Independence Testing for Bounded Degree Bayesian Networks

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

Clément L. Canonne
School of Computer Science
University of Sydney
clement.canonne@sydney.edu.au

Joy Qiping Yang
School of Computing
National University of Singapore
joy.yang@nus.edu.sg

Abstract

We study the following *independence testing* problem: given access to samples from a distribution P over $\{0,1\}^n$, decide whether P is a product distribution or whether it is ε-far in total variation distance from any product distribution. For arbitrary distributions, this problem requires $\exp(n)$ samples. We show in this work that if P has a sparse structure, then in fact only linearly many samples are required. Specifically, if P is Markov with respect to a Bayesian network whose underlying DAG has in-degree bounded by d, then $\tilde{\Theta}(2^{d/2} \cdot n/\varepsilon^2)$ samples are necessary and sufficient for independence testing.

1 Introduction

It is often convenient to model high-dimensional datasets as probability distributions. An important reason is that the language of probability formalizes what we intuitively mean by two features of the data being *independent*: the marginal distributions along the two corresponding coordinates are statistically independent. Independence is a basic probabilistic property that, when true, enables better interpretability of data as well as computationally fast inference algorithms.

In this work, we study the problem of testing whether a collection of n binary random variables are mutually independent. Independence testing is an old and foundational problem in hypothesis testing [Neyman and Pearson, 1933, Lehmann et al., 2005], and we consider it from the perspective of *property testing* [Rubinfeld and Sudan, 1996, Goldreich et al., 1998]. Our goal is to design a tester that accepts product distributions on n variables and rejects distributions that have statistical distance at least ε from every product distribution. The tester gets access to i.i.d. samples from the input distribution and can fail (in either case) with probability at most $1/3$. The objective is to minimize the number of samples as a function of n and ε.

Previous work [Diakonikolas and Kane, 2016] shows[1] that testing independence of a distribution on $\{0,1\}^n$ requires $2^{\Omega(n)}$ queries which is intractable. However, there is an aspect of this proof that is unsatisfying. The "hard distributions", which are shown to require many queries to distinguish from product distributions, cannot be described succinctly, as they require exponentially many bits to describe. Thus, a natural question arises: can we test independence efficiently for distributions on $\{0,1\}^n$ that have a sparse description?

[1] See Appendix D for a more general result.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

One of the most canonical ways to describe high-dimensional distributions is as *Bayesian networks* (or *Bayes nets* in short). A Bayes net specifies how to generate an n-dimensional sample in an iterative way and is especially useful for modeling causal relationships. Formally, a Bayes net on $\{0,1\}^n$ is given by a directed acyclic graph (DAG) G on n vertices and probability distributions $p_{i,\pi}$ on $\{0,1\}$ for all $i \in [n]$ and all assignments π to the parents of the i'th node in the graph G. An n-dimensional sample is obtained by sampling the nodes in a topological order of G, where the i'th node is sampled according to $p_{i,\pi}$ for the assignment π that is already fixed by the samples of the parent nodes of i. The generated distribution on $\{0,1\}^n$ is said to be *Markov with respect to* G.

In this work, we consider independence testing on distributions having a sparse Bayes net description, a class of distributions naturally arising in, and with numerous applications to, machine learning [Wainwright and Jordan, 2008], robotics, natural language processing, medicine, and biological settings such as gene expression data [Friedman et al., 2000, Peng et al., 2009, Gardner et al., 2003] (where it is known that most genetic networks are actually sparse). In these cases, one hopes to leverage that additional knowledge to test whether those sparse, local dependencies are actually present without having to pay the prohibitive exponential cost in the dimension n. Specifically, we analyze independence testing on distributions that are promised to be Markov with respect to a DAG of maximum in-degree d, where $d \ll n$. While the learning sample complexity, known to be $\tilde{O}(2^d \cdot n/\varepsilon^2)$ [Bhattacharyya et al., 2020], provides a baseline for the *testing* question, it is not at all obvious that this is tight, and what the correct dependence on d and even n are. Our main result essentially settles this question, and establishes the following:

Theorem 1.1 (Informal Main Theorem)**.** *Suppose an unknown distribution P on $\{0,1\}^n$ is Markov with respect to an unspecified degree-d DAG. The sample complexity of testing whether P is a product distribution or is at least ε-far from any product distribution is $\tilde{\Theta}(2^{d/2} \cdot n/\varepsilon^2)$.*

In the course of proving this theorem, we additionally derive several technical results that are of independent interest, such as bounds on the moment generating function of squared binomials and an independence testing algorithm for arbitrary distributions in Hellinger distance.

Our work explicitly initiates the study of *testing graphical structure* in the context of property testing. That is, instead of testing a statistical property of a distribution (as is typical in distribution property testing), we can interpret our problem as that of testing a graphical property of the underlying graph that describes the distribution. In particular, testing independence can be viewed as testing maximum degree-0 of a distribution's graph. This point of view opens the door to testing many other relevant graphical properties of graphical models, e.g., maximum degree k, being a forest, being connected, etc. Hence, by analyzing the "base case" of independence testing, we provide a necessary first step to solving other graphical testing problems. Information-theoretic bounds for related problems were studied recently by Neykov et al. [2019].

1.1 Related Work

Distribution testing has been an active and rapidly progressing research program for the last 20+ years; see Rubinfeld [2012] and Canonne [2020] for surveys. One of the earliest works in this history was that of Batu et al. [2001] who studied testing independence of two random variables. There followed a series of papers [Alon et al., 2007, Levi et al., 2013, Acharya et al., 2015], strengthening and generalizing bounds for this problem, culminating in the work of Diakonikolas and Kane [2016] who gave tight bounds for testing independence of distributions over $[n_1] \times \cdots \times [n_d]$. Hao and Li [2020] recently considered the (harder) problem of estimating the distance to the closest product distribution (i.e., *tolerant* testing), showing this task could, too, be performed with a sublinear sample complexity.

Though most of the focus has been on testing properties of arbitrary input distributions, it has long been recognized that distributional restrictions are needed to obtain sample complexity improvements. For example, Rubinfeld and Servedio [2009], Adamaszek et al. [2010] studied testing uniformity of monotone distributions on the hypercube. Similarly, Daskalakis et al. [2012] considered the problem of testing monotonicity of k-modal distributions. The question of independence testing of structured high-dimensional distributions was considered in Daskalakis et al. [2019] in the context of Ising models. We note that while their work is in the same spirit as ours, Ising models and Bayes nets are incomparable modeling assumptions, and their results (and techniques) and ours are mostly disjoint. Further, while the results may overlap in some special cases, the conversion between

parameterizations makes them difficult to compare even in these cases (e.g., dependence on the maximum edge value parameter β for Ising models, and max-undirected-degree vs. max-in-degree). More recently, Canonne et al. [2020], Daskalakis and Pan [2017], Bhattacharyya et al. [2020, 2021] have studied identity testing and closeness testing for distributions that are structured as degree-d Bayes nets. Our work here continues this research direction in the context of independence testing.

1.2 Our techniques

Upper bound. The starting point of our upper bound is the (standard) observation that a distribution P over $\{0,1\}^n$ is far from being a product if, and only if, it is far from the product of its marginals. By itself, this would not lead to any savings over the trivial exponential sample complexity. However, we can combine this with a localization result due to Daskalakis and Pan [2017], which then guarantees that if the degree-d Bayes net P is at *Hellinger* distance ε from the product of its marginals P', then there exists *some* vertex $i \in [n]$ such that P_{i,Π_i} (the marginalization of P onto the set of nodes consisting of i and its d parents) is at Hellinger distance at least $\frac{\varepsilon}{\sqrt{n}}$ from P'_{i,Π_i}. These two facts, combined, seem to provide exactly what is needed: indeed, given access to samples from P and any fixed set of $d+1$ vertices S, one can simulate easily samples from both P_S and P'_S (for the second, using $d+1$ samples from P to generate one from P'_S, as P'_S is the product of marginals of P_S). A natural idea is then to iterate over all $\binom{n}{d+1}$ possible subsets S of $d+1$ variables and check whether $P_S = P'_S$ for each of them using a closeness testing algorithm for arbitrary distributions over $\{0,1\}^{d+1}$: the overhead due to a union bound and the sampling process for P_S, P'_S adds a factor $O(d \cdot \log \binom{n}{d+1}) = O(d^2 \log n)$ to the closeness testing procedure. However, since testing closeness over $\{0,1\}^{d+1}$ to total variation distance ε' has sample complexity $O(2^{2d/3}/\varepsilon'^2)$ and, by the quadratic relation between Hellinger and total variation distances, we need to take $\varepsilon' = \frac{\varepsilon^2}{n}$, we would then expect the overall test to result in a $\tilde{O}(2^{2d/3}n^2/\varepsilon^4)$ sample complexity – much more than what we set out for.

A first natural idea to improve upon this is to avoid the back and forth between Hellinger and total variation distance, and save on this quadratic blowup. Namely, by using an optimal closeness testing algorithm directly for Hellinger distance [Diakonikolas and Kane, 2016], we could in the last step keep $\varepsilon' = \frac{\varepsilon}{\sqrt{n}}$ (saving on this quadratic blowup), and pay only overall $\tilde{O}(2^{3d/4}/\varepsilon'^2) = \tilde{O}(2^{3d/4}n/\varepsilon^2)$. This is better, but still falls short of our original goal.

The second idea is to forego closeness testing in the last step entirely, and instead use directly an *independence* testing algorithm for arbitrary distributions over $\{0,1\}^{d+1}$, to test if P_S is indeed a product distribution for every choice of S considered. Unfortunately, while promising, this idea suffers from a similar drawback as our very first attempt: namely, the known independence testing algorithms are all designed for testing in total variation distance (not Hellinger)! Thus, even using an optimal TV testing algorithm for independence [Acharya et al., 2015, Diakonikolas and Kane, 2016] would still lead to this quadratic loss in the distance parameter ε', and a resulting $\tilde{O}(2^{d/2}n^2/\varepsilon^4)$ sample complexity.

To combine the best of our last two approaches and achieve the claimed $\tilde{O}(2^{d/2}n/\varepsilon^2)$, we combine the two insights and perform, for each set S of $d+1$ variables, independence testing on P_S in Hellinger distance. In order to do so, however, we first need to design a testing algorithm for this task, as none was previously available in the literature. Fortunately for us, we are able to design such a testing algorithm (Lemma 3.2) achieving the desired – and optimal – sample complexity. Combining this Hellinger independence testing algorithm over $\{0,1\}^{d+1}$ with the above outline finally leads to the $\tilde{O}(2^{d/2}n/\varepsilon^2)$ upper bound of Theorem 1.1.

Lower bound. To obtain our $\Omega(2^{d/2}n/\varepsilon^2)$ lower bound on testing independence of a degree-d Bayes net, we start with the construction introduced by Canonne et al. [2020] to prove an $\Omega(n/\varepsilon^2)$ sample complexity lower bound on testing *uniformity* of degree-1 Bayes nets. At a high level, this construction relies on picking uniformly at random a perfect matching M of the n vertices, which defines the structure of the Bayes net; and, for each of the $n/2$ resulting edges, picking either a positive or negative correlation (with value $\pm\varepsilon/\sqrt{n}$) between the two vertices, again uniformly at random. One can check relatively easily that every Bayes net P_λ obtained this way, where λ encodes the matching M and the $2^{n/2}$ signs, is (1) a degree-1 Bayes net, (2) at total variation ε from the uniform distribution U. The bulk of their analysis then lies in showing that (3) $\Omega(n/\varepsilon^2)$ samples are necessary to distinguish between U and such a randomly chosen P_λ. Generalizing this lower bound

construction and analysis to independence testing (not just uniformity), and to degree-d (and not just degree-1) Bayes nets turns out to be highly non-trivial, and is our main technical contribution.

Indeed, in view of the simpler and different $\Omega(2^{d/2}\sqrt{n}/\varepsilon^2)$ sample complexity lower bound for uniformity testing degree-d Bayes nets obtained in Canonne et al. [2020] *when the structure of the Bayes net is known*,[2] one is tempted to adapt the same idea to the matching construction: that is, reserve $d-1$ out of the n vertices to "encode" a pointer towards one of 2^{d-1} independently chosen P_λ's as above (i.e., the hard instances are now uniform mixtures over 2^{d-1} independently generated degree-1 hard instances). The degree-1 hard instances lead in previous work to a tighter dependence on n (linear instead of $\sqrt{n}$) because their Bayesian structure is *unknown*. Thus, by looking at the mixture of these degree-1 Bayes nets, one could hope to extend the analysis of that second lower bound from Canonne et al. [2020] and get the desired $\Omega(2^{d/2}n/\varepsilon^2)$ lower bound. Unfortunately, there is a major issue with this idea: namely, if the 2^{d-1} matchings are chosen independently, then the resulting overall distribution is unlikely to be a degree-d Bayes net – instead, each vertex will have expected degree $\Omega(2^d)$! (This was not an issue in the corresponding lower bound of Canonne et al. [2020], since for them each of the 2^d components of the mixture was a degree-0 Bayes net, i.e., a product distribution; so degrees could not "add up" across the components).

To circumvent this, we instead choose the matching M to be common to all 2^{d-1} components of the mixture and only pick the sign of their $n/2$ correlations independently; thus ensuring that every node in the resulting distribution P has degree d. This comes at a price, however: the analysis of the $\Omega(2^{d/2}\sqrt{n}/\varepsilon^2)$ lower bound from Canonne et al. [2020] crucially relied on independence across those components, and thus can no longer be extended to our case (where the 2^{d-1} distributions P_λ share the same matching M, and thus we only have independence across components conditioned on M). Handling this requires entirely new ideas, and constitutes the core of our lower bound. In particular, from a technical point of view this requires us to handle the moment-generating-function of squares of Binomials, as well as that of (squared) truncated Binomials. To do so, we develop a range of results on Binomials and Multinomial distributions we believe are of independent interest.

Finally, after establishing that $\Omega(2^{d/2}n/\varepsilon^2)$ samples are necessary to distinguish the resulting "mixture of trees" P from the uniform distribution U (Lemma 2.3), it remains to show that this implies our stronger statement on testing *independence* (not just uniformity). To do so, we need to show that P is not only far from U, but from *every* product distribution: doing so is itself far from immediate, and is established in Lemma 2.4 by relating the distance from the mixture P to every product distribution to the distance between distinct components of the mixture, and lower bounding those directly by analyzing the concentration properties of each component P_λ of our construction.

Preliminaries. We use the standard asymptotic notation $O(\cdot)$, $\Omega(\cdot)$ $\Theta(\cdot)$, and write $\tilde{O}(\cdot)$ to omit polylogarithmic factors in the argument. Throughout, we identify probability distributions over discrete sets with their probability mass functions (pmf), and further denote by U (resp., U_d) the uniform distribution on $\{0,1\}^n$ (resp., $\{0,1\}^d$). We also write $P^{\otimes m}$ for the m-fold product of a distribution P, that is, $P \otimes \cdots \otimes P$ (the distribution of a tuple of m i.i.d. samples from P); and $[n]$ for the set $\{1,\dots,n\}$.

Bayesian networks. Given a directed acyclic graph (DAG) over n nodes, a probability distribution P over $\{0,1\}^n$ is said to be *Markov with respect to* G if P factorizes according to G; we will also say that P has *structure* G. A DAG G is said to have in-degree d, if every node has at most d parents; for convenience, we use *degree-d* exclusively as "in-degree d" throughout the paper; we will denote by $\Pi_i \subseteq [n]$ the set of parents of a node i. Finally, a distribution P over $\{0,1\}^n$ is a *degree-d Bayes net* if P is Markov w.r.t. some degree-d DAG.

Distances between distributions. Given two distributions P, Q over the same (discrete) domain $\mathcal{X}$, the *total variation distance* (TV) between P and Q is defined as

$$d_{\mathrm{TV}}(P,Q) = \sup_{S\subseteq\mathcal{X}} (P(S)-Q(S)) = \frac{1}{2}\sum_{x\in\mathcal{X}}^{n} |P(x)-Q(x)| \in [0,1]\,. \tag{1}$$

While TV distance will be our main focus, we will also rely in our proofs on two other notions of distance between distributions: the *Hellinger distance*, given by $d_{\mathrm{H}}(P,Q) = \frac{1}{\sqrt{2}}\|\sqrt{P}-\sqrt{Q}\|_2$, and

[2] We note that generalizing this (weaker, in view of the dependence on n) $\Omega(2^{d/2}\sqrt{n}/\varepsilon^2)$ sample complexity lower bound to our setting is relatively simple, and we do so in Appendix D.1 (specifically, Theorem D.3).

the *chi-squared divergence*, defined by $d_{\chi^2}(P,Q) = \sum_{x\in\mathcal{X}} (P(x)-Q(x))^2/Q(x)$. TV distance, squared Hellinger distance, and χ^2 divergence are all instances of f-divergences, and as such satisfy the data processing inequality; further, they are related by the following sequence of inequalities:

$$d_{\mathrm{H}}^2(P,Q) \leqslant d_{\mathrm{TV}}(P,Q) \leqslant \sqrt{2} d_{\mathrm{H}}(P,Q) \leqslant \sqrt{d_{\chi^2}(P,Q)} \tag{2}$$

Tools from previous work. We finally state results from the literature which we will rely upon.

Corollary 1.2 (Daskalakis and Pan [2017, Corollary 2.4])**.** *Suppose P and Q are distributions on Σ^n with common factorization structure*

$$P(x) = P_{X_1}(x_1) \prod_{i=2}^{n} P_{X_i|X_{\Pi_i}}(x_i|x_{\Pi_i}), \qquad Q(x) = Q_{X_1}(x_1) \prod_{i=2}^{n} Q_{X_i|X_{\Pi_i}}(x_i|x_{\Pi_i}), \quad x \in \Sigma^n$$

where we assume the nodes are topologically ordered, and Π_i is the set of parents of i. Then

$$d_{\mathrm{H}}^2(P,Q) \leqslant d_{\mathrm{H}}^2(P_{X_1}, Q_{X_1}) + d_{\mathrm{H}}^2\big(P_{X_2,X_{\Pi_2}}, Q_{X_2,X_{\Pi_2}}\big) + \cdots + d_{\mathrm{H}}^2\big(P_{X_n,X_{\Pi_n}}, Q_{X_n,X_{\Pi_n}}\big).$$

In particular, if $d_{\mathrm{H}}^2(P,Q) \geqslant \varepsilon$ then there exists some i such that $d_{\mathrm{H}}^2\big(P_{X_i,X_{\Pi_i}}, Q_{X_i,X_{\Pi_i}}\big) \geqslant \frac{\varepsilon}{n}$.

Lemma 1.3 (Acharya et al. [2015, Lemma 4])**.** *There exists an efficient algorithm which, given samples from a distribution P over $[n_1]\times\cdots\times[n_d]$, outputs a product distribution Q over $[n_1]\times\cdots\times[n_d]$ such that, if P is a product distribution, then $d_{\chi^2}(p,q) \leqslant \varepsilon^2$. with probability at least 5/6. The algorithm uses $O((\sum_{\ell=1}^{d} n_\ell)/\varepsilon^2)$ samples from P.*

Theorem 1.4 (Daskalakis et al. [2018, Theorem 1])**.** *There exists an efficient algorithm which, given samples from a distribution P and a known reference distribution Q over $[n]$, as well as a distance parameter $\varepsilon \in (0,1]$, distinguishes between the cases (i) $d_{\chi^2}(P,Q) \leqslant \varepsilon^2$ and (ii) $d_{\mathrm{H}}(P,Q) \geqslant \varepsilon$ with probability at least $5/6$. The algorithm uses $O\big(n^{1/2}/\varepsilon^2\big)$ samples from P.*

Lemma 1.5 (Batu et al. [2001, Proposition 1])**.** *Let P be a discrete distribution supported on $[n]\times[m]$, with marginals P_1 and P_2. Then, we have $\min_{Q_1,Q_2} d_{\mathrm{TV}}(P, Q_1\otimes Q_2) \geqslant \frac{1}{3} d_{\mathrm{TV}}(P, P_1\otimes P_2)$, where the minimum is taken over all distributions Q_1 over $[n]$ and Q_2 over $[m]$.*

2 The $\Omega(2^{d/2}n/\varepsilon^2)$ Lower Bound

In this section, we prove our main technical result: a lower bound on the sample complexity of testing whether a degree-d Bayes net is a product distribution. In the interest of space, we omit some details of the derivations: the interested reader can find them in Appendix A.

Theorem 2.1. *Let $d \leqslant c\cdot\log n$, where $c>0$ is a sufficiently small absolute constant. Then, testing whether an arbitrary degree-d Bayes net over $\{0,1\}^n$ is a product distribution or is ε-far from every product distribution requires $\Omega(2^{d/2}n/\varepsilon^2)$ samples.*

This theorem considerably generalizes the lower bound of $\Omega(n/\varepsilon^2)$ established in Canonne et al. [2020] for the case of trees[3] ($d=1$). To establish our result, we build upon (and considerably extend) their analysis; in particular, we will rely on the following "mixture of trees" construction, which can be seen as a careful mixture of (2^{d-1} of) the hard instances from the lower bound of [Canonne et al., 2020, Theorem 14]. Figure 1 shows an illustrative example of our lower bound constructions.

Notation. Throughout, for given n, d, ε, we let $N \coloneqq n-d+1$ (without loss of generality assumed to be even), $D \coloneqq 2^{d-1}$, $\delta \coloneqq \frac{\varepsilon}{\sqrt{n}}$, and

$$z_0 = \frac{1+4\delta}{1-4\delta}, \quad z_1 = \frac{1+(4\delta)^2}{1-(4\delta)^2}, \quad z_2 = \frac{1+(4\delta)^4}{1-(4\delta)^4}.$$

Definition 2.2 (Mixture of Trees)**.** Given parameters $0 \leqslant d \leqslant n$ and $\delta \in (0,1]$, we define the probability distribution $\mathcal{D}_{n,d,\delta}$ over degree-d Bayes nets by the following process.

[3] While this can be better visualized as a forest, we use the term "tree" to refer to the structure of degree-1 Bayes nets. Technically, it still factorizes as a Bayesian path (tree), but not all edges are necessary.

1. Choose a perfect matching λ of $[N]$ uniformly at random (where $N = n - d + 1$), i.e., a set of $N/2$ disjoint pairs;
2. Draw i.i.d. $\mu_1, \ldots, \mu_D$ uniformly at random in $\{0,1\}^{N/2}$;
3. For $\ell \in [D]$, let p_ℓ be the distribution over $\{0,1\}^N$ defined as the Bayes net with tree structure λ, such that if $\lambda_k = (i,j) \in \lambda$ then the corresponding covariance between variables X_i, X_j is
$$\mathrm{Cov}(X_i, X_j) = (-1)^{\mu_{\ell,k}}\delta$$
4. Let the resulting distribution $P_{\lambda,\mu}$ over $\{0,1\}^n$ be
$$P_{\lambda,\mu}(x) = \frac{1}{D}\sum_{\ell=1}^{D} p_\ell(x_d, \ldots, x_n)\mathbb{1}[\iota(x_1, \ldots, x_{d-1}) = \ell]$$
where $\iota\colon \{0,1\}^{d-1} \to [D]$ is the indexing function, mapping the binary representation to the corresponding number.

That is, $\mathcal{D}_{n,d,\delta}$ is the uniform distribution over the set of degree-d Bayes nets where the first $d-1$ coordinates form a "pointer" to one of the 2^{d-1} tree Bayes nets sharing the same tree structure (the matching λ), but with independently chosen covariance parameters (the D parameters $\mu_1, \ldots, \mu_D$).

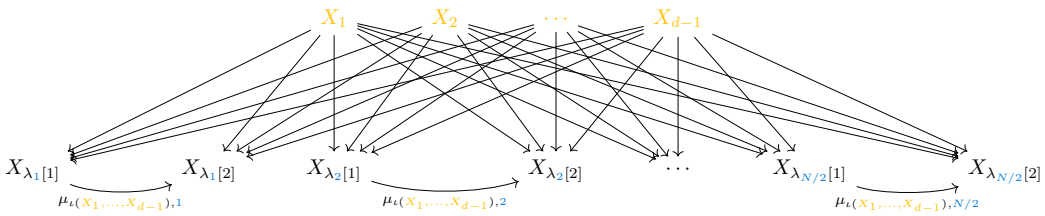

Figure 1: A depiction of our lower bound construction, where the arrows are the edges of the underlying Bayes net. The first $(d-1)$ nodes can be seen as a $(d-1)$-bit string that acts as a pointer to one of the distributions, i.e., the value of the first d nodes encodes which distribution the rest of the nodes are on. We write $\lambda_k[1] := i$ and $\lambda_k[2] := j$ when $\lambda_k = (i,j)$. There are two sources of randomness in the construction. The first one is the random matching on the $N/2$ pairs represented in graph in the form of $(\lambda_1[1], \lambda_1[2]), \ldots, (\lambda_{N/2}[1], \lambda_{N/2}[2])$, which is independent of the particular configuration $X_1, \ldots, X_d$ takes. The second one is the covariance parameters between each pair; these parameters are randomly and independently chosen for each configuration (in total 2^d of them) of the first d nodes. Color (yellow and blue) marks how the value of covariance parameter $\mu_{\ell,k}$ in each pair depends on the first d nodes' configurations as well as the matching parameter k.

With this construction in hand, Theorem 2.1 will follow from the next two lemmas:

Lemma 2.3 (Indistinguishability)**.** *There exist absolute constants $c, C > 0$ such that the following holds. For $\Omega(1) \leqslant d \leqslant c \log n$, no m-sample algorithm can distinguish with probability at least $2/3$ between a (randomly chosen) mixture of trees $P \sim \mathcal{D}_{n,d,\varepsilon/\sqrt{n}}$ and the uniform distribution U over $\{0,1\}^n$, unless $m \geqslant C \cdot 2^{d/2} n/\varepsilon^2$.*

Lemma 2.4 (Distance from product)**.** *Fix any $0 \leqslant d \leqslant \frac{n}{2}$ and $\varepsilon \in (0,1]$. With probability at least $9/10$ over the choice of $P \sim \mathcal{D}_{n,d,\varepsilon/\sqrt{n}}$, P is $\Omega(\varepsilon)$-far from every product distribution on $\{0,1\}^n$.*

Note that this guarantees farness from *every* product distribution, not just from the uniform distribution. We focus on the proof sketch of Lemma 2.3 in the next subsection, and defer the proof of Lemma 2.4 to the appendix A.2. Throughout, we fix n, d, and $\varepsilon \in (0,1]$.

2.1 Sample Complexity to distinguish from Uniform (Lemma 2.3)

Proof sketch. Following the original analogous analysis, we first use Ingster's method to upper bound square TV distance in (4) and after a series of algebraic calculations, we arrive at (6). This is where we improve upon the original analysis by substituting a tighter upper bound (7). This leads us to some of the most technical portions of the paper. To upper bound the inner expectation of (8), we use Lemma 2.5 to bound the unwieldy expression (average over $z_j^{\kappa_{j,i}}$ and raised by m) by an expectation of a multinomial variable $\vec{\alpha}$ in the expression $\prod_{i=1}^{D}\prod_{j=1}^{K}\cosh(2\alpha_i\delta_j)$ in (9); then, with some careful analysis and some tools on MGF (Moment Generating Function) of Binomial and

Multinomial from Section B, we can upper bound the latter expression in (12), which finally gives us the sample complexity lower bound.

Details. We now proceed with the proof of Lemma 2.3, starting with some convenient notations; some of the technical lemmas and facts used here are stated and proven in Appendix B. Let $\theta = (\lambda, \mu_1, \dots, \mu_{2^{d-1}})$, where each $\mu_k \in \{\pm 1\}^{N/2}$, and let P_θ be the distribution for the mixture of trees construction from Definition 2.2. The following denotes the *matching count* between (λ, μ) and x as the quantity

$$c(\lambda, \mu, x) \coloneqq \left|\left\{(i,j) \in \{d, \dots, n\}^2 : \exists 1 \leqslant k \leqslant \tfrac{N}{2}, \lambda_k = (i,j) \text{ and } (-1)^{x_i + x_j} = (-1)^{\mu_{\iota(x),k}}\right\}\right|.$$

We will also introduce an analogous quantity with an "offset", for $x_{\mathrm{ch}} = (x_d, \dots, x_n)$, referring exclusively to the child nodes of x (i.e., the last N nodes, which are the "children" of the first $d-1$ "pointer nodes" in our construction),

$$c_{\mathrm{ch}}(\lambda, \mu, x_{\mathrm{ch}}) \coloneqq \tag{3}$$
$$\left|\left\{(i,j) \in \{d, \dots, n\}^2 : \exists 1 \leqslant k \leqslant \tfrac{N}{2}, \lambda_k = (i-d+1, j-d+1) \text{ and } (-1)^{x_i + x_j} = (-1)^{\mu_{\iota(x),k}}\right\}\right|.$$

To denote the parameters of the "mixture of trees", we write $\theta_i \coloneqq (\lambda, \mu_i)$ (for $i \in [D]$), recalling that the matching parameter λ is common to all D tree components. Since each μ_i corresponds to one of the values of $(x_1, \dots, x_{d-1}) \in \{0,1\}^{d-1}$, we as before use $\iota \colon \{0,1\}^{d-1} \to [D]$ to denote the indexing function (so that, for instance, $\iota(x_1 = \dots = x_{d-1} = 0) = 0$). We finally introduce three more quantities, related to the matching and orientations parameters across the D components of the mixture:

$$A_{\theta_i, \theta'_i} \coloneqq \{(s,t) \in \{1, \dots, N/2\}^2 : \lambda[s] = \lambda'[t], \mu_i[s] = \mu'_i[t]\}$$
(common pairs, same orientation)

$$B_{\theta_i, \theta'_i} \coloneqq \{(s,t) \in \{1, \dots, N/2\}^2 : \lambda[s] = \lambda'[t], \mu_i[s] \neq \mu'_i[t]\}$$
(common pairs, different orientation)

$$C_{\theta_i, \theta'_i} \coloneqq (\lambda \cup \lambda') \setminus (A \cup B) \qquad \text{(pairs unique to } \theta \text{ or } \theta'\text{)}$$

For ease of notation, we define $A_i \coloneqq A_{\theta_i, \theta'_i}$, $B_i \coloneqq B_{\theta_i, \theta'_i}$ and $C_i \coloneqq C_{\theta_i, \theta'_i}$; and note that $C_i = C_1$, as it only depends on λ (not on the orientation μ_i).

To prove the indistinguishability, we will bound the squared total variation distance (or equivalently, squared ℓ_1) distance between the distributions of m samples from (the uniform mixture of) P_θ and U by a small constant; that is, between $Q \coloneqq \mathbb{E}_\theta[P_\theta^{\otimes m}]$ and $U^{\otimes m}$. From Ingster's method (see, e.g., [Acharya et al., 2020, Lemma III.8.]), by using chi-square divergence as an intermediate step we get

$$\left\|Q - U^{\otimes m}\right\|_1^2 \leqslant d_{\chi^2}(Q, U^{\otimes m}) = \mathbb{E}_{\theta, \theta'}[(1 + \tau(\theta, \theta'))^m] - 1, \tag{4}$$

where $\tau(\theta, \theta') \coloneqq \mathbb{E}_{x \sim U}\left[\left(\frac{P_\theta(x) - U(x)}{U(x)}\right)\left(\frac{P_{\theta'}(x) - U(x)}{U(x)}\right)\right]$. In order to get a handle on this quantity $\tau(\theta, \theta')$, we start by writing the expression for the density P_θ. For any $x \in \{0,1\}^n$, recalling Item 4 of Definition 2.2, $P_\theta(x) = \frac{1}{2^n}(1+4\delta)^{c(\lambda, \mu_{\iota(x)}, x)}(1-4\delta)^{\frac{N}{2} - c(\lambda, \mu_{\iota(x)}, x)}$.

Substituting this in the definition of τ, we get an expression depending on $z_0 \coloneqq \frac{1+4\delta}{1-4\delta}$. and the probability-generating function (PGF) of $c_{\mathrm{ch}}(\lambda, \mu, x)$. Observing that, as $x \sim U_N$ and for fixed λ, μ we have that $c_{\mathrm{ch}}(\lambda, \mu, x) \sim \mathrm{Bin}\left(\frac{N}{2}, \frac{1}{2}\right)$, we can then massage this expression to get

$$1 + \tau(\theta, \theta') = \frac{1}{D}\left\{\sum_{i=1}^{D}(1-4\delta)^N \cdot z_0^{|B_i|}\left(\frac{1+z_0^2}{2}\right)^{|A_i|} \prod_{\sigma_i : |\sigma_i| \geqslant 4} \mathbb{E}_{\alpha \sim \mathcal{B}(\sigma_i)}[z_0^\alpha]\right\}, \tag{5}$$

where the product is taken over all cycles in the multigraph $G_{\theta, \theta'}$ induced by the two matchings; and, given a cycle σ, $\mathcal{B}(\sigma)$ is the probability distribution defined as follows. Say that a cycle σ is *even* (resp., *odd*) if the number of edges with weight 1 along σ is even (resp., odd); that is, a cycle is even or odd depending on whether the number of negatively correlated pairs along the cycle is an even or odd number. If σ is an *even cycle*, then $\mathcal{B}(\sigma)$ is a Binomial with parameters $|\sigma|$ and $1/2$, conditioned on taking even values. Similarly, if σ is an *odd cycle*, $\mathcal{B}(\sigma)$ is a Binomial with parameters $|\sigma|$ and $1/2$, conditioned on taking odd values. It follows that $\mathbb{E}_{\alpha \sim \mathcal{B}(\sigma)}[z_0^\alpha]$ is given by the following expression.

$$\mathbb{E}_{\alpha \sim \mathcal{B}(\sigma)}[z_0^\alpha] = \begin{cases} \mathbb{E}_{\alpha \sim \mathrm{Bin}\left(|\sigma|, \frac{1}{2}\right)}[z^\alpha \mid \alpha \text{ even}] = \frac{(1+z_0)^{|\sigma|} + (1-z_0)^{|\sigma|}}{2^{|\sigma|}} & \text{, if } \sigma \text{ is even} \\ \mathbb{E}_{\alpha \sim \mathrm{Bin}\left(|\sigma|, \frac{1}{2}\right)}[z^\alpha \mid \alpha \text{ odd}] = \frac{(1+z_0)^{|\sigma|} - (1-z_0)^{|\sigma|}}{2^{|\sigma|}} & \text{, if } \sigma \text{ is odd} \end{cases}.$$

Denote $\mathcal{S}_{e,i} \coloneqq \mathcal{S}_e(\theta_i, \theta_i') = \{\sigma \in \mathrm{cycle}(\theta_i, \theta_i') : |\sigma| \geqslant 4, \sigma \text{ is even}\}$, and similarly $\mathcal{S}_{o,i}$ for odd cycles. We would drop λ, μ or i, when clear from context. We first expand $\mathbb{E}_{\alpha\sim\mathcal{B}(\sigma)}[z_0^\alpha]$ as follows:

$$\prod_{\sigma:|\sigma|\geqslant 4} \mathbb{E}_{\alpha\sim\mathcal{B}(\sigma)}[z_0^\alpha] \quad = \quad \frac{(1+z_0)^{|C|}}{2^{|C|}} \prod_{\sigma\in\mathcal{S}_e} (1+(-4\delta)^{|\sigma|}) \prod_{\sigma\in\mathcal{S}_o} (1-(-4\delta)^{|\sigma|}). \tag{6}$$

We now improve upon the analogous analysis from [Canonne et al., 2020, Claim 12] to obtain a better upper bound for the remaining terms; indeed, the bound they derived is $e^{O(\varepsilon^4/n)}$, which was enough for their purposes but not ours (since it does not feature any dependence on d). Crucially, we instead derive the below inequality whose proof can be found in Appendix C.1:

$$\prod_{\sigma\in\mathcal{S}_e} (1+(-4\delta)^{|\sigma|}) \prod_{\sigma\in\mathcal{S}_o} (1-(-4\delta)^{|\sigma|}) \leqslant e^{c'\frac{\varepsilon^5}{N^{3/2}}} \prod_{\sigma\in\mathcal{S}_e:|\sigma|=4} (1+(-4\delta)^{|\sigma|}) \prod_{\sigma\in\mathcal{S}_o:|\sigma|=4} (1-(-4\delta)^{|\sigma|}), \tag{7}$$

for some absolute constant $c' > 0$. We will have $e^{c'\cdot m\varepsilon^5/N^{3/2}} \leqslant e^{128m\varepsilon^2/(\sqrt{D}n)}$ if $D = O(n/\varepsilon^6)$, and this restriction on D is satisfied for the regime of parameters considered in our lower bound, $d = O(\log n)$. Letting $R \coloneqq |A_1| + |B_1| \leqslant N/2$ and $z_1 \coloneqq \frac{1+(4\delta)^2}{1-(4\delta)^2}$, with (6) and some calculations,

$$\begin{aligned}
&\mathbb{E}_{\theta,\theta'}[(1+\tau(\theta,\theta'))^m] \\
&= \mathbb{E}_{\lambda,\lambda'}\left[(1-(4\delta)^2)^{mR}\mathbb{E}_{\vec{\mu},\vec{\mu}'}\left[\left(\frac{1}{D}\sum_{i=1}^{D} z_1^{|A_i|} \prod_{\sigma\in\mathcal{S}_{e,i}} (1+(-4\delta)^{|\sigma|}) \prod_{\sigma\in\mathcal{S}_{o,i}} (1-(-4\delta)^{|\sigma|})\right)^m\right]\right].
\end{aligned} \tag{8}$$

Fix a pair λ, λ'; we have that the $|A_i|$'s are i.i.d. $\mathrm{Bin}(R, 1/2)$ random variables. We now introduce $R' \coloneqq |\{\sigma_1 : |\sigma_1| = 4\}| = |\{\sigma_i : |\sigma_i| = 4\}|$, which is the random variable denoting how many cycles have length exactly 4. In particular, we have $R' \leqslant \frac{N}{4}$, since $\sum_{\sigma:|\sigma|\geqslant 4} \sigma = |C| \leqslant N$; more specifically, we have $R' \leqslant \frac{N-2R}{4} \leqslant \frac{N}{4}$. Further, define κ_i as the number of cycles of length 4 which have an even total number of negative correlations; that is, the number of cycles σ such that μ_i, μ_i' impose either 0, 2, or 4 negatively correlated pairs along that cycle.

Since μ, μ' are uniformly distributed, being odd or even each has probability $1/2$, and thus $\kappa_i \sim \mathrm{Bin}\big(R', \frac{1}{2}\big)$. Moreover, while κ_i and A_i both depend on μ_i, μ_i', they by definition depend on disjoint subsets of those two random variables: thus, because each correlation parameter is chosen independently, we have that κ_i and A_i are independent conditioned on (R, R'). Now, recalling our setting of $z_2 = \frac{1+(4\delta)^4}{1-(4\delta)^4}$ and fixing a realization of R, R', we have

$$\begin{aligned}
&\mathbb{E}_{\vec{\mu},\vec{\mu}'}\left[\left(\frac{1}{D}\sum_{i=1}^{D} z_1^{|A_i|} \prod_{\sigma\in\mathcal{S}_e(i):|\sigma|=4} (1+(4\delta)^4) \prod_{\sigma_i\in\mathcal{S}_o(i):|\sigma|=4} (1-(4\delta)^4)\right)^m\right] \\
&\quad = (1-(4\delta)^4)^{mR'}\mathbb{E}_{\vec{\mu},\vec{\mu}'}\left[\left(\frac{1}{D}\sum_{i=1}^{D} z_1^{|A_i|} z_2^{\kappa_i}\right)^m\right] \\
&\quad \leqslant (1-(4\delta)^4)^{mR'} z_1^{\frac{mR}{2}} z_2^{\frac{mR'}{2}} \mathbb{E}_{\vec{\alpha}}\left[\prod_{i=1}^{D} (\cosh(2\alpha_i\delta^2))^R (\cosh(2\alpha_i\delta^4))^{R'}\right],
\end{aligned} \tag{9}$$

where (9) follows from the following lemma:

Lemma 2.5. *There exists an absolute constant $\delta_0 \approx 0.96$ such that the following holds. Let $K \geqslant 1$ and $R_1, \ldots, R_K$ be integers, and $\delta_1, \ldots, \delta_K \in (0, \delta_0]$. Suppose that $\kappa_{j,1}, \ldots, \kappa_{j,D} \sim \mathrm{Bin}\big(R_j, \frac{1}{2}\big)$, are i.i.d., and mutually independent across $1 \leqslant j \leqslant K$, and $z_j \coloneqq \frac{1+\delta_j}{1-\delta_j}$. Then*

$$\mathbb{E}\left[\left(\frac{1}{D}\sum_{i=1}^{D}\prod_{j=1}^{K} z_j^{\kappa_{j,i}}\right)^m\right] \leqslant \left(\prod_{j=1}^{K} z_j^{\frac{m}{2}R_j}\right)\mathbb{E}_{\vec{\alpha}}\left[\prod_{i=1}^{D}\prod_{j=1}^{K}\cosh(2\alpha_i\delta_j)\right],$$

where $(\alpha_1, \ldots, \alpha_D)$ follows a multinomial distribution with parameters m and $(1/D, \ldots, 1/D)$.

We now focus on the expectation on the right (last factor of the RHS of (9)): using that $\cosh u \leqslant \min(e^{u^2/2}, e^u)$ for $u \geqslant 0$, it can be shown, setting $\Delta := 1/\delta^2 = n/\varepsilon^2$, that

$$\begin{aligned}
&\mathbb{E}_{\vec{\alpha}}\left[\prod_{i=1}^{D}(\cosh(2\alpha_i\delta^2))^R(\cosh(2\alpha_i\delta^4))^{R'}\right] \\
&\leqslant \mathbb{E}\left[\prod_{i=1}^{D} e^{8\alpha_i\delta^2 R\mathbb{1}[\alpha_i>\Delta]}\right]^{\frac{1}{4}} \mathbb{E}\left[\prod_{i=1}^{D} e^{8\alpha_i^2\delta^4 R\mathbb{1}[\alpha_i\leqslant\Delta]}\right]^{\frac{1}{4}} \mathbb{E}\left[\prod_{i=1}^{D} e^{4\alpha_i^2\delta^8 R'}\right]^{\frac{1}{2}} \qquad (10)
\end{aligned}$$

where the threshold Δ was chosen as the value for which the term realizing the minimum changes. We first bound the product of the last two expectations:

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^{D} e^{8\alpha_i^2\delta^4 R\mathbb{1}[\alpha_i\leqslant\Delta]}\right]^{1/4} \mathbb{E}\left[\prod_{i=1}^{D} e^{4\alpha_i^2\delta^8 R'}\right]^{1/2} &\leqslant \mathbb{E}\left[\prod_{i=1}^{D} e^{8\delta^4 R\min(\alpha_i^2,\Delta^2)}\right]^{1/4} \mathbb{E}\left[\prod_{i=1}^{D} e^{4\alpha_i^2\delta^8 R'}\right]^{1/2} \\
&\leqslant (\mathbb{E}[e^{8\min(\alpha_j^2,\Delta^2)\delta^4 R}])^{D/4}(\mathbb{E}[e^{4\alpha_1^2\delta^8 R'}])^{D/2} \qquad (11) \\
&\leqslant \exp\Big(32\delta^4\frac{m^2}{D}R\Big)\exp\Big(32\delta^8\frac{m^2}{D}R'\Big). \qquad (12)
\end{aligned}$$

where we applied negative association (see, e.g., Dubhashi and Ranjan [1996, Theorem 13]) on both expectations for (11); and then got (12) by Lemmas B.7 and B.3 (for the latter, noting that $tm = 2\delta^8 mR' \leqslant 1/16$; and, for the former, assuming with little loss of generality that $\varepsilon \leqslant 1/(4\sqrt{2})$). Applying Lemma B.6 to the first (remaining) factor of the LHS above as $8\delta^2 R \leqslant 4$ and $D \geqslant \Omega(1)$, we get

$$\mathbb{E}\left[\prod_{i=1}^{D} e^{8\alpha_i\delta^2 R\mathbb{1}[\alpha_i>\Delta]}\right]^{1/4} \mathbb{E}\left[\prod_{i=1}^{D} e^{8\alpha_i^2\delta^4 R\mathbb{1}[\alpha_i\leqslant\Delta]}\right]^{1/4} \mathbb{E}\left[\prod_{i=1}^{D} e^{4\alpha_i^2\delta^8 R'}\right]^{1/2} \leqslant (1+o(1))\exp(32C'^2 R),$$

recalling that $R' \leqslant N/4 \leqslant n/4$, and our assumption that $m \leqslant C'\sqrt{D}n/\varepsilon^2$. Combining (7), (8) and (9), what we showed is

$$\begin{aligned}
\mathbb{E}_{\theta,\theta'}[(1+\tau(\theta,\theta'))^m] &\leqslant (1+o(1))e^{128\frac{m\varepsilon^2}{\sqrt{D}n}}\mathbb{E}_{\lambda,\lambda'}\left[(1-(4\delta)^2)^{mR}(1-(4\delta)^4)^{mR'}z_1^{\frac{mR}{2}}z_2^{\frac{mR'}{2}}e^{32C'^2\cdot R}\right] \\
&\leqslant (1+o(1))e^{128\cdot C'}\mathbb{E}_{\lambda,\lambda'}\left[e^{32C'^2\cdot R}\right],
\end{aligned}$$

where the equality follows from the definition of z_1, z_2. To conclude, we will use the fact that, for every $k \geqslant 0$, $\Pr[R > k] \leqslant \frac{1}{k!}$, which was established in Canonne et al. [2020, p.46]. By summation by parts, one can show that this implies $\mathbb{E}_R\left[e^{\alpha R}\right] \leqslant 1+(1-e^{-\alpha})(e^{e^\alpha}-1) \xrightarrow{\alpha\to 0} 1+\alpha(e-1)$ for any $\alpha > 0$, and so, in our case, $\mathbb{E}_{\theta,\theta'}[(1+\tau(\theta,\theta'))^m] \leqslant (1+o(1))e^{128\cdot C'}(1+(1-e^{-32C'^2})(e^{e^{32C'^2}}-1))$ In particular, the RHS can be made arbitrarily close to 1 by choosing a small enough value for the constant C' (in the bound for m). By (4), this implies the desired bound on $\|Q - U^{\otimes m}\|_1^2$, and thus establishes Lemma 2.3. □

3 Upper Bound for Independence Testing

In this section, we establish the upper bound part of Theorem 1.1, stated below:

Theorem 3.1. *There is an algorithm (Algorithm 1) with the following guarantees: given $\varepsilon \in (0,1]$ and sample access to an unknown degree-d Bayes net P, the algorithm takes $O(d^2 2^{d/2} n\log(n)/\varepsilon^2)$ samples from P, and distinguishes with probability at least $2/3$ between the cases (i) P is a product distribution, and (ii) P is at total variation distance at least ε from every product distribution.*

Algorithm 1: Independence testing for degree-d Bayes net

Input : Independent samples from a degree-d Bayes net P over $\{0,1\}^n$, $\varepsilon \in (0,1]$

Set $\delta \leftarrow \frac{1}{3\binom{n}{d+1}}$, $\varepsilon' \leftarrow \frac{\varepsilon}{\sqrt{2n}(1+\sqrt{d+1})}$, and $m \leftarrow O\Big(\frac{2^{d/2}}{\varepsilon'^2}\log\frac{1}{\delta}\Big)$

Take a multiset S of m samples from P, where $\delta \leftarrow 1/(3\binom{n}{d+1})$

for *every subset* $T \subseteq [n]$ *of* $d+1$ *nodes* **do**

  `/* We can generate one sample from` P_T `given one from` P `*/`

  Use the m samples from S to generate m i.i.d. samples from P_T

  Call Algorithm 2 with distance parameter ε' and failure probability δ, using the m samples from P_T `/* This is over` $\{0,1\}^{d+1}$`. */`

if *all* $\binom{n}{d+1}$ *tests accepted* **then** **return** *accept*

return *reject*

As discussed in the introduction, the key components of our algorithm are performing the testing in Hellinger distance, in order to use the subadditivity result of Daskalakis and Pan [2017]; and using as subroutine an independence testing algorithm under Hellinger distance. As no optimal tester for this latter task was known prior to our work, our first step is to derive such an algorithm by adapting the "testing-by-learning" framework of Acharya et al. [2015]. We emphasize that, in view of the relation $\frac{1}{2}d_{\mathrm{TV}}^2 \leqslant d_{\mathrm{H}}^2 \leqslant d_{\mathrm{TV}}$, this is a strictly stronger statement than the analogous known $O(2^{n/2}/\varepsilon^2)$-sample algorithm for testing independence under total variation distance [Diakonikolas and Kane, 2016, Acharya et al., 2015], which only implies an $O(2^{n/2}/\varepsilon^4)$ sample complexity for testing independence under Hellinger distance.

Lemma 3.2. *There exists an algorithm (Algorithm 2) with the following guarantees: given a parameter $\varepsilon \in (0,1]$ and sample access to an unknown distribution P over $\{0,1\}^n$, it takes $O(2^{n/2}/\varepsilon^2)$ samples from P, and distinguishes with probability at least $2/3$ between the cases (i) P is a product distribution, and (ii) P is at* Hellinger *distance at least ε from every product distribution.*

Algorithm 2: Hellinger independence testing for general distributions

Input : Independent samples from the target distribution over $\{0,1\}^n$, $\varepsilon \in (0,1]$

Set $m_1 \leftarrow O\big(n/\varepsilon^2\big)$, and $m_2 \leftarrow O\big(2^{n/2}/\varepsilon^2\big)$

Take two multisets S_1 and S_2 of samples with size m_1, m_2 respectively from P

$\hat{P} \leftarrow$ call the algorithm of Lemma 1.3 with the samples from S_1

Call the tester in Theorem 1.4 with samples from S_2, to test P with reference to $\hat{P}$ with distance ε

if *tester accepted* **then** **return** *accept*

return *reject*

We will next require the following lemma, whose proof is deferred to Appendix C.2:

Lemma 3.3. *Let P be a distribution on $\{0,1\}^n$, and $P' = (\pi_1 P) \otimes \cdots \otimes (\pi_n P)$ be the product of marginals of P. Denoting by $\mathcal{Q}$ the set of all product distributions on $\{0,1\}^n$, we have $\min_{Q\in\mathcal{Q}} d_{\mathrm{H}}(P,Q) \geqslant \frac{1}{1+\sqrt{n}} d_{\mathrm{H}}(P,P')$.*

We now have all the elements to prove Theorem 3.1. As discussed in the introduction, the idea is the following: when P is a product distribution, then the distribution induced by P' on any subset of $d+1$ nodes is still a product distribution. However, by the above localization Corolloary 1.2, when P is far from every product distribution, we can localize the farness between P and Q in one of the subsets of $d+1$ nodes. We also know that P', the product of marginals P, is not too bad of an approximation to the closest Q in the product space as suggested by Lemma 3.3. Thus, testing independence using P' as a proxy and paying an extra factor of $O(d)$ to compensate in accuracy suffice. Details can be found in Appendix C.1.

Acknowledgments. Yang would like to thank Vipul Arora and Philips George John for the helpful discussions on the lower bound analysis; and Vipul, specifically, for providing valuable feedback on the manuscript.

References

Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *NeurIPS*, pages 3591–3599, 2015. Preprint available at arXiv:1507.05952.

Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Inference under information constraints I: Lower bounds from chi-square contraction. *IEEE Trans. Inform. Theory*, 66(12):7835–7855, 2020. ISSN 0018-9448. doi: 10.1109/TIT.2020.3028440. URL https://doi.org/10.1109/TIT.2020.3028440.

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

Noga Alon, Alexandr Andoni, Tali Kaufman, Kevin Matulef, Ronitt Rubinfeld, and Ning Xie. Testing k-wise and almost k-wise independence. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 496–505, 2007.

Tugkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 442–451. IEEE, 2001.

Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, and N. V. Vinodchandran. Efficient distance approximation for structured high-dimensional distributions via learning. In *NeurIPS*, 2020.

Arnab Bhattacharyya, Sutanu Gayen, Saravanan Kandasamy, and NV Vinodchandran. Testing product distributions: A closer look. In *Algorithmic Learning Theory*, pages 367–396. PMLR, 2021.

Eric Blais, Clément L. Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. *ACM Transactions on Computation Theory*, 11(2):1–37, apr 2019. doi: 10.1145/3305270.

Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? *Theory of Computing*, pages 1–100, 2020.

Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing bayesian networks. *IEEE Trans. Inf. Theory*, 66(5):3132–3170, 2020. Preprint available at arXiv:1612.03156.

Constantinos Daskalakis and Qinxuan Pan. Square hellinger subadditivity for bayesian networks and its applications to identity testing. In *COLT*, volume 65 of *Proceedings of Machine Learning Research*, pages 697–703. PMLR, 2017. Preprint available at arXiv:1612.03164.

Constantinos Daskalakis, Ilias Diakonikolas, and Rocco A Servedio. Learning k-modal distributions via testing. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1371–1385. SIAM, 2012.

Constantinos Daskalakis, Gautam Kamath, and John Wright. Which distribution distances are sublinearly testable? In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2747–2764. SIAM, 2018.

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

Ilias Diakonikolas and Daniel M Kane. A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 685–694. IEEE, 2016.

Devdatt Dubhashi and Desh Ranjan. Balls and bins: A study in negative dependence. *RANDOM STRUCTURES & ALGORITHMS*, 13:99–124, 1996.

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620, 2000. doi: 10.1089/106652700750050961. PMID: 11108481.

Timothy S Gardner, Diego Di Bernardo, David Lorenz, and James J Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301(5629): 102–105, 2003.

Oded Goldreich, Shari Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM (JACM)*, 45(4):653–750, 1998.

Yi Hao and Ping Li. Bessel smoothing and multi-distribution property estimation. In *COLT*, volume 125 of *Proceedings of Machine Learning Research*, pages 1817–1876. PMLR, 2020.

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan R. Ullman. Privately learning high-dimensional distributions. In *COLT*, volume 99 of *Proceedings of Machine Learning Research*, pages 1853–1902. PMLR, 2019.

Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*, volume 3. Springer, 2005.

Reut Levi, Dana Ron, and Ronitt Rubinfeld. Testing properties of collections of distributions. *Theory of Computing*, 9(1):295–347, 2013.

Matey Neykov, Junwei Lu, and Han Liu. Combinatorial inference for graphical models. *The Annals of Statistics*, 47(2):795–827, 2019.

Jerzy Neyman and Egon Sharpe Pearson. IX. 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.

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

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

Ronitt Rubinfeld. Taming big probability distributions. *XRDS: Crossroads, The ACM Magazine for Students*, 19(1):24–28, 2012.

Ronitt Rubinfeld and Rocco A Servedio. Testing monotone high-dimensional distributions. *Random Structures & Algorithms*, 34(1):24–44, 2009.

Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, 1996.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1–2):1–305, jan 2008. ISSN 1935-8237. doi: 10.1561/2200000001. URL `https://doi.org/10.1561/2200000001`.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] The lower bound is limited in the sparse Bayes net setting as the assumption indicates, and thus the $\tilde{\Theta}(2^{d/2} \cdot n/\varepsilon^2)$ sample complexity derived is only tight in the sparse regime. More generally, as answered below all assumptions for our results are explicitly specified.
   (c) Did you discuss any potential negative societal impacts of your work? [No] In view of the theoretical nature of our paper and results, it is difficult to gauge potential negative societal impacts, as potential applications are quite far down the line.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Detailed version of Section 2

In this section, we prove our main technical result: a lower bound on the sample complexity of testing whether a degree-d Bayes net is a product distribution.

Theorem 2.1. *Let $d \leqslant c \cdot \log n$, where $c > 0$ is a sufficiently small absolute constant. Then, testing whether an arbitrary degree-d Bayes net over $\{0,1\}^n$ is a product distribution or is ε-far from every product distribution requires $\Omega(2^{d/2} n/\varepsilon^2)$ samples.*

This theorem considerably generalizes the lower bound of $\Omega(n/\varepsilon^2)$ established in Canonne et al. [2020] for the case of trees ($d = 1$). To establish our result, we build upon (and considerably extend) their analysis; in particular, we will rely on the following "mixture of trees" construction, which can be seen as a careful mixture of (2^{d-1} of) the hard instances from the lower bound of [Canonne et al., 2020, Theorem 14].

Notation. Throughout, for given n, d, ε, we let $N \coloneqq n - d + 1$ (without loss of generality assumed to be even), $D \coloneqq 2^{d-1}$, $\delta \coloneqq \frac{\varepsilon}{\sqrt{n}}$, and

$$z_0 = \frac{1+4\delta}{1-4\delta}, \quad z_1 = \frac{1+(4\delta)^2}{1-(4\delta)^2}, \quad z_2 = \frac{1+(4\delta)^4}{1-(4\delta)^4}.$$

Definition 2.2 (Mixture of Trees)**.** Given parameters $0 \leqslant d \leqslant n$ and $\delta \in (0,1]$, we define the probability distribution $\mathcal{D}_{n,d,\delta}$ over degree-d Bayes nets by the following process.

1. Choose a perfect matching λ of $[N]$ uniformly at random (where $N = n - d + 1$), i.e., a set of $N/2$ disjoint pairs;
2. Draw i.i.d. $\mu_1, \ldots, \mu_D$ uniformly at random in $\{0,1\}^{N/2}$;
3. For $\ell \in [D]$, let p_ℓ be the distribution over $\{0,1\}^N$ defined as the Bayes net with tree structure λ, such that if $\lambda_k = (i,j) \in \lambda$ then the corresponding covariance between variables X_i, X_j is
$$\mathrm{Cov}(X_i, X_j) = (-1)^{\mu_{\ell,k}} \delta$$
4. Let the resulting distribution $P_{\lambda,\mu}$ over $\{0,1\}^n$ be
$$P_{\lambda,\mu}(x) = \frac{1}{D} \sum_{\ell=1}^{D} p_\ell(x_d, \ldots, x_n) \mathbb{1}[\iota(x_1, \ldots, x_{d-1}) = \ell]$$
where $\iota \colon \{0,1\}^{d-1} \to [D]$ is the indexing function, mapping the binary representation to the corresponding number.

That is, $\mathcal{D}_{n,d,\delta}$ is the uniform distribution over the set of degree-d Bayes nets where the first $d-1$ coordinates form a "pointer" to one of the 2^{d-1} tree Bayes nets sharing the same tree structure (the matching λ), but with independently chosen covariance parameters (the D parameters $\mu_1, \ldots, \mu_D$).

With this construction in hand, Theorem 2.1 will follow from the next two lemmas:

Lemma 2.3 (Indistinguishability)**.** *There exist absolute constants $c, C > 0$ such that the following holds. For $\Omega(1) \leqslant d \leqslant c \log n$, no m-sample algorithm can distinguish with probability at least $2/3$ between a (randomly chosen) mixture of trees $P \sim \mathcal{D}_{n,d,\varepsilon/\sqrt{n}}$ and the uniform distribution U over $\{0,1\}^n$, unless $m \geqslant C \cdot 2^{d/2} n/\varepsilon^2$.*

Lemma 2.4 (Distance from product)**.** *Fix any $0 \leqslant d \leqslant \frac{n}{2}$ and $\varepsilon \in (0,1]$. With probability at least $9/10$ over the choice of $P \sim \mathcal{D}_{n,d,\varepsilon/\sqrt{n}}$, P is $\Omega(\varepsilon)$-far from every product distribution on $\{0,1\}^n$.*

Note that this guarantees farness from *every* product distribution, not just from the uniform distribution. We first establish Lemma 2.3 in the next subsection, before proving Lemma 2.4 in Section A.2. Throughout, we fix n, d, and $\varepsilon \in (0,1]$.

A.1 Sample Complexity to distinguish from Uniform (Lemma 2.3)

We here proceed with the proof of Lemma 2.3, starting with some convenient notations; some of the technical lemmas and facts used here are stated and proven in Section B. Let $\theta = (\lambda, \mu_1, \ldots, \mu_{2^{d-1}})$,

where each $\mu_k \in \{\pm 1\}^{N/2}$, and let P_θ be the distribution for the mixture of trees construction from Definition 2.2. The following denotes the *matching count* between (λ, μ) and x as the quantity

$$c(\lambda, \mu, x) \coloneqq \left|\left\{(i,j) \in \{d, \dots, n\}^2 : \exists 1 \leqslant k \leqslant \tfrac{N}{2}, \lambda_k = (i,j) \text{ and } (-1)^{x_i + x_j} = (-1)^{\mu_{\iota(x),k}}\right\}\right|.$$

We will also introduce an analogous quantity with an "offset", for $x_{\mathrm{ch}} = (x_d, \dots, x_n)$, referring exclusively to the child nodes of x (i.e., the last N nodes, which are the "children" of the first $d-1$ "pointer nodes" in our construction),

$$c_{\mathrm{ch}}(\lambda, \mu, x_{\mathrm{ch}}) \coloneqq \tag{13}$$
$$\left|\left\{(i,j) \in \{d, \dots, n\}^2 : \exists 1 \leqslant k \leqslant \tfrac{N}{2}, \lambda_k = (i-d+1, j-d+1) \text{ and } (-1)^{x_i + x_j} = (-1)^{\mu_{\iota(x),k}}\right\}\right|.$$

To denote the parameters of the "mixture of trees," we write $\theta_i \coloneqq (\lambda, \mu_i)$ (for $i \in [D]$), recalling that the matching parameter λ is common to all D tree components. Since each μ_i corresponds to one of the values of $(x_1, \dots, x_{d-1}) \in \{0,1\}^{d-1}$, we as before use $\iota \colon \{0,1\}^{d-1} \to [D]$ to denote the indexing function (so that, for instance, $\iota(x_1 = \dots = x_{d-1} = 0) = 0$). We finally introduce three more quantities, related to the matching and orientations parameters across the D components of the mixture:

$$A_{\theta_i, \theta_i'} \coloneqq \{(s,t) \in \{1, \dots, N/2\}^2 : \lambda[s] = \lambda'[t], \mu_i[s] = \mu_i'[t]\} \qquad \text{(common pairs, same orientation)}$$

$$B_{\theta_i, \theta_i'} \coloneqq \{(s,t) \in \{1, \dots, N/2\}^2 : \lambda[s] = \lambda'[t], \mu_i[s] \neq \mu_i'[t]\} \qquad \text{(common pairs, different orientation)}$$

$$C_{\theta_i, \theta_i'} \coloneqq (\lambda \cup \lambda') \setminus (A \cup B) \qquad \text{(pairs unique to } \theta \text{ or } \theta'\text{)}$$

For ease of notation, we define $A_i \coloneqq A_{\theta_i, \theta_i'}$, $B_i \coloneqq B_{\theta_i, \theta_i'}$ and $C_i \coloneqq C_{\theta_i, \theta_i'}$; and note that $C_i = C_1$, as it only depends on λ (not on the orientation μ_i).

To prove the indistinguishability, we will bound the squared total variation distance (or equivalently, squared ℓ_1) distance between the distributions of m samples from (the uniform mixture of) P_θ and U by a small constant; that is, between $Q \coloneqq \mathbb{E}_\theta[P_\theta^{\otimes m}]$ and $U^{\otimes m}$. From Ingster's method (see, e.g., [Acharya et al., 2020, Lemma III.8.]), by using chi-square divergence as an intermediate step we get

$$\left\|Q - U^{\otimes m}\right\|_1^2 \leqslant d_{\chi^2}(Q, U^{\otimes m}) = \mathbb{E}_{\theta, \theta'}[(1 + \tau(\theta, \theta'))^m] - 1, \tag{14}$$

where $\tau(\theta, \theta') \coloneqq \mathbb{E}_{x \sim U}\left[\left(\frac{P_\theta(x) - U(x)}{U(x)}\right)\left(\frac{P_{\theta'}(x) - U(x)}{U(x)}\right)\right]$. In order to get a handle on this quantity $\tau(\theta, \theta')$, we start by writing the expression for the density P_θ (for a given parameter θ of the mixture of trees). For any $x \in \{0,1\}^n$, recalling Item 4 of Definition 2.2,

$$\begin{aligned}
P_\theta(x) &= P_\theta(x_d, \dots, x_n \mid x_1, \dots, x_{d-1}) U_{d-1}(x_1, \dots, x_{d-1}) \\
&= \frac{1}{2^{d-1}} P_{\lambda, \mu_{\iota(x)}}(x_d, \dots, x_n) \\
&= \frac{1}{2^{d-1}} \cdot \frac{1}{2^N} (1 + 4\delta)^{c(\lambda, \mu_{\iota(x)}, x)} (1 - 4\delta)^{\frac{N}{2} - c(\lambda, \mu_{\iota(x)}, x)} \\
&= \frac{1}{2^n} (1 + 4\delta)^{c(\lambda, \mu_{\iota(x)}, x)} (1 - 4\delta)^{\frac{N}{2} - c(\lambda, \mu_{\iota(x)}, x)}.
\end{aligned}$$

Substituting this in the definition of τ, we get

$$\begin{aligned}
\tau(\theta, \theta') &= \mathbb{E}_{x \sim U}\left[\left(\frac{P_\theta(x)}{U(x)} - 1\right)\left(\frac{P_{\theta'}(x)}{U(x)} - 1\right)\right] \\
&= \mathbb{E}_{x \sim U}\left[\left((1 - 4\delta)^{\frac{N}{2}} \left(\frac{1 + 4\delta}{1 - 4\delta}\right)^{c(\lambda, \mu_{\iota(x)}, x)} - 1\right)\left((1 - 4\delta)^{\frac{N}{2}} \left(\frac{1 + 4\delta}{1 - 4\delta}\right)^{c(\lambda', \mu'_{\iota(x)}, x)} - 1\right)\right] \\
&= 1 + (1 - 4\delta)^N \mathbb{E}_{x \sim U}\left[z_0^{c(\lambda, \mu_{\iota(x)}, x) + c(\lambda', \mu'_{\iota(x)}, x)}\right] \\
&\quad - (1 - 4\delta)^{\frac{N}{2}} \mathbb{E}_{x \sim U}\left[z_0^{c(\lambda, \mu_{\iota(x)}, x)}\right] - (1 - 4\delta)^{\frac{N}{2}} \mathbb{E}_{x \sim U}\left[z_0^{c(\lambda', \mu'_{\iota(x)}, x)}\right],
\end{aligned} \tag{15}$$

where $z_0 \coloneqq \frac{1+4\delta}{1-4\delta}$. As $x \sim U_N$, for fixed λ, μ we have that $c_{\mathrm{ch}}(\lambda, \mu, x)$ follows a $\mathrm{Bin}\big(\frac{N}{2}, \frac{1}{2}\big)$ distribution; recalling the expression of the Binomial distribution's probability generating function, we then have

$$\begin{aligned}
(1-4\delta)^{\frac{N}{2}}\,\mathbb{E}_{x\sim U}[z_0^{c(\lambda,\mu_{\iota(x)},x)}] &= (1-4\delta)^{\frac{N}{2}}\,\mathbb{E}_{\vec{x}_1\sim U_{d-1}}\left[\mathbb{E}_{\vec{x}_2\sim U_N}\left[z_0^{c_{\mathrm{ch}}(\lambda,\mu_{\iota(\vec{x}_1)},\vec{x}_2)}\right]\right] \\
&= (1-4\delta)^{\frac{N}{2}}\,\frac{1}{D}\left\{\sum_{i=1}^{D}\mathbb{E}_{m\sim\mathrm{Bin}\left(\frac{N}{2},\frac{1}{2}\right)}[z_0^m]\right\} \\
&= (1-4\delta)^{\frac{N}{2}}\left(\frac{1+z_0}{2}\right)^{\frac{N}{2}} = 1.
\end{aligned}$$

Using this to simplify the last two terms of (15), we obtain

$$\begin{aligned}
1+\tau(\theta,\theta') &= (1-4\delta)^N\mathbb{E}_{x\sim U}[z_0^{c(\lambda,\mu_{\iota(x)},x)+c(\lambda',\mu'_{\iota(x)},x)}] \\
&= \frac{1}{D}\left\{\sum_{i=1}^{D}(1-4\delta)^N\mathbb{E}_{x\sim U_N}[z_0^{c_{\mathrm{ch}}(\lambda,\mu_i,x)+c_{\mathrm{ch}}(\lambda',\mu'_i,x)}]\right\} \\
&= \frac{1}{D}\left\{\sum_{i=1}^{D}(1-4\delta)^N\cdot z_0^{|B_i|}\mathbb{E}_{\alpha\sim\mathrm{Bin}\left(|A_i|,\frac{1}{2}\right)}[z_0^{2\alpha}]\prod_{\sigma_i:|\sigma_i|\geqslant 4}\mathbb{E}_{\alpha\sim\mathcal{B}(\sigma_i)}[z_0^{\alpha}]\right\} \\
&= \frac{1}{D}\left\{\sum_{i=1}^{D}(1-4\delta)^N\cdot z_0^{|B_i|}\left(\frac{1+z_0^2}{2}\right)^{|A_i|}\prod_{\sigma_i:|\sigma_i|\geqslant 4}\mathbb{E}_{\alpha\sim\mathcal{B}(\sigma_i)}[z_0^{\alpha}]\right\}, \qquad (16)
\end{aligned}$$

where the product is taken over all cycles in the multigraph $G_{\theta,\theta'}$ induced by the two matchings; and, given a cycle σ, $\mathcal{B}(\sigma)$ is the probability distribution defined as follows. Say that a cycle σ is *even* (resp., *odd*) if the number of edges with weight 1 along σ is even (resp., odd); that is, a cycle is even or odd depending on whether the number of negatively correlated pairs along the cycle is an even or odd number. If σ is an *even cycle*, then $\mathcal{B}(\sigma)$ is a Binomial with parameters $|\sigma|$ and $1/2$, conditioned on taking even values. Similarly, if σ is an *odd cycle*, $\mathcal{B}(\sigma)$ is a Binomial with parameters $|\sigma|$ and $1/2$, conditioned on taking odd values. It follows that $\mathbb{E}_{\alpha\sim\mathcal{B}(\sigma)}[z_0^{\alpha}]$ is given by the following expression.

$$\mathbb{E}_{\alpha\sim\mathcal{B}(\sigma)}[z_0^{\alpha}] = \begin{cases} \mathbb{E}_{\alpha\sim\mathrm{Bin}\left(|\sigma|,\frac{1}{2}\right)}[z^{\alpha}\mid \alpha \text{ even}] = \frac{(1+z_0)^{|\sigma|}+(1-z_0)^{|\sigma|}}{2^{|\sigma|}} & \text{, if } \sigma \text{ is even} \\ \mathbb{E}_{\alpha\sim\mathrm{Bin}\left(|\sigma|,\frac{1}{2}\right)}[z^{\alpha}\mid \alpha \text{ odd}] = \frac{(1+z_0)^{|\sigma|}-(1-z_0)^{|\sigma|}}{2^{|\sigma|}} & \text{, if } \sigma \text{ is odd} \end{cases}.$$

Denote

$$\begin{aligned}
\mathcal{S}_{e,i} &\coloneqq \mathcal{S}_e(\theta_i,\theta'_i) = \{\sigma\in\mathrm{cycle}(\theta_i,\theta'_i) : |\sigma|\geqslant 4, \sigma \text{ is even}\}, \\
\mathcal{S}_{o,i} &\coloneqq \mathcal{S}_o(\theta_i,\theta'_i) = \{\sigma\in\mathrm{cycle}(\theta_i,\theta'_i) : |\sigma|\geqslant 4, \sigma \text{ is odd}\}.
\end{aligned}$$

We will often drop λ, μ or i, when clear from context. We expand $\mathbb{E}_{\alpha\sim\mathcal{B}(\sigma)}[z_0^{\alpha}]$ as follows:

$$\begin{aligned}
\prod_{\sigma:|\sigma|\geqslant 4}\mathbb{E}_{\alpha\sim\mathcal{B}(\sigma)}[z_0^{\alpha}] &= \prod_{\sigma:|\sigma|\geqslant 4,\mathrm{even}}\frac{(1+z_0)^{|\sigma|}+(1-z_0)^{|\sigma|}}{2^{|\sigma|}}\prod_{\sigma:|\sigma|\geqslant 4,\mathrm{odd}}\frac{(1+z_0)^{|\sigma|}-(1-z_0)^{|\sigma|}}{2^{|\sigma|}} \\
&= \prod_{\substack{\sigma:|\sigma|\geqslant 4\\ \mathrm{even}}}\frac{(1+z_0)^{|\sigma|}}{2^{|\sigma|}}\left(1+\left(\frac{1-z_0}{1+z_0}\right)^{|\sigma|}\right)\prod_{\substack{\sigma:|\sigma|\geqslant 4\\ \mathrm{odd}}}\frac{(1+z_0)^{|\sigma|}}{2^{|\sigma|}}\left(1-\left(\frac{1-z_0}{1+z_0}\right)^{|\sigma|}\right) \\
&= \prod_{\sigma:|\sigma|\geqslant 4}\frac{(1+z_0)^{|\sigma|}}{2^{|\sigma|}}\prod_{\sigma\in\mathcal{S}_e}(1+(-4\delta)^{|\sigma|})\prod_{\sigma\in\mathcal{S}_o}(1-(-4\delta)^{|\sigma|}) \\
&= \frac{(1+z_0)^{\sum_{\sigma:|\sigma|\geqslant 4}|\sigma|}}{2^{\sum_{\sigma:|\sigma|\geqslant 4}|\sigma|}}\prod_{\sigma\in\mathcal{S}_e}(1+(-4\delta)^{|\sigma|})\prod_{\sigma\in\mathcal{S}_o}(1-(-4\delta)^{|\sigma|}) \\
&= \frac{(1+z_0)^{|C|}}{2^{|C|}}\prod_{\sigma\in\mathcal{S}_e}(1+(-4\delta)^{|\sigma|})\prod_{\sigma\in\mathcal{S}_o}(1-(-4\delta)^{|\sigma|})
\end{aligned}$$

where for the last equality we used that $\sum_{\sigma:|\sigma|\geqslant 4}|\sigma| = |C|$. We now improve upon the analogous analysis from [Canonne et al., 2020, Claim 12] to obtain a better upper bound for the remaining terms; indeed, the bound they derived is $e^{O(\varepsilon^4/n)}$, which was enough for their purposes but not ours (since it does not feature any dependence on d). Let $z_1 \coloneqq \frac{1+(4\delta)^2}{1-(4\delta)^2}$. In view of using the above expression to bound (16), we first simplify (part of) the summands of (16) by using the fact that $2|A_i| + 2|B_i| + |C_i| = N$ for all i, and following the same computations as in Canonne et al. [2020]:

$$\begin{aligned}
&(1-4\delta)^N z_0^{|B_i|}\left(\frac{1+z_0^2}{2}\right)^{|A_i|}\frac{(1+z_0)^{|C_i|}}{2^{|C_i|}} \\
&\qquad = (1-4\delta)^N z_0^{|B_i|}\left(\frac{1+z_0^2}{2}\right)^{|A_i|}\frac{(1+z_0)^{|C_i|}}{2^{|C_i|}} \\
&\qquad = ((1-4\delta)^2 z_0)^{|B_i|}\left((1-4\delta)^2\frac{1+z_0^2}{2}\right)^{|A_i|}\underbrace{\left((1-4\delta)\frac{1+z_0}{2}\right)^{|C_i|}}_{=1} \\
&\qquad = (1-(4\delta)^2)^{|B_i|}(1+(4\delta)^2)^{|A_i|} \\
&\qquad = (1-(4\delta)^2)^{|A_i|+|B_i|} z_1^{|A_i|} = (1-(4\delta)^2)^{|A_1|+|B_1|} z_1^{|A_i|},
\end{aligned}$$

where the last equality uses the fact that the sum $|A_i| + |B_i|$ only depends on the matchings λ, λ' (not the orientations μ_i, μ_i'), and thus is independent of i. Plugging this simplification into (16), and letting $R \coloneqq |A_1| + |B_1| \leqslant N/2$ for convenience, we get

$$\begin{aligned}
1+\tau(\theta,\theta') &= \frac{1}{D}\left\{\sum_{i=1}^{D}(1-4\delta)^N \cdot z_0^{|B_i|}\left(\frac{1+z_0^2}{2}\right)^{|A_i|}\prod_{\sigma_i}\mathbb{E}_{\alpha\sim\mathcal{B}(\sigma_i)}[z_0^{\alpha}]\right\} \\
&= (1-(4\delta)^2)^R \cdot \frac{1}{D}\left\{\sum_{i=1}^{D} z_1^{|A_i|}\prod_{\sigma\in\mathcal{S}_{e,i}}(1+(-4\delta)^{|\sigma|})\prod_{\sigma\in\mathcal{S}_{o,i}}(1-(-4\delta)^{|\sigma|})\right\}.
\end{aligned}$$

Next, we compute the expectation after raising the above to the power m.

$$\begin{aligned}
&\mathbb{E}_{\theta,\theta'}[(1+\tau(\theta,\theta'))^m] \\
&\quad = \mathbb{E}_{\theta,\theta'}\left[\left((1-(4\delta)^2)^R\frac{1}{D}\left\{\sum_{i=1}^{D} z_1^{|A_i|}\prod_{\sigma\in\mathcal{S}_{e,i}}(1+(-4\delta)^{|\sigma|})\prod_{\sigma\in\mathcal{S}_{o,i}}(1-(-4\delta)^{|\sigma|})\right\}\right)^m\right] \\
&\quad = \mathbb{E}_{\lambda,\lambda'}\left[(1-(4\delta)^2)^{mR}\mathbb{E}_{\vec{\mu},\vec{\mu}'}\left[\left(\frac{1}{D}\left\{\sum_{i=1}^{D} z_1^{|A_i|}\prod_{\sigma\in\mathcal{S}_{e,i}}(1+(-4\delta)^{|\sigma|})\prod_{\sigma\in\mathcal{S}_{o,i}}(1-(-4\delta)^{|\sigma|})\right\}\right)^m\right]\right].
\end{aligned} \tag{17}$$

The quantity inside the inner expectation is quite unwieldy; to proceed, we will rely on the following identity, which lets us bound the two product terms:

$$\prod_{\sigma\in\mathcal{S}_e}(1+(-4\delta)^{|\sigma|})\prod_{\sigma\in\mathcal{S}_o}(1-(-4\delta)^{|\sigma|}) \leqslant e^{c'\frac{\varepsilon^5}{N^{3/2}}}\prod_{\sigma\in\mathcal{S}_e:|\sigma|=4}(1+(-4\delta)^{|\sigma|})\prod_{\sigma\in\mathcal{S}_o:|\sigma|=4}(1-(-4\delta)^{|\sigma|}), \tag{18}$$

for some absolute constant $c' > 0$. We defer the proof of this inequality to Appendix C.4, and proceed assuming it. Note that as long as $D = O(n/\varepsilon^6)$, we will have $e^{c'\cdot m\varepsilon^5/N^{3/2}} \leqslant e^{128m\varepsilon^2/(\sqrt{D}n)}$,[4] and this restriction on D is satisfied for the regime of parameters considered in our lower bound, $d = O(\log n)$.

Fix a pair λ, λ'; we have that the $|A_i|$'s are i.i.d. $\mathrm{Bin}(R, 1/2)$ random variables. We now introduce

$$R' \coloneqq |\{\sigma_1 : |\sigma_1| = 4\}| = |\{\sigma_i : |\sigma_i| = 4\}|,$$

[4] As per the condition set in Lemma B.5, we will from now on assume that $n/\varepsilon^2 \geqslant n \geqslant 40D$, which gives us $N^{3/2} \geqslant (n/2)^{3/2} \geqslant \frac{n}{2}\cdot(20D)^{1/2} \geqslant 2n\sqrt{D}$; and some more calculations give us $c'\frac{m\varepsilon^2}{N^{3/2}} \leqslant \frac{128m\varepsilon^2}{\sqrt{D}n}$.

which is the random variable denoting how many cycles have length exactly 4. In particular, we have $R' \leqslant \frac{N}{4}$, since $\sum_{\sigma:|\sigma|\geqslant 4} \sigma = |C| \leqslant N$; more specifically, we have $R' \leqslant \frac{N-2R}{4} \leqslant \frac{N}{4}$. Further, define κ_i as the number of cycles of length 4 which have an even total number of negative correlations; that is, the number of cycles σ such that μ_i, μ_i' impose either 0, 2, or 4 negatively correlated pairs along that cycle.

Since μ, μ' are uniformly distributed, being odd or even each has probability $1/2$, and thus $\kappa_i \sim \mathrm{Bin}\big(R', \frac{1}{2}\big)$. Moreover, while κ_i and A_i both depend on μ_i, μ_i', they by definition depend on disjoint subsets of those two random variables: thus, because each correlation parameter is chosen independently, we have that κ_i and A_i are independent conditioned on (R, R'). Now, recalling our setting of $z_2 = \frac{1+(4\delta)^4}{1-(4\delta)^4}$ and fixing a realization of R, R', we have

$$
\begin{aligned}
&\mathbb{E}_{\vec{\mu},\vec{\mu}'}\left[\left(\frac{1}{D}\sum_{i=1}^{D} z_1^{|A_i|} \prod_{\sigma\in\mathcal{S}_e(i):|\sigma|=4}(1+(4\delta)^4) \prod_{\sigma_i\in\mathcal{S}_o(i):|\sigma|=4}(1-(4\delta)^4)\right)^m\right] \\
&= \mathbb{E}_{\vec{\mu},\vec{\mu}'}\left[\left(\frac{1}{D}\sum_{i=1}^{D} z_1^{|A_i|}(1+(4\delta)^4)^{\kappa_i}(1-(4\delta)^4)^{R'-\kappa_i}\right)^m\right] \\
&= (1-(4\delta)^4)^{mR'}\mathbb{E}_{\vec{\mu},\vec{\mu}'}\left[\left(\frac{1}{D}\sum_{i=1}^{D} z_1^{|A_i|} z_2^{\kappa_i}\right)^m\right] \\
&\leqslant (1-(4\delta)^4)^{mR'} z_1^{\frac{mR}{2}} z_2^{\frac{mR'}{2}}\,\mathbb{E}_{\vec{\alpha}}\left[\prod_{i=1}^{D}(\cosh(2\alpha_i\delta^2))^R(\cosh(2\alpha_i\delta^4))^{R'}\right], \qquad (19)
\end{aligned}
$$

where (19) follows from the following lemma, whose proof we defer to the end of the section:

Lemma 2.5. *There exists an absolute constant $\delta_0 \approx 0.96$ such that the following holds. Let $K \geqslant 1$ and $R_1, \ldots, R_K$ be integers, and $\delta_1, \ldots, \delta_K \in (0, \delta_0]$. Suppose that $\kappa_{j,1}, \ldots, \kappa_{j,D} \sim \mathrm{Bin}\big(R_j, \frac{1}{2}\big)$, are i.i.d., and mutually independent across $1 \leqslant j \leqslant K$, and $z_j \coloneqq \frac{1+\delta_j}{1-\delta_j}$. Then*

$$
\mathbb{E}\left[\left(\frac{1}{D}\textstyle\sum_{i=1}^{D}\prod_{j=1}^{K} z_j^{\kappa_{j,i}}\right)^m\right] \leqslant \left(\prod_{j=1}^{K} z_j^{\frac{m}{2}R_j}\right)\mathbb{E}_{\vec{\alpha}}\left[\prod_{i=1}^{D}\prod_{j=1}^{K}\cosh(2\alpha_i\delta_j)\right],
$$

where $(\alpha_1, \ldots, \alpha_D)$ follows a multinomial distribution with parameters m and $(1/D, \ldots, 1/D)$.

We now focus on the expectation on the right (last factor of the RHS of (19)): using that $\cosh u \leqslant \min(e^{u^2/2}, e^u)$ for $u \geqslant 0$, we have, setting $\Delta \coloneqq 1/\delta^2 = n/\varepsilon^2$,

$$
\begin{aligned}
&\mathbb{E}_{\vec{\alpha}}\left[\prod_{i=1}^{D}(\cosh(2\alpha_i\delta^2))^R(\cosh(2\alpha_i\delta^4))^{R'}\right] \\
&\leqslant \mathbb{E}_{\vec{\alpha}}\left[\prod_{i=1}^{D}\min(e^{2\alpha_i\delta^2 R}, e^{2\alpha_i^2\delta^4 R})e^{2\alpha_i^2\delta^8 R'}\right] \\
&\leqslant \mathbb{E}_{\vec{\alpha}}\left[\prod_{i=1}^{D} e^{2\alpha_i\delta^2 R\mathbb{1}[\alpha_i>\Delta]} e^{2\alpha_i^2\delta^4 R\mathbb{1}[\alpha_i\leqslant\Delta]} e^{2\alpha_i^2\delta^8 R'}\right] \\
&\leqslant \mathbb{E}\left[\prod_{i=1}^{D} e^{8\alpha_i\delta^2 R\mathbb{1}[\alpha_i>\Delta]}\right]^{1/4}\mathbb{E}\left[\prod_{i=1}^{D} e^{8\alpha_i^2\delta^4 R\mathbb{1}[\alpha_i\leqslant\Delta]}\right]^{1/4}\mathbb{E}\left[\prod_{i=1}^{D} e^{4\alpha_i^2\delta^8 R'}\right]^{1/2} \qquad (20)
\end{aligned}
$$

where the last step comes from the generalized Hölder inequality (or, equivalently, two applications of the Cauchy–Schwarz inequality), and the threshold Δ was chosen as the value for which the term realizing the minimum changes. We first bound the product of the last two expectations:

$$
\mathbb{E}\left[\prod_{i=1}^{D} e^{8\alpha_i^2\delta^4 R\mathbb{1}[\alpha_i\leqslant\Delta]}\right]^{1/4}\mathbb{E}\left[\prod_{i=1}^{D} e^{4\alpha_i^2\delta^8 R'}\right]^{1/2}
$$

$$\leqslant \mathbb{E}\left[\prod_{i=1}^{D} e^{8\delta^4 R \min(\alpha_i^2,\Delta^2)}\right]^{1/4} \mathbb{E}\left[\prod_{i=1}^{D} e^{4\alpha_i^2\delta^8 R'}\right]^{1/2}$$

$$\leqslant (\mathbb{E}[e^{8\min(\alpha_j^2,\Delta^2)\delta^4 R}])^{D/4}(\mathbb{E}[e^{4\alpha_1^2\delta^8 R'}])^{D/2} \tag{21}$$

$$\leqslant \exp\left(32\delta^4\frac{m^2}{D}R\right)\exp\left(32\delta^8\frac{m^2}{D}R'\right). \tag{22}$$

where we applied negative association (see, e.g., Dubhashi and Ranjan [1996, Theorem 13]) on both expectations for (21); and then got (22) by Lemmas B.7 and B.3 (for the latter, noting that $tm = 2\delta^8 mR' \leqslant 1/16$; and, for the former, assuming with little loss of generality that $\varepsilon \leqslant 1/(4\sqrt{2})$). Applying Lemma B.6 to the first (remaining) factor of the LHS above as $8\delta^2 R \leqslant 4$ and $D \geqslant \Omega(1)$, we get

$$\mathbb{E}\left[\prod_{i=1}^{D} e^{8\alpha_i\delta^2 R\mathbb{1}[\alpha_i>\Delta]}\right]^{1/4} \mathbb{E}\left[\prod_{i=1}^{D} e^{8\alpha_i^2\delta^4 R\mathbb{1}[\alpha_i\leqslant\Delta]}\right]^{1/4} \mathbb{E}\left[\prod_{i=1}^{D} e^{4\alpha_i^2\delta^8 R'}\right]^{1/2}$$
$$\leqslant (1+o(1))\cdot\exp\left(32\delta^4\frac{m^2}{D}R\right)\exp\left(32\frac{m^2}{D}\delta^8 R'\right)$$
$$= (1+o(1))\exp(32C'^2 R),$$

recalling that $R' \leqslant N/4 \leqslant n/4$, and our assumption that $m \leqslant C'\sqrt{D}n/\varepsilon^2$. Combining (17), (18), and (19), what we showed is

$$\mathbb{E}_{\theta,\theta'}[(1+\tau(\theta,\theta'))^m] \leqslant (1+o(1))e^{128\frac{m\varepsilon^2}{\sqrt{D}n}}\mathbb{E}_{\lambda,\lambda'}\left[(1-(4\delta)^2)^{mR}(1-(4\delta)^4)^{mR'}z_1^{\frac{mR}{2}}z_2^{\frac{mR'}{2}}e^{32C'^2\cdot R}\right]$$
$$= (1+o(1))e^{128\cdot C'}\mathbb{E}_{\lambda,\lambda'}\left[(1-(4\delta)^4)^{\frac{mR}{2}}(1-(4\delta)^8)^{\frac{mR'}{2}}e^{32C'^2\cdot R}\right]$$
$$\leqslant (1+o(1))e^{128\cdot C'}\mathbb{E}_{\lambda,\lambda'}\left[e^{32C'^2\cdot R}\right],$$

where the equality follows from the definition of z_1, z_2. To conclude, we will use the fact that, for every $k \geqslant 0$,

$$\Pr[R > k] \leqslant \frac{1}{k!}, \tag{23}$$

which was established in Canonne et al. [2020, p.46]. By summation by parts, one can show that this implies

$$\mathbb{E}_R\left[e^{\alpha R}\right] \leqslant 1 + (1-e^{-\alpha})(e^{e^{\alpha}}-1) \xrightarrow[\alpha\to 0]{} 1+\alpha(e-1)$$

for any $\alpha > 0$, and so, in our case,

$$\mathbb{E}_{\theta,\theta'}[(1+\tau(\theta,\theta'))^m] \leqslant (1+o(1))e^{128\cdot C'}\left(1+(1-e^{-32C'^2})(e^{e^{32C'^2}}-1)\right) \tag{24}$$

In particular, the RHS can be made arbitrarily close to 1 by choosing a small enough value for the constant C' (in the bound for m). By (14), this implies the desired bound on $\|Q - U^{\otimes m}\|_1^2$, and thus establishes Lemma 2.3. □

The remaining technical lemma. It only remains to establish Lemma 2.5, which we do now.

Lemma 2.5. *There exists an absolute constant $\delta_0 \approx 0.96$ such that the following holds. Let $K \geqslant 1$ and $R_1, \ldots, R_K$ be integers, and $\delta_1, \ldots, \delta_K \in (0, \delta_0]$. Suppose that $\kappa_{j,1}, \ldots, \kappa_{j,D} \sim \mathrm{Bin}\big(R_j, \frac{1}{2}\big)$, are i.i.d., and mutually independent across $1 \leqslant j \leqslant K$, and $z_j := \frac{1+\delta_j}{1-\delta_j}$. Then*

$$\mathbb{E}\left[\left(\frac{1}{D}\sum_{i=1}^{D}\prod_{j=1}^{K} z_j^{\kappa_{j,i}}\right)^m\right] \leqslant \left(\prod_{j=1}^{K} z_j^{\frac{m}{2}R_j}\right)\mathbb{E}_{\vec{\alpha}}\left[\prod_{i=1}^{D}\prod_{j=1}^{K}\cosh(2\alpha_i\delta_j)\right],$$

where $(\alpha_1, \ldots, \alpha_D)$ follows a multinomial distribution with parameters m and $(1/D, \ldots, 1/D)$.

Proof of Lemma 2.5. We will require the following simple fact, which follows from the multinomial theorem and the definition of the multinomial distribution:

Fact A.1. *Let D be a positive integer and m be a non-negative integer. For any $x_1, \ldots, x_D \in \mathbb{R}$, we have*

$$\left(\frac{1}{D}\sum_{i=1}^{D} x_i\right)^m = \mathbb{E}_{\alpha_1,\ldots,\alpha_D}\left[\prod_{i=1}^{D} x_i^{\alpha_i}\right],$$

where $(\alpha_1, \ldots, \alpha_D)$ follows a multinomial distribution with parameters m and $(1/D, \ldots, 1/D)$.

We now apply Fact A.1 inside the expectation of the LHS of the statement. Note that the sets of random variables $\vec{\alpha} = \{\alpha_1, \ldots, \alpha_D\}$, $\vec{\kappa}_1 = \{\kappa_{1,1}, \ldots, \kappa_{1,D}\}, \ldots, \vec{\kappa}_K = \{\kappa_{K,1}, \ldots, \kappa_{K,D}\}$ are mutually independent, since $\vec{\alpha}$ are a set of auxiliary random variables derived from an averaging operation and by the assumption on $\vec{\kappa}_j$; and we have that $\kappa_{j,1}, \ldots, \kappa_{j,D}$ are i.i.d.,

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{D}\sum_{i=1}^{D}\prod_{j=1}^{K} z_j^{\kappa_{j,i}}\right)^m\right] &= \mathbb{E}\left[\mathbb{E}_{\alpha_1,\ldots,\alpha_D}\left[\prod_{i=1}^{D}\left(\prod_{j=1}^{K} z_j^{\kappa_{j,i}}\right)^{\alpha_i}\right]\right] \\
&= \mathbb{E}_{\vec{\alpha},\vec{\kappa}_j}\left[\prod_{i=1}^{D}\prod_{j=1}^{K} z_j^{\alpha_i\kappa_{j,i}}\right] \\
&= \mathbb{E}_{\vec{\alpha}}\left[\mathbb{E}_{\vec{\kappa}_j}\left[\prod_{i=1}^{D}\prod_{j=1}^{K} z_j^{\alpha_i\kappa_{j,i}}\right]\right] \\
&= \mathbb{E}_{\vec{\alpha}}\left[\left[\prod_{i=1}^{D}\prod_{j=1}^{K} \mathbb{E}_{\vec{\kappa}_j}[z_j^{\alpha_i\kappa_{j,i}}]\right]\right] \\
&= \mathbb{E}_{\vec{\alpha}}\left[\left[\prod_{i=1}^{D}\prod_{j=1}^{K} \left(\frac{1+z_j^{\alpha_i}}{2}\right)^{R_j}\right]\right] \\
&\qquad\qquad \text{(Probability-Generating Function of a Binomial)} \\
&= \mathbb{E}_{\vec{\alpha}}\left[\prod_{i=1}^{D}\prod_{j=1}^{K} z_j^{\frac{\alpha_i R_j}{2}}\left(\frac{z_j^{-\alpha_i/2}+z_j^{\alpha_i/2}}{2}\right)^{R_j}\right] \\
&= \left(\prod_{j=1}^{K} z_j^{\frac{mR_j}{2}}\right)\mathbb{E}_{\vec{\alpha}}\left[\prod_{i=1}^{D}\prod_{j=1}^{K}\left(\frac{z_j^{-\alpha_i/2}+z_j^{\alpha_i/2}}{2}\right)^{R_j}\right].
\end{aligned}$$

Next, we will simplify the expression left inside by upper bounding it, using the fact that, given our assumption on δ_j being bounded above by δ_0, we have $z_j = \frac{1+\delta_j}{1-\delta_j} \leqslant e^{4\delta_j}$. Thus,

$$\begin{aligned}
\mathbb{E}_{\vec{\alpha}}\left[\prod_{i=1}^{D}\prod_{j=1}^{K}\left(\frac{z_j^{-\alpha_i/2}+z_j^{\alpha_i/2}}{2}\right)^{R_j}\right] &\leqslant \mathbb{E}_{\vec{\alpha}}\left[\prod_{i=1}^{D}\prod_{j=1}^{K}\left(\frac{e^{-2\alpha_i\delta_j}+e^{2\alpha_i\delta_j}}{2}\right)^{R_j}\right] \\
&= \mathbb{E}_{\vec{\alpha}}\left[\prod_{i=1}^{D}\prod_{j=1}^{K}(\cosh(2\alpha_i\delta_j))^{R_j}\right]
\end{aligned}$$

as claimed. □

A.2 Product Distributions Are Far from Mixture of Trees (Lemma 2.4)

In this subsection, we outline the proof of Lemma 2.4. Our argument starts with Lemma A.2, which allows us to relate the total variation distance between the mixture and the product of its marginals to a simpler quantity, the difference between two components of this mixture.

Lemma A.2. *Let p be a distribution on $\{0,1\}^N \times \{0,1\}^M$ (with $N, M \geqslant 2$), and denote its marginals on $\{0,1\}^N$, $\{0,1\}^M$ by p_1, p_2 respectively. Then, if p_1 is uniform,*

$$d_{\mathrm{TV}}(p, p_1 \otimes p_2) \;\geqslant d_{\mathrm{TV}}(p(\cdot \mid x_1 = 0), p(\cdot \mid x_1 = 1))\,.$$

This in turn will be much more manageable, as the parameters of these two mixture components are independent, and thus analyzing this distance can be done by analyzing Binomial-like expressions. This second step is reminiscent of [Canonne et al., 2020, Lemma 8], which can be seen as a simpler version involving only one Binomial instead of two:

Lemma A.3. *There exist $C_1, C_2 > 0$ such that the following holds. Let $\varepsilon \in (0,1]$ and $n \geqslant C_1$, and let a, b be two integers such that $a + b = n$ and $b \geqslant \frac{1}{4}n$. Then, for $\delta \coloneqq \frac{\varepsilon}{\sqrt{n}}$, we have*

$$\frac{(1-\delta)^n}{2^n} \sum_{k_1=0}^{a} \sum_{k_2=0}^{b} \binom{a}{k_1}\binom{b}{k_2} \left| \left(\left(\frac{1+\delta}{1-\delta}\right)^{k_1+k_2} - \left(\frac{1+\delta}{1-\delta}\right)^{k_1+b-k_2} \right) \right| \geqslant C_2 \varepsilon.$$

This parameter b corresponds to the difference between the orientations parameters μ, μ' being large, which happens with high constant probability as long as n is large enough. The proof of Lemma A.3 is deferred to Appendix C.3, and we hereafter proceed with the rest of the argument. For fixed θ and $x_2, \ldots, x_d$, $z \coloneqq \frac{1+4\delta}{1-4\delta}$. We will denote by μ, μ' the two (randomly chosen) orientation parameters corresponding to the mixture components indexed by $(0, x_2, \ldots, x_d)$ and $(1, x_2, \ldots, x_d)$. By Lemma A.2 and Lemma 1.5, for any product distribution q,

$$2d_{\mathrm{TV}}(p,q) \geqslant \frac{1}{3 \cdot 2^N}(1-4\delta)^{\frac{N}{2}} \sum_{x_{d+1},\ldots,x_n} |z^{c(\lambda,\mu,x)} - z^{c(\lambda,\mu',x)}| \tag{25}$$

Let S_1 denote the set of pairs in the child nodes with common parameters between μ and μ', and S_2 the set of pairs with different parameters (that is, the definition of S_1, S_2 is essentially that of A and B from the previous section (p.15), but for equal matching parameters $\lambda = \lambda'$). In particular, we have that $|S_2| = \mathrm{Hamming}(\mu, \mu') \sim \mathrm{Binomial}\left(\frac{N}{2}, \frac{1}{2}\right)$ and $|S_1 \cup S_2| = \frac{N}{2}$. Let $\tilde{c}(S, \mu, x)$ be the analogue of $c_{\mathrm{ch}}(\lambda, \mu, x)$ from (13), but only on a subset of pairs S instead of $\{d, \ldots, n\}^2$; i.e.,

$$\tilde{c}(S,\mu,x) \coloneqq \left|\left\{(i,j) \in S : \exists k \in \mathbb{N}, \lambda_k = (i-d+1, j-d+1) \text{ and } (-1)^{x_i+x_j} = (-1)^{\mu_k}\right\}\right|.$$

Given any x, μ and μ', the following holds from the definitions of $\tilde{c}$ and c_{ch}:

- Since $S_1 \cup S_2$ contains all the pairs, $c_{\mathrm{ch}}(\lambda, \mu, x) = \tilde{c}(S_1, \mu, x) + \tilde{c}(S_2, \mu, x)$ (similarly for μ').
- Since S_1 (resp., S_2) contains exactly the pairs whose orientation is the same (resp., differs) between μ and μ', we have $\tilde{c}(S_1, \mu', x) = \tilde{c}(S_1, \mu, x)$ and $\tilde{c}(S_2, \mu', x) + \tilde{c}(S_2, \mu, x) = |S_2|$
- For a fixed matching and a partition S_1, S_2 of its $N/2$ pairs, given an orientation vector $\mu \in \{0,1\}^{N/2}$, and fixed values $0 \leqslant k_1 \leqslant |S_1|$, $0 \leqslant k_2 \leqslant |S_2|$, there are $2^{N/2}\binom{|S_1|}{k_1}\binom{|S_2|}{k_2}$ different vectors $x \in \{0,1\}^N$ such that $\tilde{c}(S_1, \mu, x) = k_1$ and $\tilde{c}(S_2, \mu, x) = k_2$.

Using these properties, we have, assuming $|S_2| \geqslant \frac{1}{4} \cdot \frac{N}{2}$ and N bigger than some constant,

$$\begin{aligned}
\sum_{x \in \{0,1\}^N} |z^{c_{\mathrm{ch}}(\lambda,\mu,x)} - z^{c_{\mathrm{ch}}(\lambda,\mu',x)}| &= \sum_{x \in \{0,1\}^N} |z^{\tilde{c}(S_1,\mu,x)+\tilde{c}(S_2,\mu,x)} - z^{\tilde{c}(S_1,\mu',x)+\tilde{c}(S_2,\mu',x)}| \\
&= \sum_{x \in \{0,1\}^N} |z^{\tilde{c}(S_1,\mu,x)+\tilde{c}(S_2,\mu,x)} - z^{\tilde{c}(S_1,\mu,x)+|S_2|-\tilde{c}(S_2,\mu,x)}| \\
&= 2^{\frac{N}{2}} \sum_{k_1=0}^{|S_1|} \binom{|S_1|}{k_1} \sum_{k_2=0}^{|S_2|} \binom{|S_2|}{k_2} |z^{k_1+k_2} - z^{k_1+|S_2|-k_2}| \\
&\geqslant C \cdot \frac{2^N}{(1-4\delta)^{N/2}} \cdot \varepsilon,
\end{aligned}$$

where $C > 0$ is an absolute constant, and for the last inequality we invoked Lemma A.3. Recalling now that $|S_2| \sim \mathrm{Bin}\left(\frac{N}{2}, \frac{1}{2}\right)$, for N large enough we also have

$$\Pr\left[|S_2| \geqslant \frac{N}{8}\right] \geqslant 1 - e^{-\frac{N}{16}} > 9/10.$$

Thus, combining the two along with (25), we conclude that

$$\Pr[d_{\mathrm{TV}}(P_\theta(\cdot \mid x_1 = 0, x_2, \ldots, x_d), P_\theta(\cdot \mid x_1 = 1, x_2, \ldots, x_d)) \geqslant \Omega(\varepsilon)] \geqslant 9/10,$$

establishing Lemma 2.4. □

B Useful results on the MGFs of Binomials and Multinomials

In this section, we establish various self-contained results on the moment-generating functions (MGF) and stochastic dominance of Binomials, truncated (or "capped") Binomials, and multinomial distributions, which we used extensively in Section 2.1 and should be of independent interest. Notably, derivations from Section 2 following (10) are direct consequences of the three lemmas in the section: Lemma B.3, Lemma B.6 and Lemma B.7 below, which we restate and establish later in this section.

Lemma B.3. *Let $X \sim \mathrm{Bin}(m, p)$. Then, for any t such that $0 < tm \leqslant 1/16$,*

$$\mathbb{E}[e^{tX^2}] \leqslant \exp(16tm^2p^2 + 2tmp).$$

Lemma B.6. *Suppose $\vec{\alpha} = (\alpha_1, \ldots, \alpha_D)^T$ follows a multinomial distribution with parameters m and $(1/D, \ldots, 1/D)$, and $\Delta \geqslant 40D \gg c^4$ be such that $m \leqslant c\sqrt{D}\Delta$. Then, for any $t \leqslant 4$ and $D \geqslant \Omega(1)$, we have*

$$\mathbb{E}\left[\prod_{i=1}^{D} e^{t\alpha_i \mathbb{1}[\alpha_i > \Delta]}\right] \leqslant 1 + c\sqrt{D}\exp\left(-\frac{1}{80}\Delta \log D\right).$$

Lemma B.7. *Let $X' \sim \mathrm{Bin}(m, p)$, and $X \coloneqq \min(X', \Delta)$, for some $\Delta \leqslant m$. Then, for any t such that $0 < t\Delta \leqslant 1/8$ and $0 < tmp \leqslant 1/16$, we have*

$$\mathbb{E}[e^{tX^2}] \leqslant \exp(16tm^2p^2 + 2tmp).$$

B.1 Bounds on moment-generating functions

We start with some relatively simple statements:

Fact B.1. *If $X \sim \mathrm{Bin}(m, p)$, then, for any $0 \leqslant t \leqslant 1$, $\mathbb{E}[e^{tX}] \leqslant \exp(2tmp)$.*

Proof. This follows from computing explicitly $\mathbb{E}[e^{tX}] = (1 + p(e^t - 1))^m \leqslant (1 + 2tp)^m \leqslant e^{2tmp}$, where the first inequality uses that $t \leqslant 1$. □

We will also require the following decoupling inequality:

Lemma B.2. *Let $F \colon \mathbb{R} \to \mathbb{R}$ be a convex,* non-decreasing *function, and $X = (X_1, \ldots, X_n)$ be a vector of independent* non-negative *random variables. Then*

$$\mathbb{E}\left[F\left(\sum_{i \neq j} X_i X_j\right)\right] \leqslant \mathbb{E}\left[F\left(4\sum_{i,j} X_i Y_j\right)\right]$$

where Y is an independent copy of X.

Proof. Introduce a vector of independent (and independent of X) $\mathrm{Bern}(1/2)$ random variables $\delta = (\delta_1, \ldots, \delta_n)$; so that $\mathbb{E}[\delta_i(1 - \delta_j)] = \frac{1}{4}\mathbb{1}_{i \neq j}$. For any realization of X, we can write

$$\sum_{i \neq j} X_i X_j = 4\mathbb{E}_\delta\left[\sum_{i \neq j} \delta_i(1 - \delta_j) X_i X_j\right] = 4\mathbb{E}_\delta\left[\sum_{i,j} \delta_i(1 - \delta_j) X_i X_j\right],$$

and so, by Jensen's inequality and Fubini, as well as independence of X and δ,

$$\begin{aligned}\mathbb{E}_X\left[F\left(\sum_{i \neq j} X_i X_j\right)\right] &= \mathbb{E}_X\left[F\left(4\mathbb{E}_\delta\left[\sum_{i,j} \delta_i(1 - \delta_j) X_i X_j\right]\right)\right] \\ &\leqslant \mathbb{E}_\delta\left[\mathbb{E}_X\left[F\left(4\sum_{i,j} \delta_i(1 - \delta_j) X_i X_j\right)\right]\right]\end{aligned}$$

This implies that there exists some realization $\delta^* \in \{0,1\}^n$ such that

$$\mathbb{E}_X\left[F\left(\sum_{i\neq j} X_i X_j\right)\right] \leqslant \mathbb{E}_X\left[F\left(4\sum_{i,j}\delta_i^*(1-\delta_j^*)X_iX_j\right)\right].$$

Let $I := \{i \in [n] : \delta_i^* = 1\}$. Then $\sum_{i,j}\delta_i^*(1-\delta_j^*)X_iX_j = \sum_{(i,j)\in I\times I^c} X_iX_j$, and we get

$$\begin{aligned}
\mathbb{E}_X\left[F\left(\sum_{i\neq j} X_i X_j\right)\right] &\leqslant \mathbb{E}_X\left[F\left(4\sum_{(i,j)\in I\times I^c} X_iX_j\right)\right] \\
&= \mathbb{E}_X\left[F\left(4\sum_{(i,j)\in I\times I^c} X_iY_j\right)\right] \\
&\leqslant \mathbb{E}_X\left[F\left(4\sum_{(i,j)\in I\times I^c} X_iY_j + 4\sum_{(i,j)\notin I\times I^c} X_iY_j\right)\right] \\
&= \mathbb{E}_X\left[F\left(4\sum_{i,j} X_iY_j\right)\right],
\end{aligned} \tag{26}$$

where the equality uses the fact that $(X_i)_{i\in I}$ and $(X_j)_{j\in I^c}$ are independent (as I, I^c are disjoint), and so replacing $\sum_{j\in I^c} X_j$ by the identically distributed $\sum_{j\in I^c} Y_j$ does not change the expectation; and the second inequality uses monotonicity of F and non-negativity of X, Y, as $4\sum_{(i,j)\notin I\times I^c} X_iY_j \geqslant 0$. (Note that up to (and including) (26), the assumption that the X_i's are independent is not necessary; we will use this fact later on.) □

Note that compared to the usual version of the inequality, we do not require that the X_i's have mean zero; but instead require that they be non-negative, and that F be monotone. We will, in the next lemma, apply Lemma B.2 to the function $F(x) = e^{2tx}$, for some fixed *positive* parameter $t > 0$ (so that F is indeed non-decreasing), and to $X_1, \ldots, X_n$ independent Bernoulli r.v.'s. Specifically, we obtain the following bound on the MGF of the square of a Binomial:

Lemma B.3. *Let $X \sim \mathrm{Bin}(m,p)$. Then, for any t such that $0 < tm \leqslant 1/16$,*

$$\mathbb{E}[e^{tX^2}] \leqslant \exp(16tm^2p^2 + 2tmp).$$

Proof. Write $X = \sum_{i=1}^m X_i$, where the X_i are i.i.d. $\mathrm{Bern}(p)$ (in particular, $X_i = X_i^2$). Then, by the Cauchy–Schwarz inequality and the decoupling inequality from Lemma B.2, we have, for $t > 0$,

$$\begin{aligned}
\mathbb{E}[e^{tX^2}] &= \mathbb{E}\left[e^{t\sum_i X_i} e^{t\sum_{i\neq j} X_iX_j}\right] \\
&\leqslant \sqrt{\mathbb{E}\left[e^{2t\sum_i X_i}\right]}\sqrt{\mathbb{E}\left[e^{2t\sum_{i\neq j} X_iX_j}\right]} \\
\text{(decoupling)} &\leqslant \sqrt{\mathbb{E}\left[e^{2t\sum_i X_i}\right]}\sqrt{\mathbb{E}\left[e^{8t\sum_{i,j} X_iY_j}\right]}.
\end{aligned} \tag{27}$$

where $Y_j \sim \mathrm{Bern}(p)$ are i.i.d., and independent of the X_i's. Let $Y = \sum_{i=1}^m Y_i \sim \mathrm{Bin}(m,p)$. From Fact B.1, as long as $2t \leqslant 1$, $8tm \leqslant 1$, and $16tmp \leqslant 1$ (all conditions satisfied in view of our assumption),

$$\mathbb{E}_{X,Y}[e^{8tXY}] = \mathbb{E}_X[\mathbb{E}_Y[e^{8tXY}]] \leqslant \mathbb{E}_X[e^{16tXmp}] \leqslant e^{32tm^2p^2},$$

and $\mathbb{E}\left[e^{2tX}\right] \leqslant e^{4tmp}$. Going back to (27), this implies

$$\mathbb{E}[e^{tX^2}] \leqslant \sqrt{\exp(4tmp)}\sqrt{\exp(32tm^2p^2)} = \exp\left(2tmp + 16tm^2p^2\right),$$

concluding the proof. □

We will prove an MGF bound on the truncated Multinomial in Lemma B.6 (noting that using MGF bound of Multinomial distribution is not nearly enough), as required by our analysis on the independence testing lower bound; prior to that, we will need two important lemmas: Lemma B.4 and Lemma B.5. These two lemmas both try to bound the expression with a uniform and more manageable term.

Lemma B.4. *Fix m, Δ, D such that $\frac{m}{\Delta} \leqslant c\sqrt{D}$ for some $c > 0$ (and $D > \max(16c^4, e^{100})$). Fix any integer $k > 0$ and a tuple of non-negative integers $(a_1, \ldots, a_D)$ summing to m such that $L := \sum_{i=1}^{k} a_i > k\Delta$ (in particular, $k \leqslant c\sqrt{D}$). Suppose $(\alpha_1, \ldots, \alpha_D)$ follows a multinomial distribution with parameters m and $(1/D, \ldots, 1/D)$. Then,*

$$e^{4L} \Pr[\vec{\alpha} = (a_1, \ldots, a_D)] \leqslant m \cdot \exp(-\frac{1}{5} L \log D).$$

Proof. Via a multinomial distribution grouping argument, the probability can be bounded by considering a grouping of two random variables, $L_1 = \sum_{i=1}^{k} \alpha_i$ and $L_2 = \sum_{i=k+1}^{D} \alpha_i$, where (L_1, L_2) follows a multinomial distribution with parameters m and $\left(\frac{k}{D}, \frac{D-k}{D}\right)$, namely, recalling $L = \sum_{i=1}^{k} a_i$ and setting $T := \sum_{i=k+1}^{D} a_i$,

$$\Pr[\vec{\alpha} = (a_1, \ldots, a_D)] \leqslant \Pr[L_1 = L, L_2 = T] = \frac{m!}{L!T!} \left(\frac{k}{D}\right)^L \left(\frac{D-k}{D}\right)^T$$

Moreover, note that $m = L + T$. Via Stirling's approximation, we have

$$\frac{m!}{L!T!} \leqslant \exp(m \log m + \log m - L \log L - T \log T) \tag{28}$$

from which we can write, taking the logarithm for convenience,

$$\begin{aligned}
\log\left(\frac{m!}{L!T!}\left(\frac{k}{D}\right)^L \left(\frac{D-k}{D}\right)^T\right) &\leqslant \log m - \left(L \log \frac{LD}{mk} + T \log \frac{TD}{m(D-k)}\right) && (29)\\
&= \log m - \left(L \log \frac{LD}{mk} + T \log\left(\frac{TD}{m(D^{3/4}-k)} \frac{D^{3/4}-k}{D-k}\right)\right) && (30)\\
&\leqslant \log m - m \log(D^{1/4}) + (m-L) \log\left(\frac{D-k}{D^{3/4}-k}\right) && (31)\\
&= \log m + m \log\left(1 + k\frac{D^{1/4}-1}{D-kD^{1/4}}\right) - L \log\left(1 + \frac{D^{1/4}-1}{1-k/D^{3/4}}\right)\\
&\leqslant \log m + mk\frac{D^{1/4}-1}{D-kD^{1/4}} - L \log(D^{1/4}) && (32)\\
&\leqslant \log m + \frac{mk}{D^{1/2}} \frac{1-1/D^{1/4}}{D^{1/4}-c} - \frac{1}{4} L \log D\\
&\leqslant \log m + \frac{cL}{D^{1/4}-c} - \frac{1}{4} L \log D && (33)\\
&\leqslant \log m + L - \frac{1}{4} L \log D && (\text{as } c/D^{1/4} \leqslant 1/2)
\end{aligned}$$

where we used Gibbs' inequality for (31); we then have (32) by $\log(1+x) \leqslant x$ for the first term. (33) then follows from $k \leqslant c\sqrt{D}$ and $km \leqslant k\Delta \cdot c\sqrt{D} \leqslant L \cdot c\sqrt{D}$. Finally,

$$e^{4L} \frac{m!}{L!T!} \left(\frac{k}{D}\right)^L \left(\frac{D-k}{D}\right)^T \leqslant \exp(5L - \frac{1}{4} L \log D) \leqslant m \exp\left(-\frac{1}{5} L \log D\right),$$

the last inequality as long as $\log D > 100$. □

Lemma B.5. *Suppose $\vec{\alpha} = (\alpha_1, \ldots, \alpha_D)^T$ follows a multinomial distribution with parameters m and $(1/D, \ldots, 1/D)$, and that $\frac{m}{\Delta} \leqslant c\sqrt{D}$ for some $c > 0, \Delta \geqslant 1$ with $\Delta \geqslant 40D \geqslant \Omega(c^4)$ and*

$D > \Omega(1)$. *For any integer* $c\sqrt{D} \geqslant k \geqslant 1$ *and any* $t \leqslant 4$,

$$\mathbb{E}\left[\prod_{i:\alpha_i>\Delta} e^{t\alpha_i} \cdot \mathbb{1}[\nu(\vec{\alpha}) = k]\right] \leqslant \exp\left(-\frac{1}{80}\Delta\log(D)\right).$$

where $\nu(\vec{\alpha}) := |\{i : \alpha_i > \Delta\}|$ *denotes the number of coordinates of* $\vec{\alpha}$ *greater than* Δ.

Proof. Without loss of generality, (as later, we will sum over all combinations) assume that $\alpha_1, \ldots, \alpha_k$ are the coordinates larger than Δ, for some integer k; and denote their sum by L. Note that we then have $k\Delta < L \leqslant m \leqslant c\Delta\sqrt{D}$, and thus $0 \leqslant k \leqslant c\sqrt{D}$.

$$\mathbb{E}\left[\prod_{i:\alpha_i>\Delta} e^{4\alpha_i} \cdot \mathbb{1}[\nu(\vec{\alpha}) = k]\right] = \binom{D}{k} \sum_{\alpha_1,\ldots,\alpha_k>\Delta} \sum_{\alpha_{k+1},\ldots,\alpha_D\leqslant\Delta} e^{4\sum_{i=1}^k \alpha_i} \Pr[\vec{\alpha} = (\alpha_1, \ldots, \alpha_D)] \tag{34}$$

A uniform bound on any $\alpha_1, \ldots, \alpha_D$ as specified can be obtained from Lemma B.4; and, combining it with (34), we have an expression that does not depend on the value of $\vec{\alpha}$; from which[5]

$$\begin{aligned}
\mathbb{E}\left[e^{4\sum_{i=1}^k \alpha_i} \mathbb{1}[\nu(\vec{\alpha}) = k]\right] &\leqslant \binom{D}{k} \sum_{\alpha_1,\ldots,\alpha_k>\Delta,\alpha_{k+1},\ldots,\alpha_D\leqslant\Delta} m e^{-\frac{1}{5}\Delta\log D} \\
&\leqslant \binom{D}{k}(m-\Delta)^k \Delta^{D-k} \exp\left(\log m - \frac{1}{5}\Delta\log D\right) \\
&\leqslant \exp\left(k\log D + k\log(m-\Delta) + (D-k)\log\Delta + \log m - \frac{\Delta}{5}\log(D)\right) \\
&= \exp\left(k\log(D\cdot\frac{m-\Delta}{\Delta}) + D\log\Delta + \log m - \frac{\Delta}{5}\log(D)\right) \\
&\leqslant \exp\left(-\frac{1}{5}\Delta\log D + (D+1)\log\Delta + \log(c\sqrt{D}) + \frac{3}{2}k\log(cD)\right) \\
&\leqslant \exp\left(-\frac{1}{10}\Delta\log D + 2c\sqrt{D}\log(cD)\right) \qquad (35) \\
&\leqslant \exp\left(-\frac{1}{80}\Delta\log D\right).
\end{aligned}$$

where (35) follows from $20\frac{D}{\log D} \leqslant \frac{\Delta}{\log\Delta}$, which holds for $\Delta \geqslant 40D$ and D large enough (larger than some absolute constant); and the last inequality holds, given the above constraints, for $D \geqslant 16c^4$. □

Lemma B.6. *Suppose* $\vec{\alpha} = (\alpha_1, \ldots, \alpha_D)^T$ *follows a multinomial distribution with parameters* m *and* $(1/D, \ldots, 1/D)$, *and* $\Delta \geqslant 40D \gg c^4$ *be such that* $m \leqslant c\sqrt{D}\Delta$. *Then, for any* $t \leqslant 4$ *and* $D \geqslant \Omega(1)$, *we have*

$$\mathbb{E}\left[\prod_{i=1}^D e^{t\alpha_i \mathbb{1}[\alpha_i>\Delta]}\right] \leqslant 1 + c\sqrt{D}\exp\left(-\frac{1}{80}\Delta\log D\right).$$

Proof. Let $\nu(\vec{\alpha}) := |\{i : \alpha_i > \Delta\}|$ denote the number of coordinates of $\vec{\alpha}$ greater than Δ. Note that $\nu(\vec{\alpha}) < L := \frac{m}{\Delta}$, and that $L = c\sqrt{D}$s by assumption. We break down the expectation by enumerating over the possible values for $\nu(\vec{\alpha})$, from $0 \leqslant k \leqslant L$:

$$\mathbb{E}\left[\prod_{i=1}^D e^{t\alpha_i \mathbb{1}[\alpha_i>\Delta]}\right] = \mathbb{E}\left[\sum_{k=1}^L \prod_{i:\alpha_i>\Delta} e^{t\alpha_i} \cdot \mathbb{1}[\nu(\vec{\alpha}) = k] + \mathbb{1}[\nu(\vec{\alpha}) = 0]\right]$$

[5] We have the number of terms in the summation upper bounded by the following analysis: $(m-\Delta)^k$ is an upper bound of combinations of $\alpha_1, \ldots, \alpha_k$ with values larger than Δ; and similarly, $(\Delta+1)^{D-k}$ will be the upper bound for the combinations of $\alpha_{k+1}, \ldots, \alpha_D$ with values up to Δ.

$$
\begin{aligned}
&= \sum_{k=1}^{L} \mathbb{E}\left[\prod_{i:\alpha_i>\Delta} e^{t\alpha_i} \cdot \mathbb{1}[\nu(\vec{\alpha}) = k]\right] + 1 \cdot \Pr[\nu(\vec{\alpha}) = 0] \\
&\leqslant L \exp\left(-\frac{1}{80}\Delta \log D\right) + \Pr[\nu(\vec{\alpha}) = 0] \qquad (36) \\
&\leqslant cD^{1/2} \exp\left(-\frac{1}{80}\Delta \log D\right) + 1\,,
\end{aligned}
$$

where (36) follows from Lemma B.5. □

We now state and prove our last lemma, Lemma B.7, on the MGF of the square of a truncated Binomial:

Lemma B.7. *Let $X' \sim \mathrm{Bin}(m, p)$, and $X \coloneqq \min(X', \Delta)$, for some $\Delta \leqslant m$. Then, for any t such that $0 < t\Delta \leqslant 1/8$ and $0 < tmp \leqslant 1/16$, we have*

$$\mathbb{E}[e^{tX^2}] \leqslant \exp(16tm^2p^2 + 2tmp).$$

Proof. We will analyze the sampling process in Definition B.8:

Definition B.8. Fix integers $m \geqslant \Delta \geqslant 1$, and let $X'_1, \ldots, X'_m$ be i.i.d. $\mathrm{Bern}(p)$ random variables. Define the distribution of $X_1, \ldots, X_m$ through the following sampling process:

1. Initialize $X_i = 0$ for all $i \in [m]$; sample $\{X'_i\}_{1 \leqslant i \leqslant m}$ as m i.i.d. $\mathrm{Bern}(p)$;
2. If $\sum_{i\in[m]} X'_i < \Delta$, let $X_i = X'_i$ for all $i \in [m]$;
3. If $\sum_{i\in[m]} X'_i \geqslant \Delta$, let $\mathcal{S}' = \{i \in [m] : X'_i = 1\}$ and let $\mathcal{S}$ be a uniformly random subset of $\mathcal{S}'$ of size Δ; set $X_i = X'_i$ for $i \in \mathcal{S}$.

Consider a sequence of random variable $X_1, \ldots, X_m$ as defined in Definition B.8; each X_i (for $1 \leqslant i \leqslant m$) is supported on $\{0, 1\}$ (so that, in particular, $X_i^2 = X_i$); and $X = \sum_{i\in[m]} X_i$. By the Cauchy–Schwarz inequality,

$$
\begin{aligned}
\mathbb{E}[e^{tX^2}] &= \mathbb{E}\left[e^{t\sum_{i=1}^m X_i + t\sum_{i\neq j} X_iX_j}\right] \\
&\leqslant \sqrt{\mathbb{E}\left[e^{2t\sum_{i=1}^m X_i}\right]}\sqrt{\mathbb{E}\left[e^{2t\sum_{i\neq j} X_iX_j}\right]} \\
&\leqslant \sqrt{\mathbb{E}\left[e^{2t\sum_{i=1}^m X_i}\right]}\sqrt{\mathbb{E}\left[e^{8t\sum_{(i,j)\in I\times I^c} X_iX_j}\right]} \qquad (37) \\
&\leqslant \sqrt{\mathbb{E}\left[e^{2tX}\right]}\sqrt{\mathbb{E}[e^{8tY_1Y_2}]} \qquad (38)
\end{aligned}
$$

where $Y_1 \sim \min(\mathrm{Bin}(|I|, p), \Delta)$, $Y_2 \sim \min(\mathrm{Bin}(|I^c|, p), \Delta)$ and Y_1 is independent of Y_2 (and (I, I^c) is some fixed, but unknown partition of $[m]$). (37) follows from the intermediate step (26) in the proof of Lemma B.2 (observing that $x \mapsto e^{tx}$ is convex, and non-decreasing as $t > 0$; and using the remark from that proof about the independence of X_i's not being required up to that step) and (38) follows from Lemma B.12. We will implicitly use Facts B.9, B.10, and B.11 for the remaining calculations, eventually replacing most expressions with $X' \sim \mathrm{Bin}(m, p)$.

Recalling that $X \leqslant X'$ by definition, the first term of (38) can be bounded as $\mathbb{E}[e^{2tX}] \leqslant e^{4tmp}$. Moreover, from our assumption, $tY_1 \leqslant t\Delta \leqslant 1/8$ and $tmp \leqslant 1/16$. Combined with the fact that Y_1, Y_2 is dominated by $X \sim \min(\mathrm{Bin}(m, p), \Delta)$ and thus by $X' \sim \mathrm{Bin}(m, p)$, we have

$$\mathbb{E}[e^{8tY_1Y_2}] = \mathbb{E}_{Y_1}[\mathbb{E}_{Y_2}[e^{8tY_1Y_2}]] \leqslant \mathbb{E}_{Y_1}[e^{16tY_1mp}] \leqslant e^{32tm^2p^2}.$$

Going back to (38), this implies

$$\mathbb{E}[\exp\left(tX^2\right)] \leqslant \sqrt{\exp\left(4tmp\right)}\sqrt{\exp\left(32tm^2p^2\right)} = \exp\left(2tmp + 16tm^2p^2\right),$$

concluding the proof. □

B.2 Stochastic dominance results between truncated Binomials

Fact B.9. *Let $X \sim \mathrm{Bin}(m,p)$, and $0 < n \leqslant m$. Defining $Y \coloneqq \min(X,n)$ and $Z \coloneqq X \mid X \leqslant n$, we have, for every $k \geqslant 0$,*

$$\Pr[X \geqslant k] \geqslant \Pr[Y \geqslant k] \geqslant \Pr[Z \geqslant k],$$

i.e., $X \succeq Y \succeq Z$, where $\succeq$ denotes first-order stochastic dominance.

Proof. We can write the PMF of Z and Y, for all $0 \leqslant k \leqslant n$,

$$\Pr[Y = k] = \begin{cases} \Pr[X = k], & k < n \\ \Pr[X \geqslant n], & k = n \end{cases}, \qquad \Pr[Z = k] = \frac{\Pr[X = k]}{\Pr[X \leqslant n]}.$$

It follows that $\Pr[Y \geqslant k] = \Pr[X \geqslant k]\mathbb{1}\{k \leqslant n\}$, which gives the first part of the statement.

The second part follows from a direct comparison between the two CDF of Z, Y: indeed, for $0 \leqslant k \leqslant n$,

$$\begin{aligned} \Pr[Y \geqslant k] \geqslant \Pr[Z \geqslant k] \quad &\Leftