 1. By the above, we have that $d_{\mathrm{TV}}(F_\delta(p), F_\delta(q)) \geq (\delta + O(\delta^2))d_{\mathrm{TV}}(G(p), G(q))$. Letting p_i (resp. q_i) be the probability that p (resp. q) assigns to $i \in [n]$, we get that for any fixed $i \in [n]$ the probability that $F_\delta(p)$ returns e_i is

$$(1 - e^{-\delta p_i}) \prod_{j \neq i} e^{-\delta p_j} = (e^{\delta p_i} - 1) \prod_{j=1}^{n} e^{-\delta p_j} .$$

Therefore $G(p)$ puts on e_i probability proportional to $(e^{\delta p_i} - 1) = (\delta + O(\delta^2))p_i$. Similarly, the probability that $G(q)$ puts on e_i is proportional to $(\delta + O(\delta^2))q_i$ (where in both cases, the constant of proportionality is $(\delta + O(\delta^2))^{-1}$). Therefore,

$$\begin{aligned}
d_{\mathrm{TV}}(G(p), G(q)) &= \delta^{-1}(1 + O(\delta)) \sum_{i=1}^{n} |(\delta + O(\delta^2))p_i - (\delta + O(\delta^2))q_i| \\
&= \delta^{-1}(1 + O(\delta)) \sum_{i=1}^{n} (\delta|p_i - q_i| + O(\delta^2)(p_i + q_i)) \\
&= \delta^{-1}(1 + O(\delta))(\delta d_{\mathrm{TV}}(p, q) + O(\delta^2)) \\
&= d_{\mathrm{TV}}(p, q) + O(\delta) .
\end{aligned}$$

Thus, $d_{\mathrm{TV}}(F_\delta(p), F_\delta(q)) \geq (\delta + O(\delta^2))d_{\mathrm{TV}}(p, q)$. This completes the proof. ∎

The above claim guarantees the existence of some constant $\delta_0 \in (0, 1]$ such that $d_{\mathrm{TV}}(F_{\delta_0}(p), F_{\delta_0}(q)) \in [0.9\delta_0 d_{\mathrm{TV}}(p, q), 1.1 d_{\mathrm{TV}}(p, q)]$. However, it is known [Valiant and Valiant](2011) that for any sufficiently small $\epsilon > 0$ there exist distributions p and q over $[n]$ such that one must take at least $c\frac{n}{\log n}$

samples (where $c > 0$ is an absolute constant) to distinguish between $d_{\mathrm{TV}}(p, q) \leq \epsilon/(2 \cdot 0.9\delta_0)$ and $d_{\mathrm{TV}}(p, q) \geq \epsilon/(1.1\delta_0)$. Given q samples from p and q we can with high probability simulate $c'q$ samples from $P = F_{\delta_0}(p)$ and $Q = F_{\delta_0}(q)$ (where $c' = c'(\delta_0) > 0$ is another absolute constant). Therefore, we cannot distinguish between the cases $d_{\mathrm{TV}}(P, Q) \leq \epsilon/2$ and $d_{\mathrm{TV}}(P, Q) \geq \epsilon$ in fewer than $c' \cdot c\frac{n}{\log n}$, as doing so would enable us to distinguish between p and q with less than $c\frac{n}{\log n}$ samples – yielding a contradiction. Moreover, the above still holds when q is explicitly known, specifically even when q is taken to be the uniform distribution on $[n]$. ∎

6. From Product Distributions to Bayes nets: an Overview

In the following paragraphs, we describe how to generalize our previous results for product distributions to testing general Bayes nets. The case of known structure turns out to be manageable, and at a technical level a generalization of our testers for product distributions. The case of unknown structure poses various complications and requires a number of non-trivial new ideas.

Testing Identity and Closeness of Fixed Structure Bayes Nets. Our testers and matching lower bounds for the fixed structure regime are given in Sections 7 and 9.1.

For concreteness, let us consider the case of testing identity of a tree-structured ($d = 1$) Bayes net P against an explicit tree-structured Bayes net Q with the same structure. Recall that we are using as a proxy for the distance $\|P - Q\|_1$ an appropriate chi-squared-like quantity. A major difficulty in generalizing our identity tester for products is that the chi-squared statistic depends not on the probabilities of the various coordinates, but on the conditional probabilities of these coordinates based on all possible parental configurations. This fact produces a major wrinkle in our analysis for the following reason: while in the product distribution case each sample provides information about each coordinate probability, in the Bayes net case a sample only provides information about conditional probabilities for parental configurations that actually occurred in that sample.

This issue can be especially problematic to handle if there are uncommon parental configurations about which we will have difficulty gathering much information (with a small sized sample). Fortunately, the probabilities conditioned on such parental configurations will have a correspondingly smaller effect on the final distribution and thus, we will not need to know them to quite the same accuracy. So while this issue can be essentially avoided, we will require some technical assumptions about *balancedness* to let us know that none of the parental configurations are too rare. Using these ideas, we develop an identity tester for tree-structured Bayes nets that uses an optimal $\Theta(\sqrt{n}/\epsilon^2)$ samples. For known structure Bayes nets of degree $d > 1$, the sample complexity will also depend exponentially on the degree d. Specifically, each coordinate will have as many as 2^d parental configurations. Thus, instead of having only n coordinate probabilities to worry about, we will need to keep track of $2^d n$ conditional probabilities. This will require that our sample complexity also scale like $2^{d/2}$. The final complexity of our identity and closeness testers will thus be $O(2^{d/2}\sqrt{n}/\epsilon^2)$.

We now briefly comment on our matching lower bounds. Our sample complexity lower bound of $\Omega(\sqrt{n}/\epsilon^2)$ for the product case can be generalized in a black-box manner to yield a tight lower bound $\Omega(2^{d/2}\sqrt{n}/\epsilon^2)$ for testing uniformity of degree-d Bayes nets. The basic idea is to consider degree-d Bayes nets with the following structure: The first d nodes are all independent (with marginal probability $1/2$ each), and will form in some sense a "pointer" to one of 2^d arbitrary product distributions. The remaining $n - d$ nodes will each depend on all of the first d. The resulting

distribution is now an (evenly weighted) disjoint mixture of 2^d product distributions on the $(n-d)$-dimensional hypercube. In other words, there are 2^d product distributions $p_1, \ldots, p_{2^d}$, and our distribution returns a random i (encoded in binary) followed by a random sample form p_i. By using the fact that the p_i's can be arbitrary product distributions, we obtain our desired sample complexity lower bound.

Testing Identity and Closeness of Unknown Structure Bayes Nets. As we show in Sections 8 and 9.2, this situation changes substantially when we do not know the underlying structure of the nets involved. In particular, we show that even for Bayes nets of degree-1 uniformity testing requires $\Omega(n/\epsilon^2)$ samples.

The lower bound construction for this case is actually quite simple: The adversarial distribution P will be developed by taking a *random matching* of the vertices and making each matched pair of vertices randomly $1 \pm \epsilon/\sqrt{n}$ correlated. If the matching were known by the algorithm, the testing procedure could proceed by approximating these $n/2$ correlations. However, not knowing the structure, our algorithm would be forced to consider all $\binom{n}{2}$ pairwise correlations, substantially increasing the amount of noise involved. To actually prove this lower bound, we consider the distribution X obtained by taking k samples from a randomly chosen P and Y from taking k samples from the uniform distribution. Roughly speaking, we wish to show that $\chi^2(X, Y)$ is approximately 1. This amounts to showing that for a randomly chosen pair of distributions P and P' from this family, we have that $\mathbb{E}[P^k(x)P'^k(x)]$ is approximately 1. Intuitively, we show that this expectation is only large if P and P' share many edges in common. In fact, this expectation can be computed exactly in terms of the lengths of the cycles formed by the graph obtained taking the union of the edges from P and P'. Noting that P and P' typically share only about 1 edge, this allows us to prove our desired lower bound.

However, the hardness of the situation described above is not generic and can be avoided if the explicit distribution Q satisfies some non-degeneracy assumptions. Morally, a Bayes nets Q is non-degenerate if it is not close in variational distance to any other Bayes net of no greater complexity and non-equivalent underlying structure. For tree structures, our condition is that for each node the two conditional probabilities for that node (depending on the value of its parent) are far from each other.[2]

If this is the case, even knowing approximately what the pairwise distributions of coordinates are will suffice to determine the structure. One way to see this is the following: the analysis of the Chow-Liu algorithm Chow and Liu (1968) shows that the tree-structure for P is the maximum spanning tree of the graph whose edge weights are given by the shared information of the nodes involved. This tree will have the property that each edge, e, has higher weight than any other edge connecting the two halves of the tree. We show that our non-degeneracy assumption implies that this edge has higher weight by a noticeable margin, and thus that it is possible to verify that we have the correct tree with only rough approximations to the pairwise shared information of variables.

---

2. One may wonder how this non-degeneracy condition compares to the "faithfulness" notion from the statistics literature. Broadly speaking, faithfulness asks that the set of conditional independence relations satisfied by the distribution are exactly those implied by the structure; under this assumption, there exist algorithms able to learn the structure of the Bayes net, given enough samples. However, since we need here to get guarantees with a finite number of samples, to be sure that our samples came from a distribution with this specific structure we must enforce that additional conditional independence assumptions (that may be associated with different structures) to be far from satisfied. Non-degeneracy is exactly the condition that any such conditional independence assumptions relating nodes that are close in the graph are far from holding. Finally, faithfulness also requires that we do not have certain conditional independence conditions for distant nodes, yet we would expect them to be close to independent.

For Bayes nets of higher degree, the analysis is somewhat more difficult. We need a slightly more complicated notion of non-degeneracy, essentially boiling down to a sizeable number of not-approximately-conditionally-independent assumptions. For example, a pair of nodes can be positively identified as having an edge between them in the underlying graph if they are not conditionally independent upon any set of d other nodes. By requiring that for each edge the relevant coordinate variables are not close to being conditionally independent, we can verify the identity of the edges of $\mathcal{S}$ with relatively few samples. Unfortunately, this is not quite enough, as with higher degree Bayesian networks, simply knowing the underlying undirected graph is not sufficient to determine its structure. We must also be able to correctly identify the so-called $\vee$-structures. To do this, we will need to impose more not-close-to-conditionally-independent assumptions that allow us to robustly determine these as well.

Assuming that Q satisfies such a non-degeneracy condition, testing identity to it is actually quite easy. First one verifies that the distribution P has all of its pairwise (or $(d+2)$-wise) probabilities close to the corresponding probabilities for Q. By non-degeneracy, this will imply that P must have the same (or at least an equivalent) structure as Q. Once this has been established, the testing algorithms for the known structure can be employed.

Sample Complexity of Testing High-Degree Bayes Nets. One further direction of research is that of understanding the dependence on degree of the sample complexity of testing identity and closeness for degree-d Bayes nets without additional assumptions. For $d = 1$, we showed that these problems can be as hard as learning the distribution. For the general case, we give an algorithm with sample complexity $2^{d/2}\text{poly}(n, 1/\epsilon)$ for identity testing (and $2^{2d/3}\text{poly}(n, 1/\epsilon)$ for closeness testing). The conceptual message of this result is that, when the degree increases, testing becomes easier than learning information-theoretically. It is a plausible conjecture that the correct answer for identity testing is $\Theta(2^{d/2}n/\epsilon^2)$ and for closeness testing is $\Theta(2^{2d/3}n/\epsilon^2)$. We suspect that our lower bound techniques can be generalized to match these quantities, but the constructions will likely be substantially more intricate.

The basic idea of our $2^{d/2}\text{poly}(n, 1/\epsilon)$ sample upper bound for identity testing is this: We enumerate over all possible structures for P, running a different tester for each of them by comparing the relevant conditional probabilities. Unfortunately, in this domain, our simple formula for the KL-Divergence between the two distributions will no longer hold. However, we can show that using the old formula will be sufficient by showing that if there are large discrepancies when computing the KL-divergence, then there must be large gap between the entropies $H(P)$ and $H(Q)$ in a particular direction. As the gap cannot exist both ways, this suffices for our purposes.

7. Testing Identity of Fixed Structure Bayes Nets

In this section, we prove our matching upper and lower bounds for testing the identity of Bayes nets with known graph structure. In Section 7.1, we describe an identity testing algorithm that uses $O\left(2^{d/2}\sqrt{n}/\epsilon^2\right)$ samples, where d is the maximum in-degree and n the number of nodes (dimension). In Section 7.2, we show that this sample upper bound is tight, up to constant factors, even for uniformity testing.

7.1. Identity Testing Algorithm

In this section, we establish the upper bound part of Theorem 18 for identity, namely testing identity to a fixed Bayes net given sample access to an unknown Bayes net with the same underlying structure. In order to state our results, we recall the definition of *balancedness* of a Bayes net:

Definition 45 *A Bayes net P over $\{0,1\}^n$ with structure $\mathcal{S}$ is said to be (c, C)-balanced if, for all k, it is the case that (i) $p_k \in [c, 1 - c]$ and (ii) $\Pr_P[\Pi_k] \geq C$.*

Roughly speaking, the above conditions ensure that the conditional probabilities of the Bayes net are bounded away from 0 and 1, and that each parental configuration occurs with some minimum probability. With this definition in hand, we are ready to state and prove the main theorem of this section:

Theorem 46 *There exists a computationally efficient algorithm with the following guarantees. Given as input (i) a DAG $\mathcal{S}$ with n nodes and maximum in-degree d and a known (c, C)-balanced Bayes net Q with structure $\mathcal{S}$, where $c = \tilde{\Omega}(1/\sqrt{n})$ and $C = \tilde{\Omega}(d\epsilon^2/\sqrt{n})$; (ii) a parameter $\epsilon > 0$, and (iii) sample access to an unknown Bayes net P with structure $\mathcal{S}$, the algorithm takes $O(2^{d/2}\sqrt{n}/\epsilon^2)$ samples from P, and distinguishes with probability at least $2/3$ between the cases $P = Q$ and $\|P - Q\|_1 > \epsilon$.*

We choose $m \geq \alpha \frac{2^{d/2}\sqrt{n}}{\epsilon^2}$, where $\alpha > 0$ is an absolute constant to be determined in the course of the analysis. Let $\mathcal{S}$ and Q be as in the statement of the theorem, for $c \geq \beta \frac{\log n}{\sqrt{n}} \geq \beta \frac{\log n}{m}$ and $C \geq \beta \frac{d + \log n}{m}$, for an appropriate absolute constant $\beta > 0$.

Recall that S denotes the set $\{(i, a) : i \in [n], a \in \{0,1\}^{|\text{Parents}(i)|}\}$. By assumption, we have that $|\text{Parents}(i)| \leq d$ for all $i \in [n]$. For each $(i, a) \in S$, corresponding to the parental configuration $\Pi_{i,a} = \{X_{\text{Parents}(i)} = a\}$, we define the value $N_{i,a} \stackrel{\text{def}}{=} m \Pr_Q[\Pi_{i,a}]/\sqrt{2}$. Intuitively, $N_{i,a}$ is equal to a small constant factor times the number of samples satisfying $\Pi_{i,a}$ one would expect to see among m independent samples, if the unknown distribution P were equal to Q. We will also use the notation $p_{i,a} \stackrel{\text{def}}{=} \Pr[X_i = 1 \mid X_{\text{Parents}(i)=a}]$, where $X \sim P$, and $q_{i,a} \stackrel{\text{def}}{=} \Pr[X_i = 1 \mid X_{\text{Parents}(i)=a}]$, where $X \sim Q$.

Given m independent samples $X^{(1)}, \ldots, X^{(m)}$ from a Bayes net P with structure $\mathcal{S}$, we define the estimators $Z_{i,a}, Y_{i,a}$ for every $i \in [n]$, $a \in \{0,1\}^{|\text{Parents}(i)|}$ as follows. For every (i, a) such that the number of samples $X^{(j)}$ satisfying the configuration $\Pi_{i,a}$ is between $N_{i,a}$ and $2N_{i,a}$ (that is, neither too few nor too many), we look only at the first $N_{i,a}$ such samples $X^{(j_1)}, \ldots, X^{(j_{N_{i,a}})}$, and let

$$Z_{i,a} \stackrel{\text{def}}{=} \sum_{j=1}^{N_{i,a}} \mathbf{1}_{\left\{X_i^{(j_\ell)}=1\right\}}$$

$$Y_{i,a} \stackrel{\text{def}}{=} \sum_{j=1}^{N_{i,a}} \mathbf{1}_{\left\{X_i^{(j_\ell)}=0\right\}}.$$

We note that $Z_{i,a} + Y_{i,a} = N_{i,a}$ by construction. We then define the quantity $W_{i,a}$ as

$$\frac{((1-q_{i,a})Z_{i,a} - q_{i,a}Y_{i,a})^2 + (2q_{i,a}-1)Z_{i,a} - q_{i,a}^2(Z_{i,a}+Y_{i,a})}{N_{i,a}(N_{i,a}-1)}\mathbf{1}_{N_{i,a}>1} + (p_{i,a}-q_{i,a})^2\mathbf{1}_{N_{i,a}\leq 1}.$$

On the other hand, for every (i, a) such that the number of samples $X^{(j)}$ satisfying the configuration $\Pi_{i,a}$ is less than $N_{i,a}$ or more than $2N_{i,a}$, we continue as a thought experiment and keep on getting samples until we see $N_{i,a}$ samples with the right configuration, and act as above (although the actual algorithm will stop and output reject whenever this happens). From there, we finally consider the statistic W:

$$W \stackrel{\text{def}}{=} \sum_{i=1}^{n} \sum_{a \in \{0,1\}^{|\text{Parents}(i)|}} \frac{\Pr_Q [\Pi_{i,a}]}{q_{i,a}(1 - q_{i,a})} W_{i,a} \tag{8}$$

Observe that the algorithm will output reject as soon as at least one parental configuration $\Pi_{i,a}$ was not seen enough times, or seen too many times, among the m samples.

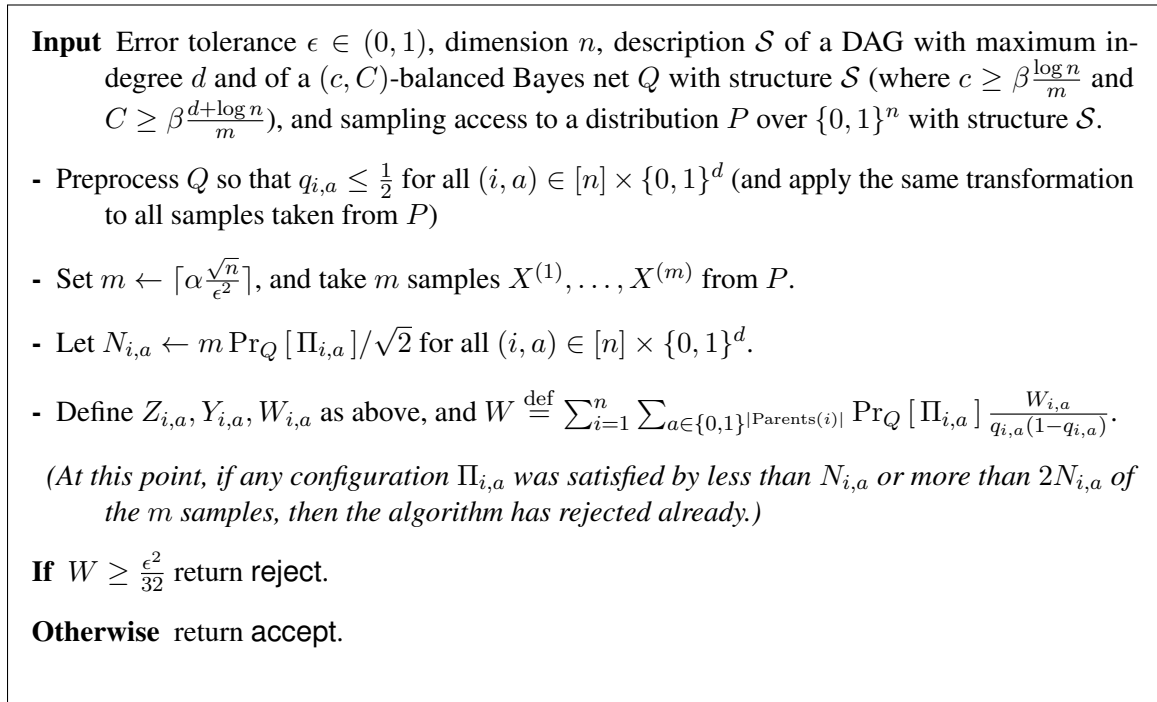The pseudocode of our algorithm is given in the following figure.

Input Error tolerance $\epsilon \in (0, 1)$, dimension n, description $\mathcal{S}$ of a DAG with maximum in-degree d and of a (c, C)-balanced Bayes net Q with structure $\mathcal{S}$ (where $c \geq \beta \frac{\log n}{m}$ and $C \geq \beta \frac{d + \log n}{m}$), and sampling access to a distribution P over $\{0, 1\}^n$ with structure $\mathcal{S}$.

- Preprocess Q so that $q_{i,a} \leq \frac{1}{2}$ for all $(i, a) \in [n] \times \{0, 1\}^d$ (and apply the same transformation to all samples taken from P)

- Set $m \leftarrow \lceil \alpha \frac{\sqrt{n}}{\epsilon^2} \rceil$, and take m samples $X^{(1)}, \ldots, X^{(m)}$ from P.

- Let $N_{i,a} \leftarrow m \Pr_Q [\Pi_{i,a}] / \sqrt{2}$ for all $(i, a) \in [n] \times \{0, 1\}^d$.

- Define $Z_{i,a}, Y_{i,a}, W_{i,a}$ as above, and $W \stackrel{\text{def}}{=} \sum_{i=1}^{n} \sum_{a \in \{0,1\}^{|\text{Parents}(i)|}} \Pr_Q [\Pi_{i,a}] \frac{W_{i,a}}{q_{i,a}(1 - q_{i,a})}$.

  (At this point, if any configuration $\Pi_{i,a}$ was satisfied by less than $N_{i,a}$ or more than $2N_{i,a}$ of the m samples, then the algorithm has rejected already.)

If $W \geq \frac{\epsilon^2}{32}$ return reject.

Otherwise return accept.

Figure 3: Testing identity against a known-structure balanced Bayes net.

Preprocessing. We will henceforth assume that $q_{i,a} \leq \frac{1}{2}$ for all $(i, a) \in [n] \times \{0, 1\}^d$. This can be done without loss of generality, as Q is explicitly known. For any i such that $q_{i,a} > \frac{1}{2}$, we replace $q_{i,a}$ by $1 - q_{i,a}$ and work with the corresponding distribution Q' instead. By flipping the corresponding bit of all samples we receive from P, it only remains to test identity of the resulting distribution P' to Q', as all distances are preserved.

First Observation. If $P = Q$, then we want to argue that with probability at least $9/10$ none of the $W_{i,a}$'s will be such that too few samples satisfied $\Pi_{i,a}$ (as this will immediately cause rejection). To see why this is the case, observe that as long as $m \Pr_Q [\Pi_{i,a}] \geq \beta(d + \log n)$ (for an appropriate choice of absolute constant $\beta > 0$), the number $m_{i,a}$ of samples satisfying $\Pi_{i,a}$ among the m we draw will, by a Chernoff bound, such that $m_{i,a} \geq m \Pr_Q [\Pi_{i,a}] \geq N_{i,a}$ with probability at least $1 - \frac{1}{2^d n} \cdot \frac{1}{10}$.

A union bound over the at most $2^d n$ possible parental configurations will yield the desired conclusion. But the fact that $P = Q$ is (c, C)-balanced indeed implies that $\Pr_Q [\, \Pi_{i,a} \,] \geq C \geq \beta \frac{d + \log n}{m}$, the last inequality by our choice of C.

Therefore, it will be sufficient to continue our analysis, assuming that none of the $W_{i,a}$'s caused rejection because of an insufficient number of samples satisfying $\Pi_{i,a}$. As we argued above, this came at the cost of only $1/10$ of probability of success in the completeness case, and can only increase the probability of rejection, i.e., success, in the soundness case.

Moreover, in the analysis of the expectation and variance of W, we assume that for every $(i, a) \in S$, we have $\Pr_P [\, \Pi_{i,a} \,] \leq 4 \Pr_Q [\, \Pi_{i,a} \,]$. This is justified by the following two lemmas, which ensure respectively that if it is not the case, then we will have rejected with high probability (this time because *too many* samples satisfied $\Pi_{i,a}$); and that we still have not rejected (with high probability) if $P = Q$.

Lemma 47 *Let P be as in the statement of Theorem [46], and suppose there exists a parental configuration $(i^*, a^*) \in S$ such that $\Pr_P [\, \Pi_{i^*, a^*} \,] > 4 \Pr_Q [\, \Pi_{i^*, a^*} \,]$. Then, with probability at least $9/10$, the number of samples m_{i^*, a^*} satisfying Π_{i^*, a^*} among the m samples taken will be more than $2 N_{i^*, a^*}$.*

Proof This follows easily from a Chernoff bound, as

$$
\Pr\left[m_{i,a} < 2m \Pr_Q [\, \Pi_{i^*, a^*} \,] \right] < \Pr\left[m_{i,a} < \frac{1}{2} m \Pr_P [\, \Pi_{i^*, a^*} \,] \right] = \Pr\left[m_{i,a} < \frac{1}{2} \mathbb{E}[m_{i,a}] \right] ,
$$

and $\mathbb{E}[m_{i,a}] > \beta(d + \log n)$. ∎

Lemma 48 *Suppose $P = Q$. Then, with probability at least $9/10$, for every parental configuration $(i, a) \in S$ the number of samples $m_{i,a}$ satisfying $\Pi_{i,a}$ among the m samples taken will be at most $2 N_{i,a}$.*

Proof This again follows from a Chernoff bound and a union bound over all $2^d n$ configurations, as we have $\Pr[\, m_{i,a} > 2m \Pr_Q [\, \Pi_{i,a} \,]\,] = \Pr[\, m_{i,a} > 2\mathbb{E}[m_{i,a}]\,]$, and $\mathbb{E}[m_{i,a}] > \beta(d + \log n)$. ∎

Expectation and Variance Analysis. We start with a simple closed form formula for the expectation of our statistic:

Lemma 49 *We have that $\mathbb{E}[W] = \sum_{i,a} \Pr_Q [\, \Pi_{i,a} \,] \frac{(p_{i,a} - q_{i,a})^2}{q_{i,a}(1 - q_{i,a})}$. (In particular, if $P = Q$ then $\mathbb{E}[W] = 0$.)*

Proof Fix any $(i, a) \in S$. Since $Z_{i,a}$ follows a $\mathrm{Bin}(N_{i,a}, p_{i,a})$ distribution, we get

$\mathbb{E}[W_{i,a}]$

$$
= \mathbb{E}\left[(Z_{i,a} - q_{i,a} N_{i,a})^2 + (2q_{i,a} - 1)Z_{i,a} - q_{i,a}^2 N_{i,a} \right] \frac{\mathbf{1}_{N_{i,a} > 1}}{N_{i,a}(N_{i,a} - 1)} + \mathbb{E}\left[(p_{i,a} - q_{i,a})^2 \right] \mathbf{1}_{N_{i,a} \leq 1}
$$
$$
= (p_{i,a} - q_{i,a})^2 \mathbf{1}_{N_{i,a} > 1} + (p_{i,a} - q_{i,a})^2 \mathbf{1}_{N_{i,a} \leq 1} = (p_{i,a} - q_{i,a})^2 ,
$$

giving the result by linearity of expectation. The last part follows from the fact that $p_{i,a} = q_{i,a}$ for all (i,a) if $P = Q$. ∎

As a simple corollary, we obtain:

Claim 50 *If $\|P - Q\|_1 \geq \epsilon$, then $\mathbb{E}[W] \geq \frac{\epsilon^2}{16}$.*

Proof The claim follows from Pinsker's inequality and Lemma 29, along with our assumption that $\Pr_P[\Pi_{i,a}] \leq 4 \cdot \Pr_Q[\Pi_{i,a}]$ for every (i,a):

$$\|P - Q\|_1^2 \leq 2\,\mathrm{D}(P\|Q) \leq 2 \sum_{(i,a)} \Pr_P[\Pi_{i,a}] \frac{(p_{i,a} - q_{i,a})^2}{q_{i,a}(1 - q_{i,a})} \leq 8 \sum_{(i,a)} \Pr_Q[\Pi_{i,a}] \frac{(p_{i,a} - q_{i,a})^2}{q_{i,a}(1 - q_{i,a})} \ .$$

∎

We now turn to bounding from above the variance of our statistic. This will be done by controlling the covariances and variances of the summands individually, and specifically showing that the former are zero. We have the following:

Claim 51 *If $(i,a) \neq (j,b)$, then $\mathrm{Cov}(W_{i,a}, W_{i,b}) = 0$; and the variance satisfies*

$$\mathrm{Var}\left[\frac{\Pr_Q[\Pi_{i,a}]}{q_{i,a}(1 - q_{i,a})} W_{i,a}\right] \leq \frac{4}{m} \Pr_Q[\Pi_{i,a}] \frac{p_{i,a}(1 - p_{i,a})}{q_{i,a}^2(1 - q_{i,a})^2}(p_{i,a} - q_{i,a})^2 + \frac{4}{m^2} \frac{p_{i,a}^2}{q_{i,a}^2(1 - q_{i,a})^2} \boldsymbol{1}_{N_{i,a} > 1} \ .$$

(Moreover, if $P = Q$ then $\mathrm{Var}\left[\frac{\Pr_Q[\Pi_{i,a}]}{q_{i,a}(1 - q_{i,a})} W_{i,a}\right] \leq \frac{4}{m^2}$.)

Proof The key point is to observe that, because of the way we defined the $Z_{i,a}$'s and $Y_{i,a}$'s (only considering the $N_{i,a}$ first samples satisfying the desired parental configuration), we have that $W_{i,a}$ and $W_{j,b}$ are independent whenever $(i,a) \neq (j,b)$. This directly implies the first part of the claim, i.e.,

$$\mathrm{Cov}(W_{i,a}, W_{i,b}) = \mathbb{E}[(W_{i,a} - \mathbb{E}[W_{i,a}])(W_{j,b} - \mathbb{E}[W_{j,b}])] = 0 \ ,$$

when $(i,a) \neq (j,b)$.

We then consider $\mathrm{Var}\left[\frac{\Pr_Q[\Pi_{i,a}]}{q_{i,a}(1 - q_{i,a})} W_{i,a}\right]$. Note that

$$\mathbb{E}\left[W_{i,a}^2\right]$$
$$= \mathbb{E}\left[((Z_{i,a} - q_{i,a}N_{i,a})^2 + (2q_{i,a} - 1)Z_{i,a} - q_{i,a}^2 N_{i,a})^2\right] \frac{\boldsymbol{1}_{N_{i,a} > 1}}{N_{i,a}^2(N_{i,a} - 1)^2} + (p_{i,a} - q_{i,a})^4 \boldsymbol{1}_{N_{i,a} \leq 1} \ ,$$

so that, writing p, q, N, Z for $p_{i,a}, q_{i,a}, N_{i,a}, Z_{i,a}$ respectively (for readability):

$$\mathrm{Var}\left[\frac{\mathrm{Pr}_Q\left[\Pi_{i,a}\right]}{q(1-q)}W_{i,a}\right]$$

$$= \left(\frac{\mathrm{Pr}_Q\left[\Pi_{i,a}\right]}{q(1-q)}\right)^2\left(\mathbb{E}\left[W_{i,a}^2\right] - \mathbb{E}[W_{i,a}]^2\right)$$

$$= \left(\frac{\mathrm{Pr}_Q\left[\Pi_{i,a}\right]}{q(1-q)}\right)^2\left(\mathbb{E}\left[W_{i,a}^2\right] - (p-q)^4\right)$$

$$= \mathbb{E}\left[((Z-qN)^2 + (2q-1)Z - q^2N)^2 - N^2(N-1)^2(p-q)^4\right]\frac{\mathrm{Pr}_Q\left[\Pi_{i,a}\right]^2\mathbf{1}_{N>1}}{N^2(N-1)^2q^2(1-q)^2}$$

$$= \frac{1}{m^2}\mathbb{E}\left[((Z-qN)^2 + (2q-1)Z - q^2N)^2 - N^2(N-1)^2(p-q)^4\right]\frac{\mathbf{1}_{N>1}}{(N-1)^2q^2(1-q)^2}$$

$$= \frac{1}{m^2}\frac{2N}{N-1}\frac{p(1-p)}{q^2(1-q)^2}\mathbf{1}_{N>1}\left((2N-3)p^2 + 2(N-1)q^2 - 4(N-1)pq + p\right)\ .$$

If $p=q$, then this becomes $\mathrm{Var}\left[\frac{\mathrm{Pr}_Q[\Pi_{i,a}]}{q(1-q)}W_{i,a}\right] = \frac{1}{m^2}\frac{2N}{N-1}\mathbf{1}_{N_{i,a}>1} \leq \frac{4}{m^2}$, providing the second part of the claim. In the general case, we can bound the variance as follows:

$$\mathrm{Var}\left[\frac{\mathrm{Pr}_Q\left[\Pi_{i,a}\right]}{q(1-q)}W_{i,a}\right] = \frac{1}{m^2}\frac{2N}{N-1}\frac{p(1-p)}{q^2(1-q)^2}\mathbf{1}_{N>1}\left(2(N-1)(p^2+q^2-2pq) - p^2 + p\right)$$

$$= \frac{1}{m^2}\frac{2N}{N-1}\frac{p(1-p)}{q^2(1-q)^2}\mathbf{1}_{N>1}\left(2(N-1)(p-q)^2 + p(1-p)\right)$$

$$= \frac{4N}{m^2}\frac{p(1-p)}{q^2(1-q)^2}(p-q)^2\mathbf{1}_{N>1} + \frac{1}{m^2}\frac{2N}{N-1}\frac{p^2(1-p)^2}{q^2(1-q)^2}\mathbf{1}_{N>1}$$

$$\leq \frac{4N}{m^2}\frac{p(1-p)}{q^2(1-q)^2}(p-q)^2 + \frac{4}{m^2}\frac{p^2(1-p)^2}{q^2(1-q)^2}\mathbf{1}_{N>1}$$

$$= \frac{4}{m}\mathrm{Pr}_Q\left[\Pi_{i,a}\right]\frac{p_{i,a}(1-p_{i,a})}{q_{i,a}^2(1-q_{i,a})^2}(p_{i,a}-q_{i,a})^2 + \frac{4}{m^2}\frac{p_{i,a}^2(1-p_{i,a})^2}{q_{i,a}^2(1-q_{i,a})^2}\mathbf{1}_{N_{i,a}>1}$$

$$\leq \frac{4}{m}\mathrm{Pr}_Q\left[\Pi_{i,a}\right]\frac{p_{i,a}(1-p_{i,a})}{q_{i,a}^2(1-q_{i,a})^2}(p_{i,a}-q_{i,a})^2 + \frac{4}{m^2}\frac{p_{i,a}^2}{q_{i,a}^2(1-q_{i,a})^2}\mathbf{1}_{N_{i,a}>1}.$$

This completes the proof. ∎

Using this claim, we now state the upper bound it allows us to obtain:

Lemma 52 *We have that* $\mathrm{Var}[W] \leq 24\frac{2^d n}{m^2} + 26\frac{\mathbb{E}[W]}{cm}$. *(Moreover, if $P = Q$ we have* $\mathrm{Var}[W] \leq 4\frac{2^d n}{m^2}$.*)*

Proof This will follow from Claim 51, which guarantees that if $P = Q$, $\mathrm{Var}[W] \leq 2^d n \cdot \frac{4}{m^2} = 4\frac{2^d n}{m^2}$. Moreover, in the general case,

$$\mathrm{Var}[W] \leq \frac{4}{m}\sum_{(i,a)}\mathrm{Pr}_Q\left[\Pi_{i,a}\right]\frac{p_{i,a}(1-p_{i,a})}{q_{i,a}^2(1-q_{i,a})^2}(p_{i,a}-q_{i,a})^2 + \frac{4}{m^2}\sum_{(i,a)}\frac{p_{i,a}^2}{q_{i,a}^2(1-q_{i,a})^2}\mathbf{1}_{N_{i,a}>1}\ .$$

We deal with the two terms separately, as follows:

- For the second term, we will show that

$$\frac{4}{m^2} \sum_{(i,a)} \frac{p_{i,a}^2}{q_{i,a}^2(1-q_{i,a})^2} \mathbf{1}_{N_{i,a}>1} \le 24\frac{2^d n}{m^2} + \frac{24\mathbb{E}[W]}{cm} \ .$$

This follows from the following sequence of (in-)equalities:

$$\sum_{(i,a)} \frac{p_{i,a}^2}{q_{i,a}^2(1-q_{i,a})^2} \mathbf{1}_{N_{i,a}>1}$$

$$= \sum_{(i,a)} \frac{(p_{i,a}-q_{i,a})^2}{q_{i,a}^2(1-q_{i,a})^2} \mathbf{1}_{N_{i,a}>1} + \sum_{(i,a)} \frac{2p_{i,a}q_{i,a}-q_{i,a}^2}{q_{i,a}^2(1-q_{i,a})^2} \mathbf{1}_{N_{i,a}>1}$$

$$= \sum_{(i,a)} \frac{(p_{i,a}-q_{i,a})^2}{q_{i,a}^2(1-q_{i,a})^2} \mathbf{1}_{N_{i,a}>1} + \sum_{(i,a)} \frac{2q_{i,a}(p_{i,a}-q_{i,a})+q_{i,a}^2}{q_{i,a}^2(1-q_{i,a})^2} \mathbf{1}_{N_{i,a}>1}$$

$$\le 4 \cdot 2^d n + \sum_{(i,a)} \frac{(p_{i,a}-q_{i,a})^2}{q_{i,a}^2(1-q_{i,a})^2} \mathbf{1}_{N_{i,a}>1} + \sum_{(i,a)} \frac{2(p_{i,a}-q_{i,a})}{q_{i,a}(1-q_{i,a})^2} \mathbf{1}_{N_{i,a}>1} \mathbf{1}_{N_{i,a}>1}$$

$$\le 4 \cdot 2^d n + \sum_{(i,a)} \frac{(p_{i,a}-q_{i,a})^2}{q_{i,a}^2(1-q_{i,a})^2} \mathbf{1}_{N_{i,a}>1} + 4 \sum_{(i,a)} \frac{p_{i,a}-q_{i,a}}{q_{i,a}(1-q_{i,a})} \mathbf{1}_{N_{i,a}>1}$$

$$\underset{\text{(AM-GM)}}{\le} 4 \cdot 2^d n + \sum_{(i,a)} \frac{(p_{i,a}-q_{i,a})^2}{q_{i,a}^2(1-q_{i,a})^2} \mathbf{1}_{N_{i,a}>1} + 2 \sum_{(i,a)} \left(1 + \frac{(p_{i,a}-q_{i,a})^2}{q_{i,a}^2(1-q_{i,a})^2}\right) \mathbf{1}_{N_{i,a}>1}$$

$$\le 6 \cdot 2^d n + 3 \sum_{(i,a)} \frac{(p_{i,a}-q_{i,a})^2}{q_{i,a}^2(1-q_{i,a})^2} \mathbf{1}_{N_{i,a}>1}$$

$$\le 6 \cdot 2^d n + \frac{6}{c} \sum_{(i,a)} \frac{(p_{i,a}-q_{i,a})^2}{q_{i,a}(1-q_{i,a})} \mathbf{1}_{N_{i,a}>1}$$

$$\le 6 \cdot 2^d n + \frac{6m}{c} \sum_{(i,a)} \frac{N_{i,a}}{m} \frac{(p_{i,a}-q_{i,a})^2}{q_{i,a}(1-q_{i,a})} \mathbf{1}_{N_{i,a}>1}$$

$$= 6 \cdot 2^d n + \frac{6m}{c} \sum_{(i,a)} \Pr_Q[\Pi_{i,a}] \frac{(p_{i,a}-q_{i,a})^2}{q_{i,a}(1-q_{i,a})} \mathbf{1}_{N_{i,a}>1} \le 6 \cdot 2^d n + \frac{6m}{c}\mathbb{E}[W] \ ,$$

using our assumption that $q_{i,a} \le \frac{1}{2}$ for all (i,a).

- For the first term, we will show that

$$\frac{4}{m} \sum_{(i,a)} \Pr_Q[\Pi_{i,a}] \frac{p_{i,a}(1-p_{i,a})}{q_{i,a}^2(1-q_{i,a})^2}(p_{i,a}-q_{i,a})^2 \le \frac{2}{cm}\mathbb{E}[W] \ .$$

This is shown as follows:

$$\sum_{(i,a)} \Pr_Q \left[\Pi_{i,a} \right] \frac{p_{i,a}(1 - p_{i,a})(p_{i,a} - q_{i,a})^2}{q_{i,a}^2 (1 - q_{i,a})^2} \leq \frac{1}{4} \sum_{(i,a)} \frac{1}{q_{i,a}(1 - q_{i,a})} \cdot \Pr_Q \left[\Pi_{i,a} \right] \frac{(p_{i,a} - q_{i,a})^2}{q_{i,a}(1 - q_{i,a})}$$

$$\leq \frac{1}{2c} \sum_{(i,a)} \Pr_Q \left[\Pi_{i,a} \right] \frac{(p_{i,a} - q_{i,a})^2}{q_{i,a}(1 - q_{i,a})}$$

$$= \frac{1}{2c} \mathbb{E}[W] \ .$$

Combining the above, we conclude that $\mathrm{Var}[W] \leq 24 \frac{2^d n}{m^2} + 26 \frac{\mathbb{E}[W]}{cm}$. ∎

We now have all the tools we require to establish the completeness and soundness of the tester.

Lemma 53 (Completeness) *If $P = Q$, then the algorithm outputs* accept *with probability at least* $2/3$.

Proof We first note that, as per the foregoing discussion and Lemma 48, with probability at least $8/10$ we have between $N_{i,a}$ and $2N_{i,a}$ samples for every parental configuration $(i, a) \in S$, and therefore have not outputted reject. By Chebyshev's inequality and Lemma 52,

$$\Pr\left[W \geq \frac{\epsilon^2}{32} \right] \leq 4096 \frac{2^d n}{m^2 \epsilon^4} \leq \frac{4}{30}$$

for a suitable choice of $\alpha > 0$. Therefore, by a union bound the algorithm will output reject with probability at most $\frac{4}{30} + \frac{2}{10} = \frac{1}{3}$. ∎

Lemma 54 (Soundness) *If $\|P - Q\|_1 \geq \epsilon$, then the algorithm outputs* reject *with probability at least* $2/3$.

Proof As noted before, it is sufficient to show that, conditioned on having between $N_{i,a}$ and $2N_{i,a}$ samples for every parental configuration and $\Pr_P \left[\Pi_{i^*,a^*} \right] \leq 4 \Pr_Q \left[\Pi_{i^*,a^*} \right]$ for all (i, a), the algorithm rejects with probability at least $2/3 + 1/10 = 23/30$. Indeed, whenever too few or too many samples from a given parental configuration are seen the algorithm rejects automatically, and by Lemma 47 this happens with probability at least $9/10$ if some parental configuration is such that $\Pr_P \left[\Pi_{i^*,a^*} \right] > 4 \Pr_Q \left[\Pi_{i^*,a^*} \right]$. Conditioning on this case, by Chebyshev's inequality,

$$\Pr\left[W \leq \frac{\epsilon^2}{32} \right] \leq \Pr\left[|W - \mathbb{E}[W]| \geq \frac{1}{2} \mathbb{E}[W] \right] \leq \frac{4 \mathrm{Var}[W]}{\mathbb{E}[W]^2} \leq 96 \frac{2^d n}{m^2 \mathbb{E}[W]^2} + 104 \frac{1}{cm \mathbb{E}[W]} \ ,$$

from Lemma 52. Since $\mathbb{E}[W] \geq \frac{\epsilon^2}{16}$ by Claim 50, we then get $\Pr\left[W \leq \frac{\epsilon^2}{32} \right] = O\left(\frac{2^d n}{m^2 \epsilon^4} + \frac{1}{cm \epsilon^2} \right) \leq \frac{17}{30}$, again for a suitable choice of $\alpha > 0$ and $\beta > 0$ (recalling that $c \geq \beta \frac{\log n}{\sqrt{n}}$). ∎

Remark 55 We note that we can reduce the problem of testing degree-d Bayes nets over alphabet Σ, to testing degree $(d+1)\lceil \log_2(|\Sigma|) \rceil - 1$ and alphabet of size 2. First consider the case where $|\Sigma| = 2^b$. Then it suffices to have nb bits in n clusters of size b. Each cluster of b will represent a single variable in the initial model with each of the 2^b possibilities denoting a single letter. Then each bit will need to potentially be dependent on each other bit in its cluster and on each bit in each cluster that its cluster is dependent on. Therefore, we need degree $(d+1)b - 1$. Note that this operation preserves balancedness.

Now if $|\Sigma|$ is not a power of 2, we need to pad the alphabet. The obvious way to do this is to create a set of unused letters until the alphabet size is a power of 2. Unfortunately, this creates an unbalanced model. To create a balanced one, we proceed as follows: we split a number of the letters in Σ in two. So, instead of having alphabet $a, b, c, \ldots$, we have $a_1, a_2, b_1, b_2, c, \ldots$. We make it so that when a word would have an a in a certain position, we map this to a new word that has either a_1 or a_2 in that position, each with equal probability. We note that this operation preserves L_1 distance, and maintains the balancedness properties.

7.2. Sample Complexity Lower Bound

Here we prove a matching information-theoretic lower bound:

Theorem 56 *There exists an absolute constant $\epsilon_0 > 0$ such that, for any $0 < \epsilon \le \epsilon_0$, the following holds: Any algorithm that has sample access to an unknown Bayes net P over $\{0,1\}^n$ with known structure $\mathcal{S}$ of maximum in-degree at most $d < n/2$, and distinguishes between the cases that $P = U$ and $\|P - U\|_1 > \epsilon$ requires $\Omega(2^{d/2} n^{1/2}/\epsilon^2)$ samples.*

Proof

Our lower bound will be derived from families of Bayes nets with the following structure: The first d nodes are all independent (and will in fact have marginal probability $1/2$ each), and will form in some sense a "pointer" to one of 2^d arbitrary product distributions. The remaining $n - d$ nodes will each depend on all of the first d. The resulting distribution is now an (evenly weighted) disjoint mixture of 2^d product distributions on the $(n - d)$-dimensional hypercube. In other words, there are 2^d product distributions $p_1, \ldots, p_{2^d}$, and our distribution returns a random i (encoded in binary) followed by a random sample form p_i. Note that the p_i can be arbitrary product distributions.

The unknown distribution P to test is obtained as follows: let X be a Bernoulli random variable with parameter $1/2$. If $X = 0$, P is the uniform distribution on $\{0,1\}^n$, i.e., each of the 2^d distributions p_i is uniform on $\{0,1\}^{n-d}$. Otherwise, if $X = 1$, then every p_i is a product distribution on $\{0,1\}^{n-d}$ with, for each coordinate, a parameter chosen uniformly and independently to be either $\frac{1}{2} + \frac{\epsilon}{\sqrt{n}}$ or $\frac{1}{2} - \frac{\epsilon}{\sqrt{n}}$.

We will show that the shared information between a sample of size $o(2^{d/2} n^{1/2}/\epsilon^2)$ and X is small. In view of this, let σ_i (for $1 \le i \le n - d$) be the set of indices of the samples that were drawn from p_i. Note that since X is uncorrelated with the σ_i's, and as the σ_i are a function of the samples, $I(X; S) = I(X; S \mid \sigma_i)$. This is because $I(X; S)) = H(X) - H(X \mid S) = H(X \mid \sigma_i) - H(X \mid S, \sigma_i) = I(X; S \mid \sigma_i)$.

Now, for fixed σ_i, the samples we draw from p_i are mutually independent of X. Let S_i denote the tuple of these $|\sigma_i|$ samples. Thus, we have that $I(X; S \mid \sigma_i) \le \sum_i I(X; S_i \mid \sigma_i)$. By the same

analysis as in the proof of Theorem 9, this latter term is $O\left(\binom{|\sigma_i|}{2}\frac{\epsilon^4}{n}\right)$. Therefore,

$$I\left(X; S \mid \sigma_i\right) \leq \mathbb{E}\left[\sum_i \binom{|\sigma_i|}{2}\right] O\left(\frac{\epsilon^4}{n}\right) = O\left(\frac{m^2\epsilon^4}{n2^d}\right),$$

where we used the fact that $|\sigma_i|$ is $\text{Bin}\left(m, 1/2^d\right)$ distributed. Note that the above RHS is $o(1)$ unless $m = \Omega(2^{d/2}n^{1/2}/\epsilon^2)$, which completes the proof. ∎

8. Testing Identity of Unknown Structure Bayes Nets

In this section, we give our algorithms and lower bounds for testing the identity of low-degree Bayes nets with unknown structure. In Section 8.1, we start by showing that – even for the case of trees – uniformity testing of n-node Bayes nets requires $\Omega(n/\epsilon^2)$ samples. In Sections 8.2, we design efficient identity testers with sample complexity sublinear in the dimension n, under some non-degeneracy assumptions on the explicit Bayes net.

8.1. Sample Complexity Lower Bound

In this section, we establish a tight lower bound on identity testing of Bayes nets in the unknown structure case. Our lower bound holds even for *balanced* Bayes nets with a *tree* structure. In order to state our theorem, we first give a specialized definition of balancedness for the case of trees. We say that a Bayes net with tree structure is *c-balanced* if it satisfies $p_k \in [c, 1-c]$ for all k (note that this immediately implies it is (c, C)-balanced).

Theorem 57 *There exists absolute constants $c > 0$ and $\epsilon_0 > 0$ such that, for any $\epsilon \in (0, \epsilon_0)$ and given samples from an unknown c-balanced Bayes net P over $\{0,1\}^n$ with unknown tree structure, distinguishing between the cases $P = U$ and $\|P - U\|_1 > \epsilon$ (where U is the uniform distribution over $\{0,1\}^n$) with probability $2/3$ requires $\Omega(n/\epsilon^2)$ samples. (Moreover, one can take $c = 1/3$.)*

Hence, without any assumptions about the explicit distribution, identity testing is information-theoretically as hard as learning. This section is devoted to the proof of Theorem 57.

Fix any integer $m \geq 1$. We will define a family of no-instances consisting of distributions $\{P_\lambda\}_\lambda$ over $\{0,1\}^n$ such that:

1. every P_λ is ϵ-far from the uniform distribution U on $\{0,1\}^n$: $\|P_\lambda - U\|_1 = \Omega(\epsilon)$;

2. every P_λ is a Bayes net with a tree structure;

3. unless $m = \Omega\left(\frac{n}{\epsilon^2}\right)$, no algorithm taking m samples can distinguish with probability $2/3$ between a uniformly chosen distribution from $\{P_\lambda\}_\lambda$ and u; or, equivalently, no algorithm taking *one* sample can distinguish with probability $2/3$ between $P_\lambda^{\otimes m}$ and $U^{\otimes m}$, when P_λ is chosen uniformly at random from $\{P_\lambda\}_\lambda$.

The family is defined as follows. We let $\delta \overset{\text{def}}{=} \frac{\epsilon}{\sqrt{n}}$, and let a *matching-orientation parameter* λ consist of (i) a matching $\lambda^{(1)}$ of $[n]$ (partition of $[n]$ in $\frac{n}{2}$ disjoint pairs (i, j) with $i < j$) and (ii) a vector $\lambda^{(2)}$ of $\frac{n}{2}$ bits. The distribution P_λ is then defined as the distribution over $\{0,1\}^n$ with uniform marginals, and tree structure with edges corresponding to the pairs $\lambda^{(1)}$; and such that for every $\lambda_k^{(1)} = (i, j) \in \lambda^{(1)}$, $\text{cov}(X_i, X_j) = (-1)^{\lambda_k^{(2)}}\delta$.

41

Notations. For $\lambda = (\lambda^{(1)}, \lambda^{(2)})$ as above and $x \in \{0,1\}^n$, we define the *agreement count of x for λ*, $c(\lambda, x)$, as the number of pairs (i,j) in $\lambda^{(1)}$ such that (x_i, x_j) "agrees" with the correlation suggested by $\lambda^{(2)}$. Specifically:

$$c(\lambda, x) \stackrel{\text{def}}{=} |\{ (i,j) \in [n]^2 : \exists \ell \in [n/2], \quad \lambda_\ell^{(1)} = (i,j) \text{ and } (-1)^{x_i + x_j} = (-1)^{\lambda_\ell^{(2)}} \}|.$$

Moreover, for λ, μ two matching-orientation parameters, we define the sets $A = A_{\lambda,\mu}, B = B_{\lambda,\mu}, C = C_{\lambda,\mu}$ as

$$A \stackrel{\text{def}}{=} \left\{ (s,t) \in [n/2]^2 : \lambda_s^{(1)} = \mu_t^{(1)}, \quad \lambda_s^{(2)} = \mu_t^{(2)} \right\} \quad \text{(common pairs with same orientations)}$$

$$B \stackrel{\text{def}}{=} \left\{ (s,t) \in [n/2]^2 : \lambda_s^{(1)} = \mu_t^{(1)}, \quad \lambda_s^{(2)} \neq \mu_t^{(2)} \right\}$$
$$\text{(common pairs with different orientations)}$$

$$C \stackrel{\text{def}}{=} (\lambda^{(1)} \cup \mu^{(1)}) \setminus (A \cup B) \qquad \text{(pairs unique to } \lambda \text{ or } \mu\text{)}$$

so that $2(|A| + |B|) + |C| = n$.

Proof of Item 1. Fix any matching-orientation parameter λ. We have

$$\|P_\lambda - U\|_1 = \sum_{x \in \{0,1\}^n} |P_\lambda(x) - U(x)| = \sum_{x \in \{0,1\}^n} |U(x)(1 + 2\delta)^{c(\lambda,x)}(1 - 2\delta)^{\frac{n}{2} - c(\lambda,x)} - U(x)|$$

$$= \frac{1}{2^n} \sum_{x \in \{0,1\}^n} |(1 + 2\delta)^{c(\lambda,x)}(1 - 2\delta)^{\frac{n}{2} - c(\lambda,x)} - 1|$$

$$= \frac{1}{2^n} \sum_{k=0}^{\frac{n}{2}} \sum_{x:\, c(\lambda,x)=k} |(1 + 2\delta)^k (1 - 2\delta)^{\frac{n}{2} - k} - 1|$$

$$= \frac{1}{2^n} \sum_{k=0}^{\frac{n}{2}} 2^{\frac{n}{2}} \binom{\frac{n}{2}}{k} |(1 + 2\delta)^k (1 - 2\delta)^{\frac{n}{2} - k} - 1|$$

$$= \sum_{k=0}^{\frac{n}{2}} \binom{\frac{n}{2}}{k} |\left(\frac{1 + 2\delta}{2}\right)^k \left(\frac{1 - 2\delta}{2}\right)^{\frac{n}{2} - k} - \frac{1}{2^{\frac{n}{2}}}|$$

$$= 2d_{\text{TV}}\left(\text{Bin}\left(\frac{n}{2}, \frac{1}{2}\right), \text{Bin}\left(\frac{n}{2}, \frac{1}{2} + \delta\right)\right) = \Omega(\epsilon),$$

where the last equality follows from Lemma 10.

Proof of Item 3. Let the distribution Q over $(\{0,1\}^n)^m$ be the uniform mixture

$$Q \stackrel{\text{def}}{=} \mathbb{E}_\lambda[P_\lambda^{\otimes m}],$$

where P_λ is the distribution on $\{0,1\}^n$ corresponding to the matching-orientation parameter λ. In particular, for any $x \in \{0,1\}^n$ we have

$$P_\lambda(x) = U(x)(1 + 2\delta)^{c(\lambda,x)}(1 - 2\delta)^{\frac{n}{2} - c(\lambda,x)}$$

with U being the uniform distribution on $\{0,1\}^n$ and $c(\lambda, x)$, the agreement count of x for λ, defined as before. Now, this leads to

$$\frac{dP_\lambda}{du}(x) = 1 + G(\lambda, x) \,,$$

where $G(\lambda, x) \stackrel{\text{def}}{=} (1-2\delta)^{\frac{n}{2}} \left(\frac{1+2\delta}{1-2\delta}\right)^{c(\lambda, x)} - 1$. For two matching-orientation parameters λ, μ, we can define the covariance $\tau(\lambda, \mu) \stackrel{\text{def}}{=} \mathbb{E}_{x \sim U}[G(\lambda, x) G(\mu, x)]$. By the minimax approach (as in (Pollard, 2003, Chapter 3)), it is sufficient to bound the L_1-distance between Q and $U^{\otimes m}$ by a small constant. Moreover, we have

$$\|Q - U^{\otimes m}\|_1 \le \mathbb{E}_{\lambda, \mu} \left[(1 + \tau(\lambda, \mu))^m\right] - 1 \tag{9}$$

and to show the lower bound it is sufficient to prove that the RHS is less than $\frac{1}{10}$ unless $m = \Omega\left(\frac{n}{\epsilon^2}\right)$. Setting $z \stackrel{\text{def}}{=} \frac{1+2\delta}{1-2\delta}$, we can derive, by expanding the definition

$$\tau(\lambda, \mu) = 1 + (1-2\delta)^n \mathbb{E}_{x \sim U}[z^{c(\lambda, x) + c(\mu, x)}] - 2(1-2\delta)^{\frac{n}{2}} \mathbb{E}_{x \sim U}[z^{c(\lambda, x)}].$$

Since, when x is uniformly drawn in $\{0,1\}^n$, $c(\lambda, x)$ follows a $\text{Bin}\left(\frac{n}{2}, \frac{1}{2}\right)$ distribution, we can compute the last term as

$$2(1-2\delta)^{\frac{n}{2}} \mathbb{E}_{x \sim U}[z^{c(\lambda, x)}] = 2(1-2\delta)^{\frac{n}{2}} \left(\frac{1+z}{2}\right)^{\frac{n}{2}} = 2(1-2\delta)^{\frac{n}{2}} \frac{1}{(1-2\delta)^{\frac{n}{2}}} = 2 \,,$$

where we used the expression of the probability-generating function of a Binomial. This leads to

$$\begin{aligned}
1 + \tau(\lambda, \mu) &= (1-2\delta)^n \mathbb{E}_{x \sim U}[z^{c(\lambda, x) + c(\mu, x)}] \\
&= (1-2\delta)^n z^{|B|} \mathbb{E}_{\alpha \sim \text{Bin}\left(|A|, \frac{1}{2}\right)}[z^{2\alpha}] \prod_{\sigma \text{ cycle}: \, |\sigma| \ge 4} \mathbb{E}_{\alpha \sim \mathcal{B}_{\lambda, \mu}(\sigma)}[z^\alpha] \,,
\end{aligned}$$

where "cycle" and the probability distribution $\mathcal{B}_{\lambda, \mu}(\sigma)$ are defined as follows. Recall that λ and μ define a weighted multigraph over n vertices, where each vertex has degree exactly 2, the edges are from the pairs $\lambda_i^{(1)}$'s and $\mu_i^{(1)}$'s, and the weights are in $\{0,1\}$ according to the $\lambda_i^{(2)}$'s and $\mu_i^{(2)}$'s. That multigraph $G_{\lambda, \mu}$ is better seen as the disjoint union of cycles (and indeed, $A \cup B$ corresponds to the cycles of length 2, while C corresponds to cycles of length at least 4).

For such a cycle σ in $G_{\lambda, \mu}$, we let $\mathcal{B}_{\lambda, \mu}(\sigma)$ be the distribution below. If the number of negative covariances – the number of edges with label $\lambda_\ell^{(2)} = 1$ or $\mu_\ell^{(2)} = 1$ – along σ is even (resp. odd), then $\mathcal{B}_{\lambda, \mu}(\sigma)$ is a $\text{Bin}\left(|\sigma|, \frac{1}{2}\right)$ conditioned on being even (resp. odd).

Instead of the above, we first consider the related quantity with the conditioning removed (indeed, as we will see in Claim 58, this new quantity is within an $1 + O(\epsilon^2)$ factor of the actual one):

$$
\begin{aligned}
1 + \tilde{\tau}(\lambda, \mu) &= (1 - 2\delta)^n z^{|B|} \mathbb{E}_{\alpha \sim \mathrm{Bin}\left(|A|, \frac{1}{2}\right)}[z^{2\alpha}] \prod_{\sigma \text{ cycle}: |\sigma| \geq 4} \mathbb{E}_{\alpha \sim \mathrm{Bin}\left(|\sigma|, \frac{1}{2}\right)}[z^{\alpha}] \\
&= (1 - 2\delta)^n z^{|B|} \mathbb{E}_{\alpha \sim \mathrm{Bin}\left(|A|, \frac{1}{2}\right)}[z^{2\alpha}] \mathbb{E}_{\alpha \sim \mathrm{Bin}\left(\sum_{\sigma: |\sigma| \geq 4} |\sigma|, \frac{1}{2}\right)}[z^{\alpha}] \\
&= (1 - 2\delta)^n z^{|B|} \mathbb{E}_{\alpha \sim \mathrm{Bin}\left(|A|, \frac{1}{2}\right)}[z^{2\alpha}] \mathbb{E}_{\alpha \sim \mathrm{Bin}\left(|C|, \frac{1}{2}\right)}[z^{\alpha}] \\
&= (1 - 2\delta)^n z^{|B|} \left(\frac{1 + z^2}{2}\right)^{|A|} \left(\frac{1 + z}{2}\right)^{|C|} \\
&= \left((1 - 2\delta)^2 z\right)^{|B|} \left((1 - 2\delta)^2 \frac{1 + z^2}{2}\right)^{|A|} \underbrace{\left((1 - 2\delta)\frac{1 + z}{2}\right)^{|C|}}_{=1} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (2|A| + 2|B| + |C| = n) \\
&= \left(1 - 4\delta^2\right)^{|B|} \left(1 + 4\delta^2\right)^{|A|}.
\end{aligned}
$$

Thus, we need to compute

$$
\begin{aligned}
\mathbb{E}_{\lambda, \mu}\left[(1 + \tilde{\tau}(\lambda, \mu))^m\right] &= \mathbb{E}_{\lambda, \mu}\left[\left(1 + 4\delta^2\right)^{m|A|} \left(1 - 4\delta^2\right)^{m|B|}\right] \\
&= \mathbb{E}_{\lambda, \mu}\left[a^{|A|} b^{|B|}\right] \qquad\qquad \left(\text{where } a \stackrel{\mathrm{def}}{=} \left(1 + 4\delta^2\right)^m, b \stackrel{\mathrm{def}}{=} \left(1 - 4\delta^2\right)^m\right) \\
&= \mathbb{E}_{\lambda, \mu}\left[\mathbb{E}\left[a^{|A|} b^{|B|} \mid |A| + |B|\right]\right] = \mathbb{E}_{\lambda, \mu}\left[b^{|A|+|B|} \mathbb{E}\left[\left(\frac{a}{b}\right)^{|A|} \mid |A| + |B|\right]\right] \\
&= \mathbb{E}_{\lambda, \mu}\left[b^{|A|+|B|} \left(\frac{1 + \frac{a}{b}}{2}\right)^{|A|+|B|}\right] \qquad\qquad \left(|A| \sim \mathrm{Bin}\left(|A| + |B|, \frac{1}{2}\right)\right) \\
&= \mathbb{E}_{\lambda, \mu}\left[\left(\frac{a + b}{2}\right)^{|A|+|B|}\right] = \mathbb{E}_{\lambda, \mu}\left[\left(\frac{\left(1 + 4\delta^2\right)^m + \left(1 - 4\delta^2\right)^m}{2}\right)^{|A|+|B|}\right].
\end{aligned}
$$

In particular, consider the following upper bound on $f(k)$, the probability that $|A| + |B| \geq k$: setting $s \stackrel{\mathrm{def}}{=} \frac{n}{2}$, for $0 \leq k \leq s$,

$$
\begin{aligned}
f(k) = \Pr[|A| + |B| \geq k] &\leq \frac{s! 2^s}{(2s)!} \cdot \binom{s}{k} \frac{(2s - 2k)!}{(s - k)! 2^{s-k}} = \frac{2^k k! \binom{s}{k}^2}{(2k)! \binom{2s}{2k}} = \frac{2^k}{k!} \frac{\binom{2(s-k)}{s-k}}{\binom{2s}{s}} \\
&= \frac{2^k}{k!} \frac{\prod_{j=0}^{k-1}(s - j)^2}{\prod_{j=0}^{2k-1}(2s - j)} = \frac{1}{k!} \frac{\prod_{j=0}^{k-1}(s - j)}{\prod_{j=0}^{k-1}(2s - 2j - 1)} \leq \frac{1}{k!}.
\end{aligned}
$$

Therefore, for any $z > 1$, we have

$$\mathbb{E}_{\lambda,\mu}\left[z^{|A|+|B|}\right] = \int_0^\infty \Pr\left[z^{|A|+|B|} \geq t\right] dt = \int_0^\infty \Pr\left[|A| + |B| \geq \frac{\ln t}{\ln z}\right] dt$$

$$= 1 + \int_1^\infty \Pr\left[|A| + |B| \geq \frac{\ln t}{\ln z}\right] dt \leq 1 + \int_1^\infty \Pr\left[|A| + |B| \geq \left\lfloor\frac{\ln t}{\ln z}\right\rfloor\right] dt$$

$$\leq 1 + \int_1^\infty \frac{dt}{\left\lfloor\frac{\ln t}{\ln z}\right\rfloor!} \leq 1 + \int_1^\infty \frac{dt}{\Gamma\left(\frac{\ln t}{\ln z}\right)} \qquad \text{(from our upper bound on } f(k))$$

$$= 1 + \int_1^\infty \frac{e^u du}{\Gamma\left(\frac{u}{\ln z}\right)}.$$

Assuming now that $1 < z \leq 1 + \gamma$ for some $\gamma \in (0,1)$, from $\ln z < \gamma$ and monotonicity of the Gamma function we obtain

$$\mathbb{E}_{\lambda,\mu}\left[z^{|A|+|B|}\right] = 1 + \int_1^\infty \frac{e^u du}{\Gamma\left(\frac{u}{\gamma}\right)} = 1 + \gamma \int_{1/\gamma}^\infty \frac{e^{\gamma v} dv}{\Gamma(v)} \leq 1 + \gamma \int_0^\infty \frac{e^v dv}{\Gamma(v)} \leq 1 + 42\gamma.$$

Suppose now $m \leq c\frac{n}{\epsilon^2} = \frac{4c}{\delta^2}$, for some constant $c > 0$ to be determined later. Then, by monotonicity

$$z \stackrel{\text{def}}{=} \frac{\left(1+4\delta^2\right)^m + \left(1-4\delta^2\right)^m}{2} \leq \frac{\left(1+4\delta^2\right)^{\frac{4c}{\delta^2}} + \left(1-4\delta^2\right)^{\frac{4c}{\delta^2}}}{2} \leq \frac{e^{16c} + e^{-16c}}{2} < 1 + \underbrace{\frac{1}{42 \cdot 20}}_{\stackrel{\text{def}}{=} \gamma}$$

for $c < \frac{3}{1000}$. Therefore,

$$\mathbb{E}_{\lambda,\mu}\left[(1 + \tilde{\tau}(\lambda,\mu))^m\right] - 1 = \mathbb{E}_{\lambda,\mu}\left[z^{|A|+|B|}\right] - 1 < \frac{1}{20},$$

as desired.

To conclude, we bound $\mathbb{E}_{\lambda,\mu}\left[(1 + \tau(\lambda,\mu))^m\right]$ combining the above with the following claim:

Claim 58 *Let $z \stackrel{\text{def}}{=} \frac{1+2\delta}{1-2\delta}$ as above. Then for any two matching-orientation parameters λ, μ, we have*

$$\prod_{\sigma:\, |\sigma| \geq 4} \mathbb{E}_{\alpha \sim \mathcal{B}_{\lambda,\mu}(\sigma)}[z^\alpha] \leq e^{\frac{8\epsilon^4}{n}} \cdot \mathbb{E}_{\alpha \sim \text{Bin}\left(\sum_{\sigma:\, |\sigma| \geq 4} |\sigma|, \frac{1}{2}\right)}[z^\alpha].$$

where $\mathcal{B}_{\lambda,\mu}(\sigma)$ is the probability distribution defined earlier.

Proof Fix λ, μ as in the statement, and any cycle σ in the resulting graph. Suppose first this is an "even" cycle:

$$\mathbb{E}_{\alpha \sim \mathcal{B}_{\lambda,\mu}(\sigma)}[z^\alpha] = \mathbb{E}_{\alpha \sim \text{Bin}\left(|\sigma|, \frac{1}{2}\right)}[z^\alpha \mid \alpha \text{ even}] = \frac{1}{1/2} \sum_{k=0}^{|\sigma|/2} \binom{|\sigma|}{2k} z^{2k} = \frac{(1+z)^{|\sigma|} + (1-z)^{|\sigma|}}{2^{|\sigma|}}.$$

Similarly, if σ is an "odd" cycle, $\mathbb{E}_{\alpha \sim \mathcal{B}_{\lambda,\mu}(\sigma)}[z^\alpha] = \frac{(1+z)^{|\sigma|} - (1-z)^{|\sigma|}}{2^{|\sigma|}}$. We then obtain $\mathbb{E}_{\alpha \sim \mathcal{B}_{\lambda,\mu}(\sigma)}[z^\alpha] \leq$
$\mathbb{E}_{\alpha \sim \mathrm{Bin}(|\sigma|,\frac{1}{2})}[z^\alpha]\left(1 + |\frac{1-z}{1+z}|^{|\sigma|}\right)$, from which

$$\prod_\sigma \mathbb{E}_{\alpha \sim \mathcal{B}_{\lambda,\mu}(\sigma)}[z^\alpha] \leq \prod_\sigma \mathbb{E}_{\alpha \sim \mathrm{Bin}(|\sigma|,\frac{1}{2})}[z^\alpha]\left(1 + |\frac{1-z}{1+z}|^{|\sigma|}\right)$$

$$= \mathbb{E}_{\alpha \sim \mathrm{Bin}(\sum_\sigma |\sigma|,\frac{1}{2})}[z^\alpha] \cdot \prod_\sigma \left(1 + |\frac{1-z}{1+z}|^{|\sigma|}\right).$$

We now bound the last factor: since $|\frac{1-z}{1+z}| = 2\delta = \frac{2\epsilon}{\sqrt{n}}$ we have at most $\frac{n}{2}$ cycles, we get

$$\prod_{\sigma:\, |\sigma| \geq 4} \left(1 + |\frac{1-z}{1+z}|^{|\sigma|}\right) = \prod_{\sigma:\, |\sigma| \geq 4} \left(1 + (2\delta)^{|\sigma|}\right) \leq \left(1 + 16\delta^4\right)^{\frac{n}{2}} \leq e^{8\frac{\epsilon^4}{n}},$$

as claimed. ∎

With this result in hand, we can get the conclusion we want: for any λ, μ,

$$1 + \tau(\lambda, \mu) = (1 - 2\delta)^n z^{|B|} \mathbb{E}_{\alpha \sim \mathrm{Bin}(|A|,\frac{1}{2})}[z^{2\alpha}] \prod_{\sigma \text{ cycle}:\, |\sigma| \geq 4} \mathbb{E}_{\alpha \sim \mathcal{B}_{\lambda,\mu}(\sigma)}[z^\alpha]$$

$$\leq e^{\frac{8\epsilon^4}{n}}(1 - 2\delta)^n z^{|B|} \mathbb{E}_{\alpha \sim \mathrm{Bin}(|A|,\frac{1}{2})}[z^{2\alpha}] \mathbb{E}_{\alpha \sim \mathrm{Bin}(\sum_{\sigma:\, |\sigma| \geq 4}|\sigma|,\frac{1}{2})}[z^\alpha] \qquad \text{(by Claim 58)}$$

$$= e^{\frac{8\epsilon^4}{n}}(1 + \tilde{\tau}(\lambda, \mu)),$$

from which

$$\mathbb{E}_{\lambda,\mu}\left[(1 + \tau(\lambda,\mu))^m\right] \leq e^{\frac{8\epsilon^4 m}{n}} \mathbb{E}_{\lambda,\mu}\left[(1 + \tilde{\tau}(\lambda,\mu))^m\right] \leq e^{\frac{8\epsilon^4 m}{n}}\left(1 + \frac{1}{20}\right) \qquad \text{(for } m \leq \frac{cn}{\epsilon^2})$$

$$\leq e^{8c\epsilon^2}\frac{21}{20} < 1 + \frac{1}{10}, \qquad\qquad \text{(as } c < \frac{3}{1000} \text{ and } \epsilon \leq 1)$$

concluding the proof: by (9), $\|Q - U^{\otimes m}\|_1 \leq \frac{1}{10}$, for any $m < c\frac{n}{\epsilon^2}$.

8.2. Identity Testing Algorithm against Non-Degenerate Bayes Nets

We start with the case of trees and then generalize to bounded degree.

8.2.1. THE CASE OF TREES

In this section, we prove our result on testing identity of a tree structured Bayes net with unknown topology. Recall from Section 8.1 that a Bayes net with tree structure is said to be c-balanced if it satisfies $p_k \in [c, 1 - c]$ for all k. We will require the following definition of *non-degeneracy* of a tree, which will be a simpler case of the definition we shall have for general Bayes nets (Theorem 66):

Definition 59 *For any $\gamma \in (0, 1]$, we say a tree Bayes net P over $\{0, 1\}^n$ is γ-non-degenerate if for all $i \in [n]$,*

$$\left|\Pr\left[X_i = 1 \mid X_{\mathrm{Parents}(i)} = 1\right] - \Pr\left[X_i = 1 \mid X_{\mathrm{Parents}(i)} = 0\right]\right| \geq \gamma$$

where $X \sim P$.

46

Roughly speaking, this definition states that the choice of the value of its parent has a significant influence on the probability of any node. With these definitions, we are ready to state and prove our result:

Theorem 60 *There exists an efficient algorithm with the following guarantees. Given as input (i) a tree $\mathcal{S}$ over n nodes and an explicit c-balanced, γ-non-degenerate Bayes net Q with structure $\mathcal{S}$, where $c, \gamma = \Omega\left(1/n^a\right)$ for some absolute constant $a > 0$; (ii) parameter $\epsilon > 0$, and (iii) sample access to a Bayes net P with unknown tree structure, the algorithm takes $O\left(\sqrt{n}/\epsilon^2\right)$ samples from P, and distinguishes with probability at least $2/3$ between $P = Q$ and $\|P - Q\|_1 > \epsilon$.*

The algorithm follows a natural idea: (1) check that the unknown distribution P indeed has, as it should, the same tree structure as the (known) distribution Q; (2) if so, invoke the algorithm of the previous section, which works under the assumption that P and Q have the same structure.

Therefore, to establish the theorem it is sufficient to show that (1) can be performed efficiently. Specifically, we will prove the following:

Theorem 61 *There exists an algorithm with the following guarantees. For $\gamma \in (0, 1)$, $c \in (0, 1/2)$, it takes as input an explicit c-balanced, γ-nondegenerate tree Bayes net Q over $\{0, 1\}^n$ with structure $\mathcal{S}(Q)$, and*

$$O\left(\frac{\log^2 \frac{1}{c}}{c^6 \gamma^4} \log n\right)$$

samples from an arbitrary tree Bayes net P over $\{0, 1\}^n$ with unknown structure $\mathcal{S}(P)$.

- *If $P = Q$, the algorithm returns* **accept** *with probability at least $4/5$;*

- *If $\mathcal{S}(P) \neq \mathcal{S}(Q)$, the algorithm returns* **reject** *with probability at least $4/5$.*

Note that the above theorem implies the desired result as long as $\frac{\log^2 \frac{1}{c}}{c^6 \gamma^4} = O\left(\frac{\sqrt{n}}{\epsilon^2 \log n}\right)$.

Proof [Proof of Theorem 61] We start by stating and proving lemmas that will be crucial in stating and analyzing the algorithm:

Fact 62 *Given $\tau > 0$ and sample access to a tree Bayes net P over $\{0, 1\}^n$, one can obtain with $O(\frac{\log n}{\tau^2})$ samples estimates $(\hat{\mu}_i)_{i \in [n]}$, $(\hat{\rho}_{i,j})_{i,j \in [n]}$ such that, with probability at least $9/10$,*

$$\max\left(\max_{i \in [n]}|\hat{\mu}_i - \mathbb{E}[X_i]|, \max_{i, \in [n]}|\hat{\rho}_{i,j} - \mathbb{E}[X_i X_j]|\right) \leq \tau.$$

Proof The fact follows immediately by an application of Chernoff bounds. ∎

Lemma 63 *Let $c \in (0, 1/2]$. There exists a constant λ and a function f such that*

$$I\left(X_i; X_j\right) = f(\mathbb{E}[X_i], \mathbb{E}[X_j], \mathbb{E}[X_i X_j]),$$

for any c-balanced tree Bayes net P over $\{0, 1\}^n$ and $X \sim P$, where f is λ-Lipschitz with respect to the $\|\cdot\|_\infty$ norm on the domain $\Omega_c \subseteq [0, 1] \times [0, 1] \times [0, 1] \to [0, 1]$ in which $(\mathbb{E}[X_i], \mathbb{E}[X_j], \mathbb{E}[X_i X_j])_{i,j}$ then take values. Moreover, one can take $\lambda = 16 \log \frac{1}{c}$.

47

Proof [Proof Sketch:] Expanding the definition of mutual influence $I(X;Y)$ of two random variables, it is not hard to write it as a function of $\mathbb{E}[X]$, $\mathbb{E}[Y]$, and $\mathbb{E}[XY]$ only. This function would not be Lipschitz on its entire domain, however. The core of the proof leverages the balancedness assumption to restrict its domain to a convenient subset $\Omega_c \subseteq [0,1] \times [0,1] \times [0,1]$, on which it becomes possible to bound the partial derivatives of f. We defer the details of the proof to Appendix D. ∎

We now show the following crucial lemma establishing the following result: For any balanced Bayes net, the shared information between any pair of non-adjacent vertices i, j is noticeably smaller than the minimum shared information between any pair of neighbors along the path that connects i, j.

Lemma 64 *Let $c \in (0, 1/2]$, and fix any c-balanced tree Bayes net P over $\{0,1\}^n$ with structure $\mathcal{S}(Q)$. Then, for any distinct $i, j \in [n]$ such that $i \neq \mathrm{Parents}(j)$ and $j \neq \mathrm{Parents}(i)$, we have*

$$I(X_i; X_j) \leq (1 - 2c^2) \min_{\{k, \mathrm{Parents}(k)\} \in \mathrm{path}(i,j)} I\left(X_k; X_{\mathrm{Parents}(k)}\right) ,$$

where $X \sim P$. (and $\mathrm{path}(i,j)$ is a path between i to j, of the form $i - \cdots - k - \cdots - j$, where each edge is of the form $(k, \mathrm{Parents}(k)$ or $(\mathrm{Parents}(k), k))$.)

Proof By induction and the data processing inequality, it is sufficient to prove the statement for a path of length 3, namely

$$X_i - X_k - X_j .$$

The result will follow from a version of the strong data processing inequality (see e.g., Polyanskiy and Wu (2015), from which we borrow the notations $\eta_{\mathrm{KL}}, \eta_{\mathrm{TV}}$): since $X_i \to X_k \to X_j$ forms a chain with the Markov property, we get $I(X_i; X_j) \leq \eta_{\mathrm{KL}}(P_{X_j|X_k}) I(X_i; X_k)$ from (Polyanskiy and Wu, 2015, Equation 17). Now, by (Polyanskiy and Wu, 2015, Theorem 1), we have

$$\eta_{\mathrm{KL}}(P_{X_j|X_k}) \leq \eta_{\mathrm{TV}}(P_{X_j|X_k}) = d_{\mathrm{TV}}(P_{X_j|X_k=0}, P_{X_j|X_k=1}).$$

If $k = \mathrm{Parents}(j)$ (in our Bayes net), then $d_{\mathrm{TV}}(P_{X_j|X_k=0}, P_{X_j|X_k=1}) = |p_{j,0} - p_{j,1}| \leq 1 - 2c$ from the c-balancedness assumption. On the other hand, if $j = \mathrm{Parents}(k)$, then by Bayes' rule it is easy to check that (again, from the c-balancedness assumption) $\Pr[X_{\mathrm{Parents}(k)} = 1 \mid X_k = a] \in [c^2, 1-c^2]$, and $d_{\mathrm{TV}}(P_{X_j|X_k=0}, P_{X_j|X_k=1}) = |\Pr[X_j = 1 \mid X_k = 0] - \Pr[X_j = 1 \mid X_k = 1]| \leq 1 - 2c^2$.

Therefore, we get $I(X_i; X_j) \leq (1 - 2c^2) I(X_i; X_k)$ as wanted; by symmetry, $I(X_i; X_j) \leq (1 - 2c^2) I(X_j; X_k)$ holds as well. ∎

Lemma 65 *Let $c \in (0, 1/2], \gamma \in (0, 1)$, and fix any c-balanced, γ-nondegenerate tree Bayes net P over $\{0,1\}^n$, with structure $\mathcal{S}(P)$. Then, there exists an absolute constant κ such that for any $i \in [n]$ one has*

$$I\left(X_i; X_{\mathrm{Parents}(i)}\right) \geq \kappa ,$$

where $X \sim P$. (Moreover, one can take $\kappa = \frac{c\gamma^2}{2\ln 2}$.)

Proof Fix any such i, and write $X = X_i$, $Y = X_{\text{Parents}(i)}$ for convenience; and set $u \overset{\text{def}}{=} \Pr[X = 1]$, $v \overset{\text{def}}{=} \Pr[Y = 1]$, $a \overset{\text{def}}{=} \Pr[X = 1 \mid Y = 1]$, and $b \overset{\text{def}}{=} \Pr[X = 1 \mid Y = 0]$. We then have

$$
\begin{aligned}
I(X;Y) &= \sum_{(x,y) \in \{0,1\}^2} \Pr[X = x, Y = y] \log \frac{\Pr[X = x, Y = y]}{\Pr[X = x]\Pr[Y = y]} \\
&= \sum_{(x,y) \in \{0,1\}^2} \Pr[X = x \mid Y = y]\Pr[Y = y] \log \frac{\Pr[X = x \mid Y = y]}{\Pr[X = x]} \\
&= (1-u)(1-b) \log \frac{1-b}{1-v} + (1-u)b \log \frac{b}{v} + u(1-a) \log \frac{1-a}{1-v} + ua \log \frac{a}{v} \\
&= u\varphi(a,v) + (1-u)\varphi(b,v) \,,
\end{aligned}
$$

where $\varphi(x,y) \overset{\text{def}}{=} x \log \frac{x}{y} + (1-x) \log \frac{1-x}{1-y} \geq 0$ for $x, y \in [0,1]$ is the KL-divergence between two Bernoulli distributions with parameters x, y. From our assumptions of c-balanced and γ-nondegeneracy, we know that u, v, a, b satisfy

$$
\begin{aligned}
c &\leq a, b, u, v \leq 1 - c \\
\gamma &\leq |a - b| \,,
\end{aligned}
$$

which leads to, noticing that $|a - b| \geq \gamma$ implies that at least one of $|a - v| \geq \frac{\gamma}{2}$, $|b - v| \geq \frac{\gamma}{2}$ holds and that $\varphi(\cdot, v)$ is convex with a minimum at v:

$$
I(X;Y) \geq c\left(\varphi(a,v) + \varphi(b,v)\right) \geq c \min\left(\varphi\left(v - \frac{\gamma}{2}, v\right), \varphi\left(v + \frac{\gamma}{2}, v\right)\right) \geq \frac{1}{2\ln 2}c\gamma^2 \,,
$$

using the standard lower bound of $\varphi(x,y) \geq \frac{2}{\ln 2}(x - y)^2$ on the KL-divergence. ∎

The Algorithm. With these in hand, we are ready to describe and analyze the algorithm underlying Theorem 61:

Let $\gamma \in (0,1)$, $c \in (0, 1/2)$ be fixed constants, and Q be a known c-balanced, γ-nondegenerate tree Bayes net over $\{0,1\}^n$, with structure $\mathcal{S}(Q)$. Furthermore, let P be an unknown tree Bayes net over $\{0,1\}^n$ with unknown structure $\mathcal{S}(P)$, to which we have sample access.

Let $\kappa = \kappa(c, \gamma) = \frac{c\gamma^2}{2\ln 2}$ as in Lemma 65, $c' \overset{\text{def}}{=} \frac{c}{2}$, and $\lambda = \lambda(c') = 16 \log \frac{2}{c}$ as in Lemma 63. In view of applying Lemma 64 later to P, set

$$
\tau \overset{\text{def}}{=} \frac{\kappa - (1 - 2c'^2)\kappa}{4\lambda} = \frac{1}{64\ln 2} \frac{c^3\gamma^2}{\log \frac{2}{c}}.
$$

The algorithm then proceeds as follows. (Below, X denotes a random variable distributed according to P.)

1. Take $m = O\left(\frac{\log n}{\tau^2}\right)$ samples from P, and use them to

   - Estimate all n^2 marginals $\Pr[X_i = 1 \mid X_j = a]$, and verify that they are all in $[c', 1 - c']$ (ensuring that P is c'-balanced), with probability at least $9/10$. Else, return reject;

49

- Estimate all $\binom{n}{2} + n$ values of $\mathbb{E}[X_i]$ and $\mathbb{E}[X_i X_j]$ to an additive τ, with probability at least $9/10$, as in Fact 62. (Call these estimates $\hat{\mu}_i$, $\hat{\rho}_{i,j}$.)

At the end of this step, we are guaranteed that P is c'-balanced (or else we have rejected with probability at least $9/10$).

2. Check that all $\hat{\mu}_i$, $\hat{\rho}_{i,j}$'s are all within an additive τ of what they should be under Q. If so, return **accept**, else return **reject**.

Clearly, the algorithm only uses $O\left(\frac{\log^2 \frac{1}{c}}{c^6 \gamma^4} \log n \right)$ samples from P. We now establish its correctness: first, with probability at least $4/5$ by a union bound, all estimates performed in the first step are correct; we condition on that.

Completeness. If $P = Q$, then P is c-balanced, and thus *a fortiori* c'-balanced: the algorithm does not reject in the first step. Moreover, clearly all $(\hat{\mu}_i)_i$, $(\hat{\rho}_{i,j})_{i,j}$ are then within an additive τ of the corresponding values of $P = Q$, so the algorithm returns **accept**.

Soundness. By contrapositive. If the algorithm returns **accept**, then P is c'-balanced by the first step. Given our setting of τ, by Lemma 63 our estimates $(\hat{\mu}_i)_i$, $(\hat{\rho}_{i,j})_{i,j}$ are such that all corresponding quantities

$$\hat{I}_{i,j} \stackrel{\text{def}}{=} f(\hat{\mu}_i, \hat{\mu}_j, \hat{\rho}_{i,j})$$

are within $\tau\lambda = \frac{\kappa - (1 - 2c'^2)\kappa}{4}$ of the mutual informations $I(X_i; X_j)$ for P. But then, by Lemma 64 this implies that the relative *order* of all $\hat{I}_{i,j}$, $\hat{I}_{i',j'}$ is the same as the relative order of $I(X_i; X_j), I(X_{i'}; X_{j'})$. This itself implies that running the Chow–Liu algorithm on input these $\hat{I}_{i,j}$'s would yield the same, uniquely determined tree structure $\mathcal{S}(P)$ as if running it on the actual $I(X_i; X_j)$'s. To see this, we note that the Chow-Liu algorithm works by computing a maximum-weight spanning tree (MST) with respect to the weights given by the pairwise mutual information. The claim follows from the fact that the MST only depends on the relative ordering of the edge-weights.

But since the $(\hat{\mu}_i)_i$, $(\hat{\rho}_{i,j})_{i,j}$ are also within an additive τ of the corresponding quantities for Q (per our check in the second step), the same argument shows that running the Chow–Liu algorithm would result in the same, uniquely determined tree structure $\mathcal{S}(Q)$ as if running it on the actual mutual informations from Q. Therefore, $\mathcal{S}(P) = \mathcal{S}(Q)$, concluding the proof. ∎

8.2.2. THE CASE OF BOUNDED DEGREE

In this section, we show how to test identity of unknown structure Bayesian networks with maximum in-degree d under some non-degeneracy conditions. Intuitively, we want these conditions to ensure *identifiability* of the structure: that is, that any (unknown) Bayes net close to a non-degenerate Bayes net Q must also share the same structure. To capture this notion, observe that, by definition, non-equivalent Bayes net structures satisfy different conditional independence constraints: our non-degeneracy condition is then to rule out some of these possible new conditional independence constraints, as *far* from being satisfied by the non-degenerate Bayes net. Formally, we have the following definition:

Definition 66 (Non-degeneracy) *For nodes X_i, X_j, set of nodes S, and a distribution P over $\{0,1\}^n$, we say that X_i and X_j are γ-far from independent conditioned on X_S if for all distributions Q over $\{0,1\}^n$ such that $d_{\mathrm{TV}}(P,Q) < \gamma$, it holds that X_i and X_j are not independent conditioned on X_S.*

A Bayes net P is then called γ-non-degenerate with respect to structure $\mathcal{S}$ and degree d if for any nodes X_i, X_j and set of nodes S of size $|S| \leq d$ not containing i or j satisfying one of the following:

(i) X_i is a parent of X_j,

(ii) S contains a node X_k that is a child of both X_j and X_j,

(iii) X_i is a grandparent of X_j and there is a child of X_i and parent of X_j, X_k, that is not in S,

(iv) X_i and X_j have a common parent X_k that is not in S

we have that X_i and X_j are γ-far from independent conditioned on X_S (where all relations are under structure $\mathcal{S}$).
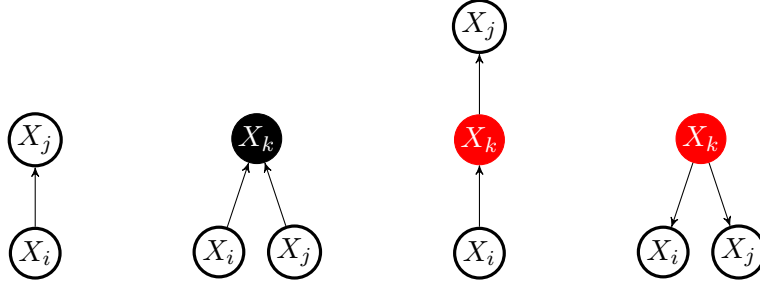
Figure 4: The four possible conditions of Theorem 66, from left (i) to right (iv). The black nodes are the ones in S, the red ones (besides X_i, X_j) are not in S.

We shall also require some terminology: namely, the definition of the *skeleton* of a Bayesian network as the underlying undirected graph of its structure. We can now state the main result of this section:

Theorem 67 *There exists an algorithm with the following guarantees. Given the full description of a Bayes net Q of degree at most d which is (c, C) balanced and γ-non-degenerate for $c = \tilde{\Omega}\left(1/\sqrt{n}\right)$ and $C = \tilde{\Omega}\left(d\epsilon^2/\sqrt{n}\right)$, parameter $\epsilon \in (0,1]$, and sample access to a distribution P, promised to be a Bayes net of degree at most d whose skeleton has no more edges than Q's, the algorithm takes $O\left(2^{d/2}\sqrt{n}/\epsilon^2 + (2^d + d\log n)/\gamma^2\right)$ samples from P, runs in time $O(n)^{d+3}(1/\gamma^2 + 1/\epsilon^2)$, and distinguishes with probability at least $2/3$ between (i) $P = Q$ and (ii) $\|P - Q\|_1 > \epsilon$.*

In Lemma 71, we show that these non-degeneracy conditions are enough to ensure identifiability of the structure, up to equivalence. In Theorem 73, we give a test for conditional independence specialized to Bernoulli random variables. In the last part, we provide a test for showing whether a non-degenerate Bayes net has a given structure using this conditional independence test, establishing Theorem 68. We then can combine this structure test with our test for Bayes nets with known structure to obtain Theorem 67. This structure tester, which may be of independent interest, has the following guarantees,

Theorem 68 *Let $\mathcal{S}$ be a structure of degree at most d and P be a Bayesian network with structure $\mathcal{S}'$ that also has degree at most d and whose skeleton has no more edges than $\mathcal{S}$. Suppose that P either (i) can be expressed as a Bayesian network with structure $\mathcal{S}$ that is γ-non-degenerate with degree d; or (ii) cannot be expressed as a Bayesian network with structure $\mathcal{S}$. Then there is an algorithm which can decide which case holds with probability $99/100$, given $\mathcal{S}$, γ, and sample access to P. The algorithm takes $O\left((2^d + d\log n)/\gamma^2\right)$ samples and runs in time $O\left(n^{d+3}/\gamma^2\right)$.*

Using the above theorem, we can prove the main result of this section:

Proof [Proof of Theorem 67] We first invoke the structure test given in Fig. 6. If it accepts, we run the known structure test given in Theorem 46. We accept only if both accept.

The correctness and sample complexity now both follow from Theorem 68 and Theorem 46. Specifically, if the structure test accepts, then with high probability, we have that Q can be expressed as a Bayes net with the same structure as P, and thus we have the pre-conditions for the known structure test. If either test rejects, then $P \neq Q$. ∎

Non-degeneracy and Equivalent Structures. The motivation behind the γ-non-degeneracy condition is the following: if Q is γ-non-degenerate, then for any Bayesian network P with degree at most d that has $d_{\mathrm{TV}}(P, Q) < \gamma$ we will argue that P can be described using the same structure $\mathcal{S}$ as we are given for Q. Indeed, the structure $\mathcal{S}'$ of P will have the property that $\mathcal{S}$ and $\mathcal{S}'$ both can be used to describe the same Bayesian networks, a property known as *I-equivalence*. It will then remain to make this algorithmic, that is to describe how to decide whether P can be described with the same structure as Q or whether $d_{\mathrm{TV}}(P, Q) \geq \gamma$. Assuming we have this decision procedure, then if the former case happens to hold we can invoke our existing known-structure tester (or reject if the latter case holds).

We will require for our proofs the following definition:

Definition 69 ($\vee$-**structure**) *For a structure $\mathcal{S}$, a triple (i, j, k) is a $\vee$-structure (also known as an immorality) if i and j are parents of k but neither i nor j is a parent of the other.*

The following result, due to Verma and Pearl Verma and Pearl (1991), will play a key role:

Lemma 70 *Two structures $\mathcal{S}$ and $\mathcal{S}'$ are I-equivalent if and only if they have the same skeleton and the same $\vee$-structures.*

Note that, for general structures $\mathcal{S}$, $\mathcal{S}'$, it may be possible to represent all Bayesian networks with structure $\mathcal{S}$ as ones with structure $\mathcal{S}'$, but not vice versa. Indeed, this can easily be achieved by adding edges to $\mathcal{S}$ to any node (if any) with less than d parents. This is the rationale for the assumption in Theorem 68 that $\mathcal{S}'$ has no more edges than $\mathcal{S}$: as this assumption is then required for $\mathcal{S}$ and $\mathcal{S}'$ to be I-equivalent unless $d_{\mathrm{TV}}(P, Q) \geq \gamma$.

We now prove that any Bayesian network Q satisfiying the conditions of Theorem 68 and being non-degenerate with respect to a structure can in fact be expressed as having that structure.

Lemma 71 *Fix $\gamma > 0$. If Q is a Bayesian network with structure $\mathcal{S}'$ of degree at most d that is γ-non-degenerate with respect to a structure $\mathcal{S}$ with degree at most d and $\mathcal{S}'$ has no more edges than $\mathcal{S}$, then $\mathcal{S}$ and $\mathcal{S}'$ are I-equivalent.*

Note that Q being γ-non-degenerate for some $\gamma > 0$ is equivalent to a set of conditional independence conditions all being false, since if X_i and X_j are not conditionally independent with respect to X_S, then there is a configuration a such that $\Pr_Q[X_S = a] > 0$ and $I(X_i; X_j \mid X_S = a) \geq 0$.

Proof We first show that $\mathcal{S}$ and $\mathcal{S}'$ have the same skeleton and then that they have the same $\vee$-structures. We need the following:

Claim 72 *Let S be the set of parents of X_i in a Bayesian network Q with structure $\mathcal{S}$. Let X_j be a node that is neither in S nor a descendant of X_i. Then X_i and X_j are independent conditioned on X_S.*

Proof Firstly, we note that there is a numbering of the nodes which is consistent with the DAG of $\mathcal{S}$ such that any $j \in S$ has $j < i$. Explicitly, we can move X_i and all its descendants to the end of the list of nodes to obtain this numbering.

Letting $D \stackrel{\text{def}}{=} \{1, \ldots, i-1\}$, we have that, from the definition of Bayesian networks, $\Pr_Q[X_i = 1 \mid X_D = b] = \Pr_Q[X_i = 1 \mid X_S = b_S]$ for all configurations b of D. Then for any configuration a of $S' \stackrel{\text{def}}{=} S \cup \{j\}$, we have

$$\Pr_P[X_j = 1 \mid X_{S'} = a] = \sum_{b:b_S=a} \Pr_P[X_j = 1 \mid X_D = b] \Pr_P[X_D = b \mid X_{S'} = a]$$

$$= \Pr_P[X_j = 1 \mid X_S = a_S] \sum_{b:b_S=a} \Pr_P[X_D = b \mid X_{S'} = a]$$

$$= \Pr_P[X_j = 1 \mid X_S = a_S]$$

concluding the proof. ■

Suppose for a contradiction that (i, j) is an edge in the skeleton of $\mathcal{S}$ but not in $\mathcal{S}'$. Without loss of generality, we may assume that X_j is not a descendant of X_i in $\mathcal{S}'$ (since otherwise we can swap the roles of i and j in the argument). Then as X_i is not in S, the set of parents of X_j in $\mathcal{S}'$, either, by Claim 72 X_i and X_j are independent conditioned on X_S. However since one of X_i and X_j is a parent of the other in $\mathcal{S}$, condition (i) of γ-non-degeneracy gives that X_i and X_j are γ-far from independent conditioned on X_S. This is a contradiction, so all edges in the skeleton of $\mathcal{S}$ must be edges of $\mathcal{S}'$. But by assumption $\mathcal{S}'$ has no more edges than $\mathcal{S}$, and so they have the same skeleton.

Next we show that i and j have the same $\vee$-structures. Assume by contradiction that (i, j, k) is a $\vee$-structure in $\mathcal{S}$ but not $\mathcal{S}'$. Since $\mathcal{S}$ and $\mathcal{S}'$ have the same skeleton, this cannot be because X_i is the parent of X_j or vice versa. Therefore, must be that at least one of X_i or X_j is the child of X_k rather than its parent in $\mathcal{S}'$. As before, without loss of generality we may assume that X_j is not a descendant of X_i in $\mathcal{S}'$. This implies that X_k cannot be a child of X_i, since then X_j must be a child of X_k and so a descendant of X_i. Thus S, the set of parents of X_i in $\mathcal{S}'$, contains X_k but not X_j; and Claim 72 then implies that X_i and X_j are independent conditioned on X_S. However, in $\mathcal{S}$ X_k is the child of both X_i and X_j and so by condition (ii) of γ-non-degeneracy, we have that X_i and X_j are γ-far from independent conditioned on X_S. This contradiction shows that all $\vee$-structures in $\mathcal{S}$ are $\vee$-structures in $\mathcal{S}'$ as well.

Finally, we assume for the sake of contradiction that (i, j, k) is a $\vee$-structure in $\mathcal{S}'$ but not $\mathcal{S}$. Again without loss of generality, we assume that X_j is not a descendant of X_i in $\mathcal{S}'$; and let S be the

parents of X_i in $\mathcal{S}'$. Note that neither X_k nor X_j is in S since this is a $\vee$-structure. Now by Claim 72, X_i and X_j are independent conditioned on X_S. In $\mathcal{S}$, however, (i, j, k) is not a $\vee$-structure yet (i, k), (j, k) (but not (i, j)) are in the skeleton of $\mathcal{S}$. Thus at least one of X_i, X_j is a child of X_k. If only one is a child, then the other must be X_k's parent. In the case of two children, we apply condition (iv) and in the case of a parent and a child, we apply condition (iii) of γ-non-degeneracy. Either way, we obtain that, since X_k is not in S, X_i and X_j are γ-far from independent conditioned on X_S. This contradiction shows that all $\vee$-structures in $\mathcal{S}'$ are also $\vee$-structures in $\mathcal{S}$.

We thus have all the conditions for Lemma 70 to apply and conclude that $\mathcal{S}$ and $\mathcal{S}'$ are I-equivalent. ∎

Conditional Independence Tester. We now turn to establishing the following proposition:

Proposition 73 *There exists an algorithm that, given parameters $\gamma, \tau > 0$, set of coordinates $S \subseteq [n]$ and coordinates $i, j \in [n] \setminus S$, as well as sample access to a distribution P over $\{0, 1\}^n$, satisfies the following. With probability at least $1 - \tau$, the algorithm accepts when X_i and X_j are independent conditioned on X_S and rejects when no distribution Q with $d_{\mathrm{TV}}(P, Q) < \gamma$ has this property (and may do either if neither cases holds). Further, the algorithm takes $O((2^d + \log(1/\tau))/\gamma^2)$ samples from P and runs in time $O((2^d + \log(1/\tau))/\gamma^2)$.*

Input $\gamma, \tau > 0$, $i, j \in \{0, 1\}^n$, $S \subseteq \{0, 1\}^n$ with $i, j \notin S$, and sample access to a distribution P on $\{0, 1\}^n$.

- Take $O((2^d + \log(1/\tau))/\gamma^2)$ samples from P. Let $\tilde{P}$ be the resulting empirical distribution.

For each configuration $a \in \{0, 1\}^{|S|}$ of S,

- Compute the empirical conditional means $\mu_{i,a} = \mathbb{E}_{X \sim \tilde{P}}[X_i \mid X_S = a]$ and $\mu_{j,a} = \mathbb{E}_{X \sim \tilde{P}}[X_j \mid X_S = a]$.
- Compute the conditional covariance $\mathrm{Cov}_{\tilde{P}}[X_i, X_j \mid X_S = a] = \mathbb{E}_{X \sim \tilde{P}}[(X_i - \mu_{i,a})(X_j - \mu_{j,a}) \mid X_S = a]$.

Compute the expected absolute value of the conditional covariance $\beta = \mathbb{E}_{Y \sim \tilde{P}}[|\mathrm{Cov}_{\tilde{P}}[X_i, X_j \mid X_S = Y_S]|]$.

If $\beta \leq \gamma/3$, return accept

Else return reject.

Figure 5: Testing whether X_i and X_j are independent conditioned on S or are γ-far from being so.

Proof The algorithm is given in Fig. 5. Its sample complexity is immediate, and that the algorithm takes linear time in the number of samples is not hard to see. It remains to prove correctness.

To do so, define $D \stackrel{\text{def}}{=} S \cup \{i, j\}$. Let $P_D, \tilde{P}_D$ be the distributions of X_S for X distributed as $P, \tilde{P}$ respectively. Since P_D is a discrete distribution with support size 2^{d+2}, by standard results the

empirical $\tilde{P}_D$ obtained from our $O((2^{d+2} + \log 1/\tau)/\gamma^2)$ samples is such that $d_{TV}(P_D, \tilde{P}) \leq \gamma/10$ with probability at least $1 - \tau$. We hereafter assume that this holds.

Note that the distribution P_D determines whether P is such that X_i and X_j are independent conditioned on S or is δ-far from being so for any δ. Thus if these two nodes are γ-far from being conditionally independent in P, then they are γ-far in P_D and therefore are $9\gamma/10$-far in $\tilde{P}_D$. We now need to show that the expected absolute value of the conditional covariance is a good approximation of the distance from conditional independence, which is our next claim:

Claim 74 *For a distribution Q on $\{0,1\}^n$, let γ be the minimum $\gamma > 0$ such that X_i and X_j are γ-far from independent conditioned on X_S in Q. Let $\beta = \mathbb{E}_{Y \sim Q}[|\mathrm{Cov}_Q[X_i, X_j \mid X_S = Y_S]|]$. Then we have $\beta/3 \leq \gamma \leq 2\beta$.*

Proof For simplicity, we assume that $|D| = n$ and that we have only coordinates i, j and S.

Firstly, we show that $\beta \leq \gamma$. By assumption, there is a distribution R with $d_{TV}(Q, R) = \gamma$ which has that X_i and X_j are independent conditioned on X_S. Thus R has $|\mathrm{Cov}_R[X_i, X_j \mid X_S = a]| = 0$ for all configurations a. Since $0 \leq |\mathrm{Cov}_Q[X_i, X_j \mid X_S = a]| \leq 1$, it follows that $|\beta - \mathbb{E}_{Y \sim R}[|\mathrm{Cov}_Q[X_i, X_j \mid X_S = Y_S]|]| \leq 3d_{TV}(Q, R)$ as $\mathrm{Cov}_Q[X_i, X_j \mid X_S = Y_S] = \mathbb{E}[X_i X_j \mid X_S = Y_S] - \mathbb{E}[X_i \mid X_S = Y_S]\mathbb{E}[X_j \mid X_S = Y_S]$ and so $\beta \leq 3\gamma$.

Next we need to show that $\beta \leq 2\gamma$. To show this, we construct a distribution S on $\{0,1\}^n$ with $d_{TV}(Q, S) = 2\beta$ in which X_i and X_j are independent conditioned on X_S. Explicitly, for a configuration a of S and $b, c \in \{0, 1\}$, we set

$$\Pr_S[X_S = a, X_i = b, X_j = c]$$

$$\stackrel{\text{def}}{=} \Pr_Q[X_S = a, X_i = b, X_j = c] + (-1)^{b+c}\mathrm{Cov}_Q[X_i, X_j \mid X_S = a]\Pr_Q[X_S = a].$$

For each configuration a, this increases two probabilities by $|\mathrm{Cov}_Q[X_i, X_j \mid X_S = a]|\Pr_Q[X_S = a]$ and decrease two probabilities by the same amount. Thus, provided that all probabilities are still non-negative (which we show below), S is a distribution with $d_{TV}(Q, S) = \sum_a 2|\mathrm{Cov}_Q[X_i, X_j \mid X_S = a]|\Pr_Q[X_S = a] = 2\beta$.

Now consider the conditional joint distribution of X_i, X_j for a given a. Let $p_{b,c} \stackrel{\text{def}}{=} \Pr_Q[X_i = b, X_j = c \mid X_S = a]$. Then the conditional covariance $\mathrm{Cov}_Q[X_i, X_j \mid X_S = a]$, which we denote by α for simplicity here, is

$$\alpha = \mathbb{E}[X_i X_j \mid X_S = a] - \mathbb{E}[X_i \mid X_S = a]\mathbb{E}[X_j \mid X_S = a]$$
$$= p_{1,1} - (p_{1,0} + p_{1,1})(p_{0,1} + p_{1,1})$$
$$= p_{1,1}(1 - p_{1,0} - p_{0,1} - p_{1,1}) - p_{1,0}p_{0,1}$$
$$= p_{1,1}p_{0,0} - p_{1,0}p_{0,1} .$$

In S, these probabilities change by α. $p_{1,1}$ and $p_{0,0}$ are increased by α and $p_{0,1}$ and $p_{1,0}$ are decreased by it. Note that if $\alpha > 0$, $p_{1,1}$ and $p_{0,0}$ are at least $p_{1,1}p_{0,0} \geq \alpha$ and when $\alpha < 0$, $p_{0,1}$ and $p_{1,0}$ are at least $p_{1,0}p_{0,1} \geq -\alpha$. Thus all probabilities in S are in $[0, 1]$, as claimed.

A similar expression for the conditional covairnce in S to that for α above yields

$$\mathrm{Cov}_S[X_i, X_j \mid X_S = a] = (p_{1,1} - \alpha)(p_{0,0} - \alpha) - (p_{1,0} + \alpha)(p_{0,1} + \alpha)$$
$$= 0\alpha^2 - (p_{0,0} + p_{1,1} + p_{0,1} + p_{1,0})\alpha + p_{1,1}p_{0,0} - p_{1,0}p_{0,1}$$
$$= p_{1,1}p_{0,0} - p_{1,0}p_{0,1} - \alpha = 0 .$$

Since X_i and X_j are Bernoulli random variables, the conditional covariance being zero implies that they are conditionally independent. ■

Completeness. Suppose by contrapositive that the algorithm rejects. Claim 74 implies that in $\tilde{P}$, X_i and X_j are $\gamma/9$-far from independent conditioned on X_S. Thus they are $\gamma/9$ far in $\tilde{P}_D$ and, since $d_{TV}(P_D, \tilde{P}_D) \leq \gamma/10$, this implies that they are not conditionally independent in P_D. Thus, in P, X_i and X_j are not independent conditioned on X_S.

Soundness. Now suppose that X_i and X_j are γ-far from independent conditioned on X_S in P. Per the foregoing discussion, this implies that they are $(9\gamma/10)$-far from being so in $\tilde{P}_D$. Now Claim 74 guarantees that $\mathbb{E}_{Y \sim \tilde{P}}[|\mathrm{Cov}_{\tilde{P}}[X_i, X_j | X_S = Y_S]|] \leq 9\gamma/20 > \gamma/3$, and therefore the algorithm rejects in this case. This completes the proof of correctness.

■

Structure Tester. Finally, we turn to the proof of Theorem 68, analyzing the structure testing algorithm described in Fig. 6.

Input $\gamma > 0$, a structure $\mathcal{S}$ and a Bayesian network P

- Draw $O((2^d + d \log n)/\gamma^2)$ samples from P. Call this set of samples S.

For each nodes X_i, X_j and set S of nodes with $|S| \leq d$ and $i, j \neq S$

    If one of the following conditions holds in structure $\mathcal{S}$

        (i) X_i is the parent of X_j,

        (ii) S contains a node X_k that is a child of both X_j and X_j,

        (iii) X_i is a grandparent of X_j and there is a child of X_i and parent of X_j, X_k that is not in S,

        (iv) X_i and X_j have a common parent X_k that is not in S

    Then run the conditional independence tester of Theorem 73 (Fig. 5) using the set of samples S to test whether X_i and X_j are independent conditioned on X_S.

    If the conditional indpendence tester accepts, return reject.

Otherwise return accept.

Figure 6: Testing whether P has structure as $\mathcal{S}$

Proof [Proof of Theorem 68] We first show correctness. There are at most n^{d+2} possible choices of X_i, X_j and $|S|$ and thus we run the conditional independence tester at most n^{d+2} times. With $O((2^d + d \log n)/\gamma^2)$ samples, each test gives an incorrect answer with probability no more than

$\tau = n^{-\Omega(d)}$. With appropriate choice of constants we therefore have that all conditional independence tests are correct with probability $99/100$. We henceforth condition on this, i.e., that all such tests are correct.

Completeness. If P is γ-non-degenerate with respect to structure $\mathcal{S}$ and degree d, then by the definition of non-degeneracy, for any X_i, X_j and S that satisfy one of conditions (i)–(iv) we have that X_i and X_j are γ-far from independent conditioned on X_S. Thus every conditional independence test rejects and the algorithm accepts.

Soundness. Now suppose by contrapositive that the algorithm accepts. For any X_i, X_j, and S that satisfy one of conditions (i)–(iv), the conditional independence test must have rejected, that is any such X_i and X_j are not independent conditioned on such an X_S. Let γ' be the mimuimum over all X_i, X_j, and S that satisfy one of conditions (i)–(iv) and distributions Q over $\{0, 1\}$ such that X_i and X_j are independent conditioned on X_S in Q, of the total variation distance between P and Q. Since there are only finitely many such combinations of X_i, X_j, and S, this γ' is positive. Thus P is γ'-non-degenerate with respect to $\mathcal{S}$ and d. Since we assumed that P has a structure $\mathcal{S}'$ with degree at most d and whose skeleton has no more edges than that of $\mathcal{S}$, we can apply Lemma 71, which yields that $\mathcal{S}$ and $\mathcal{S}'$ are I-equivalent. Thus P can indeed be expressed as a Bayesian network with structure $\mathcal{S}$. This completes the proof of correctness.

To conclude, observe that we run the loop at most n^{d+2} times, each using time at most $O((2^d + d \log n)/\gamma^2)$. The total running time is thus $O(n^{d+3}/\gamma^2)$. ∎

9. Testing Closeness of Bayes Nets

9.1. Fixed Structure Bayes Nets

We now establish the upper bound part of Theorem 18 for closeness, namely testing closeness between two unknown Bayes nets with the same (known) underlying structure.

Theorem 75 *There exists a computationally efficient algorithm with the following guarantees. Given as input (i) a DAG $\mathcal{S}$ with n nodes and maximum in-degree d, (ii) a parameter $\epsilon > 0$, and (iii) sample access to two unknown (c, C)-balanced Bayes nets P, Q with structure $\mathcal{S}$, where $c = \tilde{\Omega}(1/\sqrt{n})$ and $C = \tilde{\Omega}(d\epsilon^2/\sqrt{n})$; the algorithm takes $O(2^{d/2}\sqrt{n}/\epsilon^2)$ samples from P and Q, and distinguishes with probability at least $2/3$ between the cases $P = Q$ and $\|P - Q\|_1 > \epsilon$.*

Proof We choose $m \geq \alpha \frac{2^{d/2}\sqrt{n}}{\epsilon^2}$, where $\alpha > 0$ is an absolute constant to be determined in the course of the analysis. Let $\mathcal{S}$ and P, Q be as in the statement of the theorem, for $c \geq \beta \frac{\log n}{\sqrt{n}} \geq \beta \frac{\log n}{m}$ and $C \geq \beta \frac{d + \log n}{m}$, for some other absolute constant $\beta > 0$.

The algorithm proceeds as follows: first, taking m samples from both P and Q, it computes for each parental configuration $(i, a) \in [n] \times \{0, 1\}^d$ the number of times $\hat{N}_{i,a}$ and $\hat{M}_{i,a}$ this configuration was observed among the samples, for respectively P and Q. If for any (i, a) it is the case that $\hat{N}_{i,a}$ and $\hat{M}_{i,a}$ are not within a factor 4 of each other, the algorithm returns reject. (Using the same number of samples, it also estimates $p_{i,a}$ and $q_{i,a}$ within an additive $1/3$, and applies the same standard transformation as before so that we can hereafter assume $p_{i,a}, q_{i,a} \leq 2/3$ for all (i, a).)

Note that $\mathbb{E}\left[\hat{N}_{i,a}\right] = m \Pr_P\left[\Pi_{i,a}\right]$ and $\mathbb{E}\left[\hat{M}_{i,a}\right] = m \Pr_Q\left[\Pi_{i,a}\right]$; given the C-balancedness assumption and by Chernoff and union bounds, with probability at least $9/10$ we have that $\hat{N}_{i,a}$ and $\hat{M}_{i,a}$ are within a factor 2 of their expectation simultaneously for all $n2^d$ parental configurations. We hereafter condition on this (and observe that this implies that if $P = Q$, then the algorithm rejects in the step above with probability at most $1/10$).

The algorithm now draws independently $n2^d$ values $(M_{i,a})_{(i,a)}$, where $M_{i,a} \sim \mathrm{Poi}\left(\hat{N}_{i,a}\right)$; and takes fresh samples from P, Q until it obtains $M_{i,a}$ samples for each parental configuration $\Pi_{i,a}$ (for each of the two distributions). If at any point the algorithm takes more than $10m$ samples, it stops and returns reject.

(Again, note that by concentration (this time of Poisson random variables)[3], our assumption that $\hat{N}_{i,a} \geq m \Pr_P\left[\Pi_{i,a}\right]/2 \geq mC/2 = \beta(d + \log n)$ and a union bound, the algorithm will reject at this stage with probability at most $1/10$.)

Conditioning on not having rejected, we define for each parental configuration $\Pi_{i,a}$ the quantity $U_{i,a}$ (resp. $V_{i,a}$) as the number of samples from P (resp. Q) among the first $M_{i,a}$ satisfying $\Pi_{i,a}$ for which $X_i = 1$. In particular, this implies that $U_{i,a} \sim \mathrm{Poi}\left(p_{i,a}\hat{N}_{i,a}\right)$, $V_{i,a} \sim \mathrm{Poi}\left(q_{i,a}\hat{N}_{i,a}\right)$ (and are independent), and that the random variables $W_{i,a}$ defined below:

$$W_{i,a} \overset{\mathrm{def}}{=} \frac{(U_{i,a} - V_{i,a})^2 - (U_{i,a} + V_{i,a})}{U_{i,a} + V_{i,a}}$$

are independent. We then consider the statistic W:

$$W \overset{\mathrm{def}}{=} \sum_{i=1}^{n} \sum_{a \in \{0,1\}^d} W_{i,a}.$$

Claim 76 *If $P = Q$, then $\mathbb{E}[W] = 0$. Moreover, if $\|P - Q\|_1 > \epsilon$ then $\mathbb{E}[W] > \frac{m\epsilon^2}{144}$.*

Proof We start by analyzing the expectation of $W_{i,a}$, for any fixed $(i, a) \in [n] \times \{0,1\}^d$. The same argument as Claim 76 leads to conclude that $\mathbb{E}[W_{i,a}] = 0$ if $P = Q$ (proving the first part of the claim), and that otherwise we have

$$\mathbb{E}[W_{i,a}] \geq \frac{\min(1, mc)}{3} \hat{N}_{i,a} \frac{(p_{i,a} - q_{i,a})^2}{p_{i,a} + q_{i,a}} = \frac{1}{3}\hat{N}_{i,a} \frac{(p_{i,a} - q_{i,a})^2}{p_{i,a} + q_{i,a}}$$

$$\geq \frac{2}{9}\hat{N}_{i,a} \frac{(p_{i,a} - q_{i,a})^2}{(p_{i,a} + q_{i,a})(2 - p_{i,a} - q_{i,a})} \tag{10}$$

(since $mc \geq \beta \log n \gg 1$ and $0 < p_{i,a}, q_{i,a} \leq 2/3$). Summing over all (i, a)'s and recalling that $\hat{N}_{i,a} \geq m \Pr_P\left[\Pi_{i,a}\right]/2$, $\hat{N}_{i,a} \geq m \Pr_Q\left[\Pi_{i,a}\right]/2$ yields the bound:

$$\mathbb{E}[W] \geq \frac{m}{9} \sum_{(i,a)} \sqrt{\Pr_P\left[\Pi_{i,a}\right]\Pr_Q\left[\Pi_{i,a}\right]} \frac{(p_{i,a} - q_{i,a})^2}{(p_{i,a} + q_{i,a})(2 - p_{i,a} - q_{i,a})} \geq \frac{m}{18} \mathrm{d_H}(P, Q)^2$$

$$\geq \frac{m}{18}\left(1 - \sqrt{1 - \frac{1}{4}\|P - Q\|_1^2}\right)$$

---

3. Specifically, if $X \sim \mathrm{Poi}(\lambda)$ then we have $\Pr[|X - \lambda| > \lambda/2] = e^{-\Omega(\lambda)}$.

(where we relied on Lemma 30 for the second-to-last inequality). This gives the last part of the claim, as the RHS is at least $\frac{m\epsilon^2}{144}$ whenever $\|P - Q\|_1^2 > \epsilon^2$. ∎

We now bound the variance of our estimator:

Claim 77 $\mathrm{Var}[W] \leq n2^{d+1} + 5\sum_{(i,a)} \hat{N}_{i,a}\frac{(p_{i,a}-q_{i,a})^2}{p_{i,a}+q_{i,a}} = O(n2^d + \mathbb{E}[W])$. *In particular, if $P = Q$ then* $\mathrm{Var}[W] \leq n2^{d+1}$.

Proof We follow the proof of Claim 38 to analyze the variance of $W_{i,a}$, obtaining a bound of $\mathrm{Var}[W_{i,a}] \leq 2 + 5\hat{N}_{i,a}\frac{(p_{i,a}-q_{i,a})^2}{p_{i,a}+q_{i,a}}$. Invoking Eq. (10) and summing over all $(i,a) \in [n] \times \{0,1\}^d$ then lead to the desired conclusion. ∎

The correctness of our algorithm will then follow for the two claims above:

Lemma 78 *Set* $\tau \overset{\text{def}}{=} \frac{\epsilon^2}{288}$. *Then we have the following:*

- *If* $\|P - Q\|_1 = 0$, *then* $\Pr[W \geq \tau am] \leq \frac{1}{10}$.

- *If* $\|P - Q\|_1 > \epsilon$, *then* $\Pr[W < \tau m] \leq \frac{1}{10}$.

Proof We start with the soundness case, i.e. assuming $\|P - Q\|_1 > \epsilon$, which by Claim 76 implies $\mathbb{E}[W] > 2\tau$. Then, by Chebyshev's inequality,

$$
\begin{aligned}
\Pr[W < \tau m] \leq \Pr\left[\mathbb{E}[W] - W > \frac{1}{2}\mathbb{E}[W]\right] &\leq \frac{4\,\mathrm{Var}[W]}{\mathbb{E}[W]^2} \\
&\leq \frac{8n2^d}{\mathbb{E}[W]^2} + \frac{12}{5\mathbb{E}[W]} \qquad \text{(Claim 77)} \\
&= O\left(\frac{n2^d}{\epsilon^4 m^2} + \frac{1}{m\epsilon^2}\right) .
\end{aligned}
$$

We want to bound this quantity by $1/10$, for which it is enough to have $\frac{n2^d}{\epsilon^4 m^2} \ll 1$ and $\frac{1}{m\epsilon^2} \ll 1$, which both hold for an appropriate choice of the absolute constant $\alpha > 0$ in our setting of m.

Turning to the completeness, we suppose $\|P - Q\|_1 = 0$. Then, by Chebyshev's inequality, and invoking Claim 77,

$$
\Pr[W \geq \tau m] = \Pr[W \geq \mathbb{E}[W] + \tau m] \leq \frac{\mathrm{Var}[W]}{\tau^2 m^2} = O\left(\frac{n2^d}{\epsilon^4 m^2}\right)
$$

which is no more than $1/10$ for the same choice of m. ∎

Combining all the elements above concludes the proof, as by a union bound the algorithm is correct with probability at least $1 - \left(\frac{1}{10} + \frac{1}{10} + \frac{1}{10}\right) > \frac{2}{3}$. ∎

9.2. Unknown Structure Bayes Nets

As for the case identity testing, we give a closeness tester for balanced non-degenerate Bayes Nets. An additional assumption that we require is that the ordering of the nodes in the corresponding DAGs is known to the algorithm. Formally, we show:

Theorem 79 *There exists a computationally efficient algorithm with the following guarantees. Given as input (i) a parameter $\epsilon > 0$, (ii) an ordering of nodes π, and (ii) sample access to unknown γ-non-degenerate, (c, C)-balanced Bayes nets P, Q such the structures of P and Q give the same ordering π to nodes, where $c = \tilde{\Omega}\left(1/\sqrt{n}\right)$ and $C = \tilde{\Omega}\left(d\epsilon^2/\sqrt{n}\right)$; the algorithm takes $N = O(2^{d/2}\sqrt{n}/\epsilon^2 + 2^d/\gamma^2 + d\log(n)/\gamma^2)$ samples from P and Q, runs in time $n^d \text{poly}(N)$, and distinguishes with probability at least $2/3$ between the cases $P = Q$ and $\|P - Q\|_1 > \epsilon$.*

Proof The argument's idea is the following: we first test that P and Q have the same skeleton. Since they have the same ordering, that suffices to show that they have the same structure. If this is the case, then we use our known-structure tester.

In more detail, given the γ-non-degeneracy assumption, for each pair of coordinates i, j and set of coordinates S with $|S| \leq d$, we can, using the conditional independence tester from Theorem 73 to test whether each of P and Q has X_i and X_j conditionally independent on X_S or γ-far from it with $n^{-d-2}/100$ probability of error in $O((2^d + d\log(n))/\gamma^2)$ samples. Running tests on the same samples for all n^{d+2} combinations of i, j, S, we can with probability at least $99/100$ correctly classify which of the two cases holds, for all i, j, S that are either conditionally independent or γ-far. We note that by non-degeneracy, there is an edge between i and j in the structure defining P only if X_i and X_j are γ-far from independent conditioned on X_S for all S (i.e., if there is no edge then there must exist a S such that X_i and X_j are conditionally independent on X_S). Therefore, assuming our conditional independence testers all answered as they should, we can use this to successfully identify the set of edges in the structure of P (and thus, since we know the ordering, the entire structure).

Having determined the underlying structures of P and Q, our tester rejects if these structures differ (as using Lemma 71, γ-non-degeneracy implies that neither can equal a Bayes net with non-equivalent structure and fewer edges). Otherwise, we run the tester from Theorem 75 (since we satisfy its assumptions) and return the result. ∎

10. Identity and Closeness Testing for High-Degree Bayes Nets

Finally, in this section we give testing algorithms for identity and closeness of degree-d Bayes nets with unknown structure, *without* balancedness assumptions. Compared to the testing algorithm of Theorem 21 and Theorem 79 (which work under such assumptions) the dependence on the number of nodes n the testers in this section are suboptimal, they achieve the "right" dependence on the degree d (specifically, $2^{d/2}$ for identity and $2^{2d/3}$ for closeness). Hence, these testers achieve sub-learning sample complexity for the case that $d = \Omega(\log n)$.

Theorem 80 *There exists two algorithms with the following guarantees:*

- *(Identity) Given the full description of a Bayes net Q of degree at most d, parameter $\epsilon \in (0, 1]$, and sample access to a distribution P promised to be a Bayes net (i) of degree at most*

d and (ii) such that the structures of P and Q give the same ordering to nodes, the first takes $N = 2^{d/2}\text{poly}(n/\epsilon)$ samples from P, runs in time $n^d\text{poly}(N)$, and distinguishes with probability at least $2/3$ between (i) $P = Q$ and (ii) $\|P - Q\|_1 > \epsilon$.

- *(Closeness) Given parameter $\epsilon \in (0, 1]$, and sample access to two distributions P, Q promised to be Bayes nets (i) of degree at most d and (ii) such that the structures of P and Q give the same ordering to nodes, the second takes $N = 2^{2d/3}\text{poly}(n/\epsilon)$ samples from P and Q, runs in time $n^d\text{poly}(N)$, and distinguishes with probability at least $2/3$ between (i) $P = Q$ and (ii) $\|P - Q\|_1 > \epsilon$.*

Proof We first establish the first part of the theorem, namely the existence of an identity testing algorithm with optimal dependence on the degree d. The algorithm is quite simple: it goes over each set $S \subseteq [n]$ of at most $d + 1$ coordinates, and checks that for each of them it holds that the conditional distributions P, S, Q_S are equal (versus $\|P_S - Q_S\|_1 > \text{poly}(\frac{\epsilon}{n})$).

Since P_S and Q_S are supported on sets of size $O(2^d)$, and as there are only $O(n^{d+1})$ such sets to consider, the claimed sample complexity suffices to run all tests correctly with probability $9/10$ overall (by a union bound).

The more difficult part is to argue correctness, that is to show that if the test accepts then one must have $\|P - Q\|_1 < \epsilon$. To do so, assume (without loss of generality) that $H(P) \leq H(Q)$: we will show that $\text{D}(P\|Q)$ is small, which implies that the L_1 distance is as well.

Let the ordering of P be coordinates $1, 2, 3, \ldots$. We note that $\text{D}(P\|Q) = \sum_i \text{D}(P_i\|Q_i \mid P_1, \ldots, P_{i-1})$ (i.e. the expectation over $P_1, \ldots, P_{i-1}$ of the KL-divergence of the conditional distributions of P_i and Q_i, conditioned on these $(i - 1)$ coordinates). It thus suffices to show that each of these terms is small.

Let S_i be the set of parents of node i under P. We have that:

$$\text{D}(P_i\|Q_i \mid P_1, \ldots, P_{i-1}) = \text{D}(P_i\|Q_i \mid P_{S_i}) + \mathbb{E}_{P_1, \ldots, P_{i-1}}[\text{D}(Q_i \mid P_{S_i}\|Q_i \mid P_1, \ldots, P_{i-1})] .$$

Further, note that the fact that the tester accepted implies that $\text{D}(P_i\|Q_i \mid P_{S_i})$ is small. Now, we have that

$$H(P) = \sum_i H(P_i \mid P_1, \ldots, P_{i-1}) = \sum_i H(P_i \mid P_{S_i}) ,$$

$$H(Q) = \sum_i H(Q_i \mid Q_1, \ldots, Q_{i-1}) = \sum_i H(Q_i \mid Q_{S_i}) - I(Q_i; Q_1, \ldots, Q_{i-1} \mid Q_{S_i}) .$$

But since the $(d + 1)$-wise probabilities are close, we have that $H(P_i \mid P_{S_i})$ is close to $H(Q_i \mid Q_{S_i})$ (up to an additive $\text{poly}(\epsilon/n)$). Therefore, for each i, we have that $I(Q_i; Q_1, \ldots, Q_{i-1} \mid Q_{S_i}) = \text{poly}(\epsilon/n)$. In order to conclude, let us compare $I(Q_i; Q_1, \ldots, Q_{i-1} \mid Q_{S_i})$ and $\mathbb{E}_{P_1, \ldots, P_{i-1}}[\text{D}(Q_i \mid P_{S_i}\|Q_i \mid P_1, \ldots, P_{i-1})]$. The former is the sum, over assignments $y \in \{0, 1\}^{i-1}$ consistent with an assignment $x \in \{0, 1\}^{S_i}$, of

$$\Pr[Q_{S_i} = x] H(Q_i \mid Q_{S_i} = x) + \Pr[Q_{1, \ldots, i-1} = y] H(Q_i \mid Q_{1, \ldots, i-1} = y).$$

The latter is the sum over the same y's of

$$\Pr[P_{S_i} = x] H(Q_i \mid Q_{S_i} = x) + \Pr[P_{1, \ldots, i-1} = y] H(Q_i \mid Q_{1, \ldots, i-1} = y) .$$

But because of the d-way probability similarities, the terms $\Pr[\,P_{S_i} = x\,]$ and $\Pr[\,Q_{S_i} = x\,]$ terms are very close, within an additive $\mathrm{poly}(\epsilon/n)$.

(Here we use the extra assumption that P and Q use the same ordering.) Denote by T_i the parents of i under the topology of Q. Then $H(Q_i \mid Q_{1,\ldots,i-1} = y)$ depends only on the values of the coordinates in T_i. Thus the last part of the sum is a sum over z of $\Pr[\,Q_{T_i} = z\,] H(Q_i \mid Q_{T_i} = z)$ and $\Pr[\,P_{T_i} = z\,] H(Q_i \mid Q_{T_i} = z)$, which are also close by a similar argument. Thus,

$$\mathbb{E}_{P_1,\ldots,P_{i-1}}[\mathrm{D}(Q_i \mid P_{S_i} \| Q_i \mid P_1,\ldots,P_{i-1})] = I\left(Q_i; Q_1,\ldots,Q_{i-1} \mid Q_{S_i}\right) + \mathrm{poly}\left(\frac{\epsilon}{n}\right) = \mathrm{poly}\left(\frac{\epsilon}{n}\right).$$

This implies that P, Q are close in KL divergence, and therefore in L_1.

The second part of the theorem, asserting the existence of a closeness testing algorithm with optimal dependence on d, will be very similar. Indeed, by the proof above it suffices to check that the restrictions of P and Q to any set of $(d+3)$-coordinates are $\mathrm{poly}(\epsilon/n)$-close. Using known results Chan et al. (2014c), this can be done for any specific collection of $d + 3$ coordinates with N samples in $\mathrm{poly}(N)$ time, and high probability of success, implying the second part of the theorem. ∎

References

P. Abbeel, D. Koller, and A. Y. Ng. Learning factor graphs in polynomial time and sample complexity. *J. Mach. Learn. Res.*, 7:1743–1788, 2006.

J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Competitive closeness testing. *Journal of Machine Learning Research - Proceedings Track*, 19:47–68, 2011.

J. Acharya, C. Daskalakis, and G. Kamath. Optimal testing for properties of distributions. *CoRR*, abs/1507.05952, 2015a.

J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt. Sample-optimal density estimation in nearly-linear time. *CoRR*, abs/1506.00671, 2015b.

M. Adamaszek, A. Czumaj, and C. Sohler. Testing monotone continuous distributions on high-dimensional real cubes. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010*, pages 56–65, 2010.

A. Almudevar. A hypothesis test for equality of bayesian network models. *EURASIP J. Bioinformatics and Systems Biology*, 2010, 2010.

A. Anandkumar, D. J. Hsu, F. Huang, and S. Kakade. Learning mixtures of tree graphical models. In *NIPS*, pages 1061–1069, 2012.

S. Arora and R. Kannan. Learning mixtures of arbitrary Gaussians. In *Proceedings of the 33rd Symposium on Theory of Computing*, pages 247–257, 2001.

Z. Bai and H. Saranadasa. Effect of high dimension: by an example of a two sample problem. *Statist. Sinica,*, 6:311–329, 1996.

T. Batu. *Testing Properties of Distributions*. PhD thesis, Cornell University, 2001.

T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000. URL citeseer.ist.psu.edu/batu00testing.html.

T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *Proc. 42nd IEEE Symposium on Foundations of Computer Science*, pages 442–451, 2001.

T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating entropy. In *ACM Symposium on Theory of Computing*, pages 678–687, 2002.

T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *ACM Symposium on Theory of Computing*, pages 381–390, 2004.

M. Belkin and K. Sinha. Polynomial learning of distribution families. In *FOCS*, pages 103–112, 2010.

A. Bhattacharyya, E. Fischer, R. Rubinfeld, and P. Valiant. Testing monotonicity of distributions over general partial orders. In *ICS*, pages 239–252, 2011.

P. J. Bickel. A distribution free version of the smirnov two sample test in the p-variate case. *Ann. Math. Statist.*, 40(1):1–23, 02 1969.

T. T. Cai and Z. Ma. Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli*, 19(5B):2359–2388, 2013.

C. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. Testing shape restrictions of discrete distributions. In *33rd Symposium on Theoretical Aspects of Computer Science, STACS 2016*, pages 25:1–25:14, 2016.

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

S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Learning mixtures of structured distributions over discrete domains. In *SODA*, pages 1380–1394, 2013.

S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In *STOC*, pages 604–613, 2014a.

S. Chan, I. Diakonikolas, R. Servedio, and X. Sun. Near-optimal density estimation in near-linear time using variable-width histograms. In *NIPS*, pages 1844–1852, 2014b.

S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014c.

S. X. Chen and Y. L. Qin. A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.*, 38(2):808–835, 04 2010.

J. Cheng, R. Greiner, J. Kelly, D. Bell, and W. Liu. Learning bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, 137(1):43 – 90, 2002.

D. M. Chickering. Learning equivalence classes of bayesian-network structures. *J. Mach. Learn. Res.*, 2:445–498, 2002.

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theor.*, 14(3):462–467, 1968.

M. Cryan, L. Goldberg, and P. Goldberg. Evolutionary trees can be learned in polynomial time in the two state general Markov model. *SIAM Journal on Computing*, 31(2):375–397, 2002.

R. Daly, Q. Shen, and S. Aitken. Learning bayesian networks: approaches and issues. *The Knowledge Engineering Review*, 26:99–157, 2011. ISSN 1469-8005.

S. Dasgupta. The sample complexity of learning fixed-structure bayesian networks. *Machine Learning*, 29(2-3):165–180, 1997.

S. Dasgupta. Learning mixtures of Gaussians. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, pages 634–644, 1999.

C. Daskalakis and Q. Pan. Square hellinger subadditivity for bayesian networks and its applications to identity testing. *CoRR*, abs/1612.03164, 2016.

C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-modal distributions via testing. In *SODA*, pages 1371–1385, 2012a.

C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning Poisson Binomial Distributions. In *STOC*, pages 709–728, 2012b.

C. Daskalakis, I. Diakonikolas, R. O'Donnell, R.A. Servedio, and L. Tan. Learning Sums of Independent Integer Random Variables. In *FOCS*, pages 217–226, 2013a.

C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *SODA*, pages 1833–1852, 2013b.

C. Daskalakis, A. De, G. Kamath, and C. Tzamos. A size-free CLT for poisson multinomials and its applications. In *Proceedings of STOC'16*, 2016a.

Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing ising models. *CoRR*, abs/1612.03147, 2016b. URL http://arxiv.org/abs/1612.03147.

A. De, I. Diakonikolas, and R. Servedio. Learning from satisfying assignments. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, pages 478–497, 2015.

I. Diakonikolas and D. M. Kane. A new approach for testing properties of discrete distributions. In *FOCS*, pages 685–694, 2016. Full version available at abs/1601.05557.

I. Diakonikolas, D. M. Kane, and V. Nikishkin. Testing Identity of Structured Distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, 2015a.

I. Diakonikolas, D. M. Kane, and V. Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *56th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2015*, 2015b.

I. Diakonikolas, D. M. Kane, and A. Stewart. Nearly optimal learning and sparse covers for sums of independent integer random variables. *CoRR*, abs/1505.00662, 2015c.

I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. Collision-based testers are optimal for uniformity and closeness. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:178, 2016a.

I. Diakonikolas, D. M. Kane, and A. Stewart. The fourier transform of poisson multinomial distributions and its algorithmic applications. In *Proceedings of STOC'16*, 2016b. Available at https://arxiv.org/abs/1511.03592.

I. Diakonikolas, D. M. Kane, and A. Stewart. Robust learning of fixed-structure bayesian networks. *CoRR*, abs/1606.07384, 2016c.

Y. Freund and Y. Mansour. Estimating a mixture of two product distributions. In *Proceedings of the 12th Annual COLT*, pages 183–192, 1999.

N. Friedman and Z. Yakhini. On the sample complexity of learning bayesian networks. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, pages 274–282, 1996.

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29 (2):131–163, 1997.

N. Friedman, M. Linial, and I. Nachman. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620, 2000.

O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, Electronic Colloquium in Computational Complexity, 2000.

M. Gonen, P. H. Westfall, and W. O. Johnson. Bayesian multiple testing for two-sample multivariate endpoints. *Biometrics*, 59(1):76–82, 2003. ISSN 1541-0420.

M. Hardt and E. Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015*, pages 753–760, 2015.

H. Hotelling. The generalization of student's ratio. *Ann. Math. Statist.*, 2(3):360–378, 08 1931.

P. Indyk, R. Levi, and R. Rubinfeld. Approximating and Testing k-Histogram Distributions in Sub-linear Time. In *PODS*, pages 15–22, 2012.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15(1):2869–2909, 2014.

F. V. Jensen and T. D. Nielsen. *Bayesian Networks and Decision Graphs*. Springer Publishing Company, Incorporated, 2nd edition, 2007.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, 2005.

R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. In *ICS*, pages 179–194, 2011.

R. Y. Liu and K. Singh. A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*, 88(421):252–260, 1993. ISSN 01621459. URL http://www.jstor.org/stable/2290720.

P. L. Loh and M. J. Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. In *NIPS*, pages 2096–2104, 2012.

D. Margaritis. *Learning Bayesian Network Model Structure From Data*. PhD thesis, CMU, 2003.

A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010.

E. Mossel and S. Roch. Learning nonsingular phylogenies and Hidden Markov Models. In *To appear in Proceedings of the 37th Annual Symposium on Theory of Computing (STOC)*, 2005.

R. E. Neapolitan. *Learning Bayesian Networks*. Prentice-Hall, Inc., 2003.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933. doi: 10.1098/rsta.1933.0009. URL http://rsta.royalsocietypublishing.org/content/231/694-706/289.short.

H.-T. Nguyen, P. Leray, and G. Ramstein. *Multiple Hypothesis Testing and Quasi Essential Graph for Comparing Two Sets of Bayesian Networks*, pages 176–185. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

L. Paninski. A coincidence-based test for uniformity given very sparsely-sampled discrete data. *IEEE Transactions on Information Theory*, 54:4750–4755, 2008.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

D. Pollard. Asymptopia. http://www.stat.yale.edu/~pollard/Books/Asymptopia, 2003. Manuscript.

Y. Polyanskiy and Y. Wu. Strong data-processing inequalities for channels and Bayesian networks. *ArXiv e-prints*, August 2015.

Y. Rahmatallah, F. Emmert-Streib, and G. Glazko. Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets. *Bioinformatics*, 30(3):360–368, 2014.

A. Ramdas, D. Isenberg, A. Singh, and L. A. Wasserman. Minimax lower bounds for linear independence testing. In *IEEE International Symposium on Information Theory, ISIT 2016*, pages 965–969, 2016.

R. Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

R. Rubinfeld and R. Servedio. Testing monotone high-dimensional distributions. In *Proc. 37th Annual ACM Symposium on Theory of Computing (STOC)*, pages 147–156, 2005.

N. P. Santhanam and M. J. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Trans. Information Theory*, 58(7):4117–4134, 2012.

D. M. Sobel and T. Kushnir. Interventions do not solely benefit causal learning: Being told what to do results in worse learning than doing it yourself. In *Proceedings of the twenty-fifth annual meeting of the Cognitive Science Society*, pages 1100–1105, 2003.

M. S. Srivastava and M. Du. A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis*, 99(3):386 – 402, 2008.

N. Städler and S. Mukherjee. Multivariate gene-set testing based on graphical models. *Biostatistics*, 16(1):47–59, 2015.

G. Valiant and P. Valiant. Estimating the unseen: an $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *STOC*, pages 685–694, 2011.

G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *FOCS*, 2014.

S. Vempala and G. Wang. A spectral algorithm for learning mixtures of distributions. In *Proceedings of the 43rd Annual Symposium on Foundations of Computer Science*, pages 113–122, 2002.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '90, pages 255–270, 1991.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, 2008.

M. J. Wainwright, P. Ravikumar, and J. D. Lafferty. High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *NIPS*, pages 1465–1472, 2006.

L. Weiss. Two-sample tests for multivariate distributions. *Ann. Math. Statist.*, 31(1):159–164, 03 1960.

W. Yin, S. Garimalla, A. Moreno, M. R. Galinski, and M. P. Styczynski. A tree-like bayesian structure learning algorithm for small-sample datasets from complex biological model systems. *BMC Systems Biology*, 9(1):1–18, 2015.

APPENDIX

Appendix A. Sample Complexity of Learning Bayesian Networks

In this section, we derive tight bounds on the sample complexity of learning Bayes nets. Recall that n will denote the number of nodes and d the maximum in-degree; before stating the results and providing their proofs, we outline the high-level idea of the argument.

If the structure is known, there is a very simple learning algorithm involving finding the empirical values for the relevant conditional probabilities and constructing the Bayes net using those terms. By computing the expected KL-Divergence between this hypothesis and the truth, we can show that it is possible to learn an ϵ-aproximation in $\tilde{O}\left(2^d n/\epsilon^2\right)$ samples. Learning a Bayes net with unknown structure seems substantially more challenging, but at least the sample complexity is no greater. In particular, we can simply go over all possible topologies and come up with one hypothesis for each, and use a standard tournament to pick out the correct one. Appendix A.1 contains the details of both.

We also prove in Appendix A.2 a matching lower bound (up to logarithmic factors). For this we can even consider Bayesian networks of fixed topology. In particular, we consider the topology where each of the last $(n-d)$-coordinates depend on all of the first d (which we can assume to be uniform and independent). The distribution we end up with is what we call a disjoint mixture of 2^d product distributions. In particular for each of the 2^d possible combinations of the first d coordinates, we have a (potentially different) product distribution over the remaining coordinates. In order to learn our final distribution we must learn at least half of these product distributions to $O(\epsilon)$ error. This requires that we obtain at least $\Omega((n-d)/\epsilon^2)$ samples from $\Omega(2^d)$ of the parts of our mixture. Thus, learning will require $\Omega(2^d(n-d)/\epsilon^2))$ total samples.

A.1. Sample Complexity Upper Bound

Known Structure. The algorithm will return a Bayes net Q with the same (known) structure as the unknown P. Define $p_{i,a}$ as the probability (under P) that the i-th coordinate is 1, conditioned on the parental configuration for i being equal to a (denoted $\Pi_{i,a}$); and $q_{i,a}$ the corresponding parameter for our hypothesis Q.

Given this, the algorithm is simple. First, by a standard argument we can assume that all $p_{i,a}$ are in $[0, 2/3]$ (as one can detect whether it is the case by taking $O(\log n)$ samples, and swap the corresponding coordinate of each sample if it is not the case).

We set $m = O(\frac{2^d n}{\epsilon^2} \log(2^d n))$, and consider separately two sets of configurations (i, a):

- the *light* ones, for which $p_{i,a} \Pr_P\left[\Pi_{i,a}\right] \leq \frac{\epsilon}{2n2^d}$;

- the *heavy* ones, for which $p_{i,a} \Pr_P\left[\Pi_{i,a}\right] > \frac{\epsilon}{2n2^d}$.

We take m samples from P, and let $q_{i,a}$ be the empirical conditional probabilities. For any (i, a) for which we see less than $\tau \stackrel{\text{def}}{=} O((\log n)/\epsilon)$ samples, we set $q_{i,a} = 0$. Note that with high probability, for the right choice of constant in the definition of τ, for all heavy configurations simultaneously the estimate $q_{i,a}$ will (i) not be set to zero; and (ii) be within a factor 2 of the real value $p_{i,a}$. Conversely, each light configuration will either have $q_{i,a}$ be zero, or within a factor 2 of $p_{i,a}$.

Conditioned on this happening, we can analyze the error. Note first that by our definition of light configuration and the triangle inequality, the sum of L_1 error over all light configurations will be at

most $\epsilon/2$. We now bound the contribution to the error of the heavy configurations. (In what follows, we implicitly restrict ourselves to these.)

By Pinsker's inequality, $\|P - Q\|_1^2 \leq 2\,\mathrm{D}(P\|Q) \leq \sum_{(i,a)} \mathrm{Pr}_P\left[\,\Pi_{i,a}\,\right] \frac{(p_{i,a}-q_{i,a})^2}{q_{i,a}(1-q_{i,a})}$. It only remains to bound the expected KL divergence *via* the above bound. For each pair (i,a), on expectation we see $m\,\mathrm{Pr}_P\left[\,\Pi_{i,a}\,\right]$ samples satisfying the parental configuration. The expected squared error $(p_{i,a} - q_{i,a})^2$ between $p_{i,a}$ and our estimate $q_{i,a}$ is then $O(\frac{p_{i,a}}{m\,\mathrm{Pr}_P[\Pi_{i,a}]})$ by a standard variance argument. This implies that the expected KL-divergence is bounded as

$$\mathbb{E}[\mathrm{D}(P\|Q)] = \frac{1}{m}\sum_{(i,a)} O\left(\mathbb{E}\left[\frac{p_{i,a}(1-p_{i,a})}{q_{i,a}(1-q_{i,a})}\right]\right)$$

and it will be sufficient to argue that $\frac{p_{i,a}(1-p_{i,a})}{q_{i,a}(1-q_{i,a})} = O(1)$ with high probability (as then the RHS will be bounded as $O\left(\frac{2^d n}{m}\right)$, and thus taking $m = O(2^d n/\epsilon^2)$ will be sufficient). Since $p_{i,a} \in [0, 2/3]$, it is enough to show that, with high probability, $q_{i,a} \geq p_{i,a}/2$ simultaneously for all (i,a); but this is exactly the guarantee we had due to the "heaviness" of the configuration. Therefore, the contribution to the squared L_1 error from the heavy configuration is $O\left(\frac{2^d n}{m}\right) = O\left(\frac{\epsilon^2}{d+\log n}\right) \ll \epsilon^2$. Gathering the two sources of error, we get an L_1-error at most ϵ with high probability, as claimed.

Unknown Structure. For this upper bound, we reduce to the previous case of known structure. Indeed, there are only $N = n^{O(dn)}$ possible max-degree-d candidate structures: one can therefore run the above algorithm (with probability of failure $9/10$) for each candidate structure on a common sample from P of size $O(2^d n/\epsilon^2)$, before running a tournament (cf. Daskalakis et al. (2012b); De et al. (2015)) to select a hypothesis with error $O(\epsilon)$ (which is guaranteed to exist as, with probability at least $9/10$, the "right" candidate structure will generate a good hypothesis with error at most ϵ). The total sample complexity will then be

$$O\left(\frac{2^d n}{\epsilon^2} + \frac{\log N}{\epsilon^2}\right) = O\left(\frac{2^d n}{\epsilon^2} + \frac{dn\log n}{\epsilon^2}\right),$$

which is $O\left(2^d n/\epsilon^2\right)$ for $d = \Omega(\log n)$.

A.2. Sample Complexity Lower Bound

Our lower bound will be derived from families of Bayes nets with the following structure: The first d nodes are all independent (and will in fact have marginal probability $1/2$ each), and will form in some sense a "pointer" to one of 2^d arbitrary product distributions. The remaining $n - d$ nodes will each depend on all of the first d. The resulting distribution is now an (evenly weighted) disjoint mixture of 2^d product distributions on the $(n-d)$-dimensional hypercube. In other words, there are 2^d product distributions $p_1, \ldots, p_{2^d}$, and our distribution returns a random i (encoded in binary) followed by a random sample form p_i. Note that the p_i can be arbitrary product distributions.

We show a lower bound of $\Omega(2^d n/\epsilon^2)$ lower bound for learning, whenever $d < (1 - \Omega(1))n$.

Let $\mathcal{C}_\epsilon$ be a family of product distributions over $\{0,1\}^{n-d}$ which is hard to learn, i.e., such that any algorithm learning $\mathcal{C}_\epsilon$ to accuracy 4ϵ which succeeds with probability greater than $1/2$ must have sample complexity $\Omega(n/\epsilon^2)$. We will choose each p_i independently and uniformly at random from $\mathcal{C}_\epsilon$.

Assume for the sake of contradiction that there exists an algorithm $\mathcal{A}$ to learn the resulting disjoint mixture of p_i's to error ϵ with $o(2^d(n-d)/\epsilon^2)$ samples. Without loss of generality, this algorithm can be thought of returning as hypothesis a disjoint union of some other distributions $q_1, \ldots, q_{2^d}$, in which case the error incurred is $\|p - q\|_1 = \frac{1}{2^d}\sum_{i=1}^{n-d}\|p_i - q_i\|_1$.

By assumption on its sample complexity, for at least half of the indices $1 \leq i \leq n - d$ the algorithm obtains $o(n/\epsilon^2)$ samples from p_i. This implies that in expectation, by the fact that $\mathcal{C}_\epsilon$ was chosen hard to learn, for at least half of these indices it will be the case that $\|p_i - q_i\|_1 > 16\epsilon$ (as each of these i's is such that $\|p_i - q_i\|_1 > 16\epsilon$ with probability at least $1/2$). This in turn shows that in expectation, $\|p - q\|_1 = \frac{1}{2^d}\sum_{i=1}^{n-d}\|p_i - q_i\|_1 > \frac{4\epsilon}{4} = \epsilon$, leading to a contradiction.

Appendix B. Omitted Proofs from Section 2.2

We give in this section the proofs of the corresponding lemmas from Section 2.2.

Proof [Proof of Lemma 10] By symmetry, it is sufficient to consider the distribution $P \overset{\text{def}}{=} \bigotimes_{j=1}^n \text{Bern}\left(\frac{1}{2} + \frac{\epsilon}{\sqrt{n}}\right)$. We explicitly bound from below the expression of $\|P - U\|_1$:

$$
\begin{aligned}
\|P - U\|_1 &= \sum_{x \in \{0,1\}^n} \left| \left(\frac{1}{2} + \frac{\epsilon}{\sqrt{n}}\right)^{|x|} \left(\frac{1}{2} - \frac{\epsilon}{\sqrt{n}}\right)^{n-|x|} - \frac{1}{2^n} \right| \\
&= \frac{1}{2^n} \sum_{k=0}^n \binom{n}{k} \left| \left(1 + \frac{2\epsilon}{\sqrt{n}}\right)^k \left(1 - \frac{2\epsilon}{\sqrt{n}}\right)^{n-k} - 1 \right| \\
&\geq \frac{1}{2^n} \sum_{k=\frac{n}{2}+\sqrt{n}}^{\frac{n}{2}+2\sqrt{n}} \binom{n}{k} \left| \left(1 + \frac{2\epsilon}{\sqrt{n}}\right)^k \left(1 - \frac{2\epsilon}{\sqrt{n}}\right)^{n-k} - 1 \right| \\
&\geq \frac{C}{\sqrt{n}} \sum_{k=\frac{n}{2}+\sqrt{n}}^{\frac{n}{2}+2\sqrt{n}} \left| \left(1 + \frac{2\epsilon}{\sqrt{n}}\right)^k \left(1 - \frac{2\epsilon}{\sqrt{n}}\right)^{n-k} - 1 \right|,
\end{aligned}
$$

where $C > 0$ is an absolute constant. We bound from below each summand separately: fixing k, and writing $\ell = k - \frac{n}{2} \in [\sqrt{n}, 2\sqrt{n}]$,

$$
\left(1 + \frac{2\epsilon}{\sqrt{n}}\right)^k \left(1 - \frac{2\epsilon}{\sqrt{n}}\right)^{n-k} = \left(1 - \frac{4\epsilon^2}{n}\right)^{n/2} \left(\frac{1 + \frac{2\epsilon}{\sqrt{n}}}{1 - \frac{2\epsilon}{\sqrt{n}}}\right)^\ell \geq \left(1 - \frac{4\epsilon^2}{n}\right)^{n/2} \left(\frac{1 + \frac{2\epsilon}{\sqrt{n}}}{1 - \frac{2\epsilon}{\sqrt{n}}}\right)^{\sqrt{n}}
$$
$$
\xrightarrow[n \to \infty]{} e^{4\epsilon - 2\epsilon^2},
$$

so that each summand is bounded by a quantity that converges (when $n \to \infty$) to $e^{4\epsilon - 2\epsilon^2} - 1 > 4\epsilon - 2\epsilon^2 > 2\epsilon$, implying that each is $\Omega(\epsilon)$. Combining the above gives

$$
\|P - U\|_1 \geq \frac{C}{\sqrt{n}} \sum_{k=\frac{n}{2}+\sqrt{n}}^{\frac{n}{2}+2\sqrt{n}} \Omega(\epsilon) = \Omega(\epsilon)
$$

as claimed. ∎

Proof [Proof of Lemma 12] By symmetry it suffices to consider only the case of $i = 1$, so that we let $A = N_1$. The first step is to bound from above $I(X; A)$ by a more manageable quantity:

Fact 81 *We have that*

$$I(X; A) \leq \sum_{a=0}^{\infty} \Pr[A = a] \left(1 - \frac{\Pr[A = a \mid X = 1]}{\Pr[A = a \mid X = 0]}\right)^2. \tag{11}$$

The proof of this fact is given in Appendix D. It will then be sufficient to bound the RHS, which we do next. Since $A \sim \mathrm{Poi}(kp_1)$ with $p_1 = 1/2$ if $X = 0$ and uniformly $\frac{1}{2} \pm \frac{\epsilon}{\sqrt{n}}$ if $X = 1$, a simple computation yields that

$$\Pr[A = \ell \mid X = 0] = e^{-k/2} \frac{(k/2)^\ell}{\ell!}$$

$$\Pr[A = \ell \mid X = 1] = \left(e^{-k/2} \frac{(k/2)^\ell}{\ell!}\right) \left(\frac{e^{-k\epsilon/\sqrt{n}}(1 + 2\frac{\epsilon}{\sqrt{n}})^\ell + e^{k\epsilon/\sqrt{n}}(1 - 2\frac{\epsilon}{\sqrt{n}})^\ell}{2}\right).$$

Writing out $\varphi(\epsilon, \ell) = \frac{\Pr[A = \ell \mid X = 1]}{\Pr[A = \ell \mid X = 0]}$ as a function of $\epsilon/\sqrt{n}$, we see that it is even. Thus, expanding it as a Taylor series in $\alpha \stackrel{\text{def}}{=} \epsilon/\sqrt{n}$, the odd degree terms will cancel. Moreover, we can write

$$\sum_{\ell=0}^{\infty} \Pr[A = \ell] (1 - \varphi(\epsilon, A))^2 = \mathbb{E}_A \left[(1 - \varphi(\epsilon, A))^2\right]$$

$$= \frac{1}{2} \mathbb{E}_{A \sim \mathrm{Poi}(k/2)} \left[(1 - \varphi(\epsilon, A))^2\right]$$

$$+ \frac{1}{4} \mathbb{E}_{A \sim \mathrm{Poi}(k(1/2+\alpha))} \left[(1 - \varphi(\epsilon, A))^2\right]$$

$$+ \frac{1}{4} \mathbb{E}_{A \sim \mathrm{Poi}(k(1/2-\alpha))} \left[(1 - \varphi(\epsilon, A))^2\right].$$

Now, we can rewrite

$$(1 - \varphi(\epsilon, A))^2 = \left(1 - \frac{e^{-k\alpha}(1 + 2\alpha)^\ell + e^{k\alpha}(1 - 2\alpha)^\ell}{2}\right)^2$$

$$= 1 - \left(e^{-k\alpha}(1 + 2\alpha)^\ell + e^{k\alpha}(1 - 2\alpha)^\ell\right)$$

$$+ \frac{e^{-2k\alpha}(1 + 2\alpha)^{2\ell} + 2(1 - 4\alpha^2)^\ell + e^{2k\alpha}(1 - 2\alpha)^{2\ell}}{4}.$$

For $b \in \{-1, 0, 1\}$, we have $\mathbb{E}_{A \sim \mathrm{Poi}(k(1/2+b\alpha))}[1] = 1$ (!), and (from the MGF of a Poisson distribution)

$$e^{-k\alpha} \mathbb{E}_{A \sim \mathrm{Poi}(k(1/2+b\alpha))} \left[(1 + 2\alpha)^A\right] = e^{-k\alpha} e^{k(1/2+b\alpha) \cdot 2\alpha} = e^{b \cdot 2\alpha^2 k}$$

$$e^{k\alpha} \mathbb{E}_{A \sim \mathrm{Poi}(k(1/2+b\alpha))} \left[(1 - 2\alpha)^A\right] = e^{k\alpha} e^{k(1/2+b\alpha) \cdot -2\alpha} = e^{-b \cdot 2\alpha^2 k},$$

as well as

$$e^{-2k\alpha}\mathbb{E}_{A\sim\text{Poi}(k(1/2+b\alpha))}\left[(1+2\alpha)^{2A}\right] = e^{-2k\alpha}e^{k(1/2+b\alpha)\cdot(4\alpha+4\alpha^2)} = e^{2k\alpha^2(1+2b+2b\alpha)}$$

$$e^{2k\alpha}\mathbb{E}_{A\sim\text{Poi}(k(1/2+b\alpha))}\left[(1-2\alpha)^{2A}\right] = e^{2k\alpha}e^{k(1/2+b\alpha)\cdot(-4\alpha+4\alpha^2)} = e^{2k\alpha^2(1-2b+2b\alpha)}$$

$$2\mathbb{E}_{A\sim\text{Poi}(k(1/2+b\alpha))}\left[(1-4\alpha^2)^{A}\right] = 2e^{k(1/2+b\alpha)\cdot-4\alpha^2} = 2e^{-2k\alpha^2-4kb\alpha^3} \ .$$

Gathering the terms, we get

$$\mathbb{E}_A\left[(1-\varphi(\epsilon,A))^2\right] = \frac{1}{4}\Bigg(2\left(1-2+\frac{e^{2k\alpha^2}+e^{-2k\alpha^2}}{2}\right)$$

$$+\left(1-(e^{2k\alpha^2}+e^{-2k\alpha^2})+\frac{e^{2k\alpha^2(3+2\alpha)}+e^{2k\alpha^2(-1+2\alpha)}+2e^{-2k\alpha^2-4k\alpha^3}}{4}\right)$$

$$+\left(1-(e^{-2k\alpha^2}+e^{2k\alpha^2})+\frac{e^{-2k\alpha^2(1+2\alpha)}+e^{2k\alpha^2(3-2\alpha)}+2e^{-2k\alpha^2+4k\alpha^3}}{4}\right)\Bigg)$$

$$= \frac{1}{16}\Bigg(-4(e^{2k\alpha^2}+e^{-2k\alpha^2})+e^{2k\alpha^2(3+2\alpha)}+e^{2k\alpha^2(-1+2\alpha)}+2e^{-2k\alpha^2-4k\alpha^3}$$

$$+e^{-2k\alpha^2(1+2\alpha)}+e^{2k\alpha^2(3-2\alpha)}+2e^{-2k\alpha^2+4k\alpha^3}\Bigg)$$

$$= O(k^2\alpha^4) \ , \hspace{3cm} \text{(Taylor series expansion in }\alpha)$$

giving that indeed $\sum_{\ell=0}^{\infty}\Pr[A=\ell](1-\varphi(\epsilon,A))^2 = O\left(\frac{\epsilon^4 k^2}{n^2}\right)$. This completes the proof. ∎

Proof [Proof of Lemma 14] Using Hellinger distance as a proxy will only result in an $\Omega(\epsilon^2)$ lower bound on the distance, so we compute it explicitly instead: in what follows, $e^{(j)}\in\{0,1\}^n$ denotes the basis vector with $e_i^{(j)}=\mathbf{1}_{\{i=j\}}$. Fix any vector $b=(b_1,\ldots,b_n)\in\{0,1\}^n$ such that $|b|\in[n/3,2n/3]$, and let P be the corresponding distribution from the support of $\mathcal{N}$.

$$\|P-P^*\|_1 \geq \sum_{j=1}^{n}|P(e^{(j)})-P^*(e^{(j)})|$$

$$= \sum_{j=1}^{n}\left|\frac{1+(-1)^{b_j}\epsilon}{n}\prod_{i\neq j}\left(1-\frac{1+(-1)^{b_i}\epsilon}{n}\right)-\frac{1}{n}\left(1-\frac{1}{n}\right)^{n-1}\right|$$

$$= \frac{1}{n}\left(1-\frac{1}{n}\right)^{n-1}\sum_{j=1}^{n}\left|(1+(-1)^{b_j}\epsilon)\prod_{i\neq j}\left(1-\frac{(-1)^{b_i}\epsilon}{n-1}\right)-1\right| \ .$$

Each summand can be bounded from above as follows:

$$\left(1-\frac{\epsilon}{n-1}\right)^{2n/3} \leq \prod_{i\neq j}\left(1-\frac{(-1)^{b_i}\epsilon}{n-1}\right) \leq \left(1+\frac{\epsilon}{n-1}\right)^{2n/3} \ ,$$

where the last inequality follows from our assumption on $|b|$. In turn, this gives that

- If $b_j = 0$,

$$(1 + (-1)^{b_j}\epsilon) \prod_{i \neq j} \left(1 - \frac{(-1)^{b_i}\epsilon}{n-1}\right) - 1 \geq (1+\epsilon)\left(1 - \frac{\epsilon}{n-1}\right)^{2n/3} - 1 = \Omega(\epsilon) \ .$$

- If $b_j = 1$,

$$1 - (1 + (-1)^{b_j}\epsilon) \prod_{i \neq j} \left(1 - \frac{(-1)^{b_i}\epsilon}{n-1}\right) \geq 1 - (1-\epsilon)\left(1 + \frac{\epsilon}{n-1}\right)^{2n/3} = \Omega(\epsilon) \ .$$

Since $\frac{1}{n}\left(1 - \frac{1}{n}\right)^{n-1} = \frac{e^{-1}+o(1)}{n}$, we get $\|P - P^*\|_1 = \Omega(\epsilon)$. The lemma now follows from the fact that a uniformly random $b \in \{0,1\}^n$ satisfies $|b| \in [n/3, 2n/3]$ with probability $1 - 2^{-\Omega(n)}$. $\blacksquare$

Appendix C. Omitted Proofs from Section 4

We provide in this section the proofs of the inequalities stated in the preliminaries (Section 4).
Proof [Proof of Lemma 26] Using properties of the KL-divergence:

$$\mathrm{D}(P\|Q) = \mathrm{D}(P_1 \otimes \cdots \otimes P_n \| Q_1 \otimes \cdots \otimes Q_n) = \sum_{i=1}^n \mathrm{D}(P_i\|Q_i)$$

so it suffices to show that $2(p_i - q_i)^2 \leq \mathrm{D}(P_i\|Q_i) \leq \frac{(p_i - q_i)^2}{q_i(1-q_i)}$ for all $i \in [n]$. Since $P_i = \mathrm{Bern}(p_i)$ and $Q_i = \mathrm{Bern}(q_i)$, we can write

$$\mathrm{D}(P_i\|Q_i) = p_i \ln \frac{p_i}{q_i} + (1-p_i) \ln \frac{1-p_i}{1-q_i}$$

Defining $f \colon (0,1)^2 \to \mathbb{R}$ as $f(x,y) \stackrel{\mathrm{def}}{=} x \ln \frac{x}{y} + (1-x) \ln \frac{1-x}{1-y}$, we thus have to show that $2(x-y)^2 \leq f(x,y) \leq \frac{(x-y)^2}{y(1-y)}$ for all $x, y \in (0,1)$.

We begin by the upper bound: fixing any $x, y \in (0,1)$ and setting $\delta \stackrel{\mathrm{def}}{=} x - y$, we have

$$f(x,y) = x \ln\left(1 + \frac{\delta}{y}\right) + (1-x)\ln\left(1 - \frac{\delta}{1-y}\right) \leq \frac{x}{y}\delta - \frac{1-x}{1-y}\delta = \frac{\delta^2}{y(1-y)}$$

from $\ln(1+u) \leq u$ for all $u \in (-1, \infty)$.

Turning to the lower bound, we fix any $y \in (0,1)$ and consider the auxiliary function $h_y \colon (0,1) \to \mathbb{R}$ defined by $h_y(x) = f(x,y) - 2(x-y)^2$. From $h_y''(x) = \frac{(1-2x)^2}{x(1-x)} \geq 0$, we get that h_y is convex, i.e. h_y' is non-decreasing. Since $h_y'(x) = \ln \frac{x(1-y)}{(1-x)y} - 4(x-y)$, we have $h_y'(y) = 0$, and in turn we get that h_y is non-increasing on $(0,y]$ and non-decreasing on $[y,1)$. Since $h_y(y) = 0$, this leads to $h_y(x) \geq 0$ for all $x \in (0,1)$, i.e. $f(x,y) \geq 2(x-y)^2$. $\blacksquare$

Proof [Proof of Lemma 28] Recall that for any pair of distributions, we have that $d_H(P,Q)^2 \leq \frac{1}{2}\|P - Q\|_1 \leq \sqrt{2} \, d_H(P,Q)$, where $d_H(P,Q)$ denotes the Hellinger distance between P, Q. Therefore, it is enough to show that

$$\min\left(c', \frac{1}{4}\sum_{i=1}^{n}(p_i - q_i)^2\right) \leq d_H(P,Q)^2 \leq \sum_{i=1}^{n}\frac{(p_i - q_i)^2}{q_i(1 - q_i)} \tag{12}$$

for some absolute constant $c' > 0$. (We will show $c' = 1 - e^{-3/2} \simeq 0.78$.)

Since P, Q are product measures,

$$d_H(P,Q)^2 = 1 - \prod_{i=1}^{n}(1 - d_H(P_i, Q_i)^2) = 1 - \prod_{i=1}^{n}\left(\sqrt{p_i q_i} + \sqrt{(1 - p_i)(1 - q_i)}\right).$$

We start with the lower bound. Noting that for any $x, y \in (0,1)$ it holds that

$$\sqrt{xy} + \sqrt{(1 - x)(1 - y)} \leq 1 - \frac{1}{2}(x - y)^2$$

(e.g., by observing that the function $x \in (0,1) \mapsto 1 - \frac{1}{2}(x - y)^2 - (\sqrt{xy} + \sqrt{(1-x)(1-y)})$ is minimized at y, where it takes value 0), we get

$$d_H(P,Q)^2 \geq 1 - \prod_{i=1}^{n}\left(1 - \frac{1}{2}(p_i - q_i)^2\right) \geq \min(1 - e^{-3/2}, \frac{1}{4}\sum_{i=1}^{n}(p_i - q_i)^2)$$

$$\geq \min(1 - e^{-3/2}, \frac{1}{4}\|p - q\|_2^2)$$

where we relied on the inequality $1 - \prod_{i=1}^{n}(1 - x_i) \geq \frac{1}{2}\sum_{i=1}^{n} x_i$ for $(x_1, \dots, x_n) \in [0,1]$:

$$1 - \prod_{i=1}^{n}(1 - x_i) = 1 - e^{\sum_{i=1}^{n}\ln(1 - x_i)} \geq 1 - e^{-\sum_{i=1}^{n} x_i} \geq 1 - \left(1 - \frac{1}{2}\sum_{i=1}^{n} x_i\right).$$

(the last inequality being true for $\sum_{i=1}^{n} x_i \leq \frac{3}{2}$, i.e. $\|p - q\|_2^2 \leq 3$).

Turning to the upper bound, the elementary inequality $2\sqrt{xy} = x + y - (\sqrt{x} - \sqrt{y})^2$, $x, y > 0$, gives that

$$\sqrt{p_i q_i} + \sqrt{(1 - p_i)(1 - q_i)} \geq 1 - \frac{(p_i - q_i)^2}{(p_i + q_i)(2 - p_i - q_i)} = 1 - z_i.$$

Therefore,

$$d_H(P,Q)^2 \leq (1 - \prod_{i=1}^{n}(1 - z_i)) \leq \sum_{i=1}^{n} z_i = \sum_{i=1}^{n}\frac{(p_i - q_i)^2}{(p_i + q_i)(2 - p_i - q_i)} \leq \sum_{i=1}^{n}\frac{(p_i - q_i)^2}{q_i(1 - q_i)}, \tag{13}$$

where the third-to-last inequality follows from the union bound, and the last from the simple fact that $(p_i + q_i)(2 - p_i - q_i) \geq q_i(1 - q_i)$. ∎

Proof [Proof of Lemma 29] Let A and B be two distributions on $\{0,1\}^d$. Then we have:

$$D(A\|B) = \sum_{x\in\{0,1\}^d} \Pr_A[x] \ln \frac{\Pr_A[x]}{\Pr_B[x]};. \tag{14}$$

For a fixed $i \in [d]$, the events $\Pi_{i,a}$ form a partition of $\{0,1\}^d$. Dividing the sum above into this partition, we obtain

$$
\begin{aligned}
D(A\|B) &= \sum_a \sum_{x\in\Pi_{i,a}} \Pr_A[x] \ln \frac{\Pr_A[x]}{\Pr_B[x]} \\
&= \sum_a \Pr_A[\Pi_{i,a}] \sum_{x\in\{0,1\}^d} \Pr_{A|\Pi_{i,a}}[x] \left(\ln \frac{\Pr_A[\Pi_{i,a}]}{\Pr_B[\Pi_{i,a}]} + \ln \frac{\Pr_{A|\Pi_{i,a}}[x]}{\Pr_{B|\Pi_{i,a}}[x]} \right) \\
&= \sum_a \Pr_A[\Pi_{i,a}] (\ln \frac{\Pr_A[\Pi_{i,a}]}{\Pr_B[\Pi_{i,a}]} + D(A \mid \Pi_{i,a}\|B \mid \Pi_{i,a}) .
\end{aligned}
$$

Let $P_{\leq i}$ be the distribution over the first i coordinates of P and define $Q_{\leq i}$ similarly for Q. Let P_i and Q_i be the distribution of the i-th coordinate of P and Q respectively. We will apply the above to $P_{\leq i-1}$ and $P_{\leq i}$. First note that the i-th coordinate of $P_{\leq i} \mid \Pi_{i,a}$ and $Q_{\leq i} \mid \Pi_{i,a}$ is independent of the others, thus we have (which likely follows from standard results):

$$
\begin{aligned}
&D(P_{\leq i} \mid \Pi_{i,a}\|Q_{\leq i} \mid \Pi_{i,a}) \\
&= \sum_{x\in\{0,1\}^i} \Pr_{P_{\leq i}|\Pi_{i,a}}[x] \ln \frac{\Pr_{P_{\leq i}|\Pi_{i,a}}[x]}{\Pr_{Q_{\leq i}|\Pi_{i,a}}[x]} \\
&= \sum_{x\in\{0,1\}^i} \Pr_{P_{\leq i-1}|\Pi_{i,a}}[x_{\leq i-1}] \Pr_{P_i|\Pi_{i,a}}[x_i] \left(\ln \frac{\Pr_{P_{\leq i-1}|\Pi_{i,a}}[x_{\leq i-1}]}{\Pr_{Q_{\leq i-1}|\Pi_{i,a}}[x_{\leq i-1}]} + \ln \frac{\Pr_{P_i|\Pi_{i,a}}[x_i]}{\Pr_{Q_{\leq i}|\Pi_{i,a}}[x_i]} \right) \\
&= \sum_{x\in\{0,1\}^{i-1}} \Pr_{P_{\leq i-1}|\Pi_{i,a}}[x] \ln \frac{\Pr_{P_{\leq i-1}|\Pi_{i,a}}[x]}{\Pr_{Q_{\leq i-1}|\Pi_{i,a}}[x]} + \sum_{y\in\{0,1\}} \Pr_{P_i|\Pi_{i,a}}[y] \ln \frac{\Pr_{P_i|\Pi_{i,a}}[y]}{\Pr_{Q_{\leq i}|\Pi_{i,a}}[y]} \\
&= D(P_{\leq i-1} \mid \Pi_{i,a}\|Q_{\leq i-1} \mid \Pi_{i,a}) + D(P_i \mid \Pi_{i,a}\|Q_i \mid \Pi_{i,a})
\end{aligned}
$$

Thus, we have:

$$
\begin{aligned}
&D(P_{\leq i}\|Q_{\leq i}) \\
&= \sum_a \Pr_{P_{\leq i}}[\Pi_{i,a}] (\ln \frac{\Pr_{P_{\leq i}}[\Pi_{i,a}]}{\Pr_{Q_{\leq i}}[\Pi_{i,a}]} + D(P_{\leq i} \mid \Pi_{i,a}\|Q_{\leq i} \mid \Pi_{i,a}) \\
&= \sum_a \Pr_{P_{\leq i}}[\Pi_{i,a}] \left(\ln \frac{\Pr_{P_{\leq i}}[\Pi_{i,a}]}{\Pr_{Q_{\leq i}}[\Pi_{i,a}]} + D(P_{\leq i-1} \mid \Pi_{i,a}\|Q_{\leq i-1} \mid \Pi_{i,a}) + D(P_i \mid \Pi_{i,a}\|Q_i \mid \Pi_{i,a}) \right) \\
&= D(P_{\leq i-1}\|Q_{\leq i-1}) + \sum_a \Pr_P[\Pi_{i,a}] D(P_i \mid \Pi_{i,a}\|Q_i \mid \Pi_{i,a})
\end{aligned}
$$

By induction on i, we get

$$D(P\|Q) = \sum_{(i,a)\in S} \Pr_P[\Pi_{i,a}] D(P_i \mid \Pi_{i,a}\|Q_i \mid \Pi_{i,a}).$$

Now the distributions $P_i \mid \Pi_{i,a}$ and $Q_i \mid \Pi_{i,a}$ are Bernoulli random variables with means $p_{i,a}$ and $q_{i,a}$. For $p, q \in [0, 1]$, we have:

$$D(\mathrm{Bern}(p) \| \mathrm{Bern}(q)) = p \ln \frac{p}{q} + (1 - p) \ln \frac{1-p}{1-q} \leq \frac{(p-q)^2}{q(1-q)}$$

as in the proof of Claim 4. On the other hand, studying for instance the function $f_q \colon (0, 1) \to \mathbb{R}$ defined by $f_q(p) = \frac{p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}}{(p-q)^2}$ (extended by continuity at q), we get

$$f_q(p) \geq f_q(1 - q) \geq 2$$

for all $p, q \in (0, 1)^2$. This shows the lower bound. ∎

Appendix D. Omitted Proofs from Sections 2.2 and 8.2

Fact *[Fact 81]*

$$I(X; A) \leq \sum_{a=0}^{\infty} \Pr[A = a] \left(1 - \frac{\Pr[A = a \mid X = 1]}{\Pr[A = a \mid X = 0]} \right)^2. \tag{15}$$

Proof For $a \in \mathbb{N}$, we write $p_a = \Pr[X = 0 \mid A = a]$ and $q_a = \Pr[X = 1 \mid A = a]$, so that $p_a + q_a = 1$. By definition,

$$
\begin{aligned}
I(X; A) &= \sum_{a=0}^{\infty} \Pr[A = a] \sum_{x \in \{0,1\}} \Pr[X = x \mid A = a] \log \frac{\Pr[X = x \mid A = a]}{\Pr[X = x]} \\
&= \sum_{a=0}^{\infty} \Pr[A = a] \left(p_a \log \frac{p_a}{\Pr[X = 1]} + q_a \log \frac{q_a}{\Pr[X = 0]} \right) \\
&= \sum_{a=0}^{\infty} \Pr[A = a] \left(p_a \log (2p_a) + q_a \log (2q_a) \right) \\
&= \sum_{a=0}^{\infty} \Pr[A = a] \left((1 - q_a) \log (1 - q_a) + q_a \log (q_a) + 1 \right) \\
&\leq \sum_{a=0}^{\infty} \Pr[A = a] \left(1 - \frac{q_a}{1 - q_a} \right)^2 = \sum_{a=0}^{\infty} \Pr[A = a] \left(1 - \frac{q_a}{p_a} \right)^2
\end{aligned}
$$

where for the last inequality we rely on the fact that the binary entropy satisfies $h(x) \geq 1 - \left(1 - \frac{x}{1-x} \right)^2$ for all $x \in [0, 1)$. ∎

Proof [Alternative proof of Lemma 15] We proceed as in the proof of Lemma 12, first writing

$$I(X; A) = \sum_{\ell=0}^{\infty} \Pr[A = a] \left(1 - \frac{\Pr[A = a \mid X = 1]}{\Pr[A = a \mid X = 0]} \right)^2 = \mathbb{E}_A \left[\left(1 - \frac{\Pr[A = a \mid X = 1]}{\Pr[A = a \mid X = 0]} \right)^2 \right]$$

and noticing that $z_a \overset{\text{def}}{=} \frac{\Pr[A=a \mid X=1]}{\Pr[A=a \mid X=0]} = \frac{e^{-k\epsilon/n}(1+\epsilon)^a + e^{k\epsilon/n}(1-\epsilon)^a}{2}$. This leads to

$$4(1-z_a)^2 = 4 - 4\left(e^{-\frac{k}{n}\epsilon}(1+\epsilon)^a + e^{\frac{k}{n}\epsilon}(1-\epsilon)^a\right) + \left(e^{-2\frac{k}{n}\epsilon}(1+\epsilon)^{2a} + e^{2\frac{k}{n}\epsilon}(1-\epsilon)^{2a} + 2(1-\epsilon^2)^a\right) \tag{16}$$

We also have, by definition of A, that for any $z \in \mathbb{R}$

$$\mathbb{E}_A[z^A] = \frac{1}{2}\mathbb{E}_{A\sim\text{Poi}\left(\frac{k}{n}\right)}[z^A] + \frac{1}{4}\mathbb{E}_{A\sim\text{Poi}\left(\frac{k}{n}(1+\epsilon)\right)}[z^A] + \frac{1}{4}\mathbb{E}_{A\sim\text{Poi}\left(\frac{k}{n}(1-\epsilon)\right)}[z^A]$$

$$= \frac{1}{2}e^{\frac{k}{n}(z-1)} + \frac{1}{4}e^{\frac{k}{n}(1+\epsilon)(z-1)} + \frac{1}{4}e^{\frac{k}{n}(1-\epsilon)(z-1)} \tag{17}$$

from the expression of the probability generating function of a Poisson random variable. For any $\beta \in \{-1, 1\}$, we therefore have

$$4e^{-\frac{k}{n}\beta\epsilon}\mathbb{E}_A[(1+\beta\epsilon)^A] = 2e^{\frac{k}{n}\beta\epsilon}e^{-\frac{k}{n}\beta\epsilon} + e^{\frac{k}{n}(1+\epsilon)\beta\epsilon}e^{-\frac{k}{n}\beta\epsilon} + e^{\frac{k}{n}(1-\epsilon)\beta\epsilon}e^{-\frac{k}{n}\beta\epsilon}$$

$$= 2 + e^{\frac{k}{n}\beta\epsilon^2} + e^{-\frac{k}{n}\beta\epsilon^2}$$

$$4e^{-2\frac{k}{n}\beta\epsilon}\mathbb{E}_A[(1+\beta\epsilon)^{2A}] = 2e^{\frac{k}{n}(2\beta\epsilon+\epsilon^2)}e^{-2\frac{k}{n}\beta\epsilon} + e^{\frac{k}{n}(1+\epsilon)(2\beta\epsilon+\epsilon^2)}e^{-2\frac{k}{n}\beta\epsilon} + e^{\frac{k}{n}(1-\epsilon)(2\beta\epsilon+\epsilon^2)}e^{-2\frac{k}{n}\beta\epsilon}$$

$$= 2e^{\frac{k}{n}\epsilon^2} + e^{\frac{k}{n}((1+2\beta)\epsilon^2+\epsilon^3)} + e^{\frac{k}{n}((1-2\beta)\epsilon^2-\epsilon^3)}$$

$$4 \cdot 2\mathbb{E}_A[(1-\epsilon^2)^A] = 4e^{\frac{k}{n}\epsilon^2} + 2e^{\frac{k}{n}(1+\epsilon)\epsilon^2} + 2e^{\frac{k}{n}(1-\epsilon)\epsilon^2} = 4e^{\frac{k}{n}\epsilon^2} + 2e^{\frac{k}{n}(\epsilon^2+\epsilon^3)} + 2e^{\frac{k}{n}(\epsilon^2-\epsilon^3)}$$

Combining Eq. (16) and the above, we obtain

$$16\mathbb{E}_A\left[(1-z_A)^2\right] = 16 - 4\left(4e^{-\frac{k}{n}\epsilon}\mathbb{E}_A[(1+\epsilon)^A] + 4e^{\frac{k}{n}\epsilon}\mathbb{E}_A[(1-\epsilon)^A]\right)$$

$$+ \left(4e^{-2\frac{k}{n}\epsilon}\mathbb{E}_A[(1+\epsilon)^{2A}] + 4e^{2\frac{k}{n}\epsilon}\mathbb{E}_A[(1-\epsilon)^{2A}] + 8\mathbb{E}_A[(1-\epsilon^2)^A]\right)$$

$$= 16 - 8\left(2 + e^{\frac{k}{n}\epsilon^2} + e^{-\frac{k}{n}\epsilon^2}\right) + \left(4e^{\frac{k}{n}\epsilon^2} + e^{\frac{k}{n}(3\epsilon^2+\epsilon^3)} + e^{\frac{k}{n}(-\epsilon^2-\epsilon^3)}\right.$$

$$+ e^{\frac{k}{n}(-\epsilon^2+\epsilon^3)} + e^{\frac{k}{n}(3\epsilon^2-\epsilon^3)} + 4e^{\frac{k}{n}\epsilon^2} + 2e^{\frac{k}{n}(\epsilon^2+\epsilon^3)} + 2e^{\frac{k}{n}(\epsilon^2-\epsilon^3)}\Big)$$

$$= 8 - 8\left(e^{\frac{k}{n}\epsilon^2} + e^{-\frac{k}{n}\epsilon^2}\right) + \left(e^{\frac{k}{n}3\epsilon^2} + e^{-\frac{k}{n}\epsilon^2} + 2e^{\frac{k}{n}\epsilon^2}\right)\left(e^{\frac{k}{n}\epsilon^3} + e^{-\frac{k}{n}\epsilon^3}\right).$$

A Taylor expansion of this expression (in $\frac{k\epsilon^2}{n}$ for the first two parentheses, and $\frac{k\epsilon^3}{n}$ for the last) shows that

$$\mathbb{E}_A\left[(1-z_A)^2\right] = O\left(\frac{k^2\epsilon^4}{n^2}\right)$$

as claimed. ∎

∎

Proof [Proof of Lemma 63] From the definition of

$$I(X;Y) = \sum_{(x,y)\in\{0,1\}} \Pr[X=x, Y=y]\log\frac{\Pr[X=x, Y=y]}{\Pr[X=x]\Pr[Y=y]},$$

it is straightforward to check that for X, Y taking values in $\{0, 1\}$

$$
\begin{aligned}
I\left(X;Y\right) &= \mathbb{E}[XY]\log\frac{\mathbb{E}[XY]}{\mathbb{E}[X]\,\mathbb{E}[Y]} \\
&+ (\mathbb{E}[X] - \mathbb{E}[XY])\log\frac{\mathbb{E}[X] - \mathbb{E}[XY]}{\mathbb{E}[X]\,(1 - \mathbb{E}[Y])} \\
&+ (\mathbb{E}[Y] - \mathbb{E}[XY])\log\frac{\mathbb{E}[Y] - \mathbb{E}[XY]}{(1 - \mathbb{E}[X])\mathbb{E}[Y]} \\
&+ (1 - \mathbb{E}[X] - \mathbb{E}[Y] + \mathbb{E}[XY])\log\frac{1 - \mathbb{E}[X] - \mathbb{E}[Y] + \mathbb{E}[XY]}{(1 - \mathbb{E}[X])(1 - \mathbb{E}[Y])} \\
&= f(\mathbb{E}[X], \mathbb{E}[Y], \mathbb{E}[XY])
\end{aligned}
$$

for f defined by $f(x, y, z) \overset{\text{def}}{=} z\log\frac{z}{xy} + (x - z)\log\frac{x-z}{x(1-y)} + (y - z)\log\frac{y-z}{(1-x)y} + (1 - x - y + z)\log\frac{1-x-y+z}{(1-x)(1-y)}$.

The domain of definition of f is the subset $\Omega \subseteq [0, 1]^3$ defined by (recalling that x, y, z correspond to $\mathbb{E}[X], \mathbb{E}[Y], \mathbb{E}[XY]$ for $X, Y \in \{0, 1\}$)

$$
\begin{aligned}
&0 \le x, y \le 1 \\
&0 \le z \le \min(x, y) \\
&0 \le z \le \sqrt{xy} \qquad\qquad\qquad\text{(Cauchy–Schwarz)} \\
&0 \le 1 + z - x - y
\end{aligned}
$$

Given the c-balancedness assumption on P, $\Omega_c \subseteq \Omega$ satisfies the further following constraints:

$$
\begin{aligned}
&c \le x, y \le 1 - c \\
&c^2 \le \frac{z}{x}, \frac{z}{y} \le 1 - c^2 \\
&c^2 \le \frac{1+z-x-y}{1-x}, \frac{1+z-x-y}{1-y} \le 1 - c^2
\end{aligned}
$$

by Baye's rule and recalling that $1 + z - x - y$ corresponds to

$$
\Pr[X = 0, Y = 0] = \Pr[X = 0 \mid Y = 0]\Pr[Y = 0] = \Pr[Y = 0 \mid X = 0]\Pr[X = 0]
$$

while $1 - x, 1 - y$ correspond to $\Pr[X = 0], \Pr[Y = 0]$ respectively.
One can then check that

$$
\begin{aligned}
\frac{\partial f}{\partial x}(x, y, z) &= \log\frac{(1-x)(x-z)}{x(1+z-x-y)}, \\
\frac{\partial f}{\partial y}(x, y, z) &= \log\frac{(1-y)(y-z)}{y(1+z-x-y)}, \\
\frac{\partial f}{\partial z}(x, y, z) &= \log\frac{z(1+z-x-y)}{(x-z)(y-z)}
\end{aligned}
$$

and therefore, on Ω_c, that

$$
\left\|\frac{\partial f}{\partial x}\right\|_\infty = \left\|\frac{\partial f}{\partial y}\right\|_\infty \le \sup_{(x,y,z)\in\Omega_c}|\log\frac{x-z}{x}| + |\log\frac{1+z-x-y}{1-x}| \le 2\log\frac{1}{c^2} = 4\log\frac{1}{c}.
$$

78

Similarly,

$$
\begin{aligned}
\|\frac{\partial f}{\partial z}\|_\infty &\leq \sup_{(x,y,z)\in\Omega_c} |\log \frac{(x-z)(y-z)}{z}| + |\log(1+z-x-y)| \\
&\leq \sup_{(x,y,z)\in\Omega_c} |\log \frac{(1-c^2)y}{c^4 xy}| + |\log \frac{1}{c^2 x}| \\
&\leq \log \frac{1}{c^5} + \log \frac{1}{c^3} = 8\log \frac{1}{c} \ .
\end{aligned}
$$

So overall, f is λ-Lipschitz (with regard to the $\|\cdot\|_\infty$ norm) on Ω_c, for $\lambda = \|\frac{\partial f}{\partial x}\|_\infty + \|\frac{\partial f}{\partial y}\|_\infty + \|\frac{\partial f}{\partial z}\|_\infty \leq 16\log \frac{1}{c}$. ∎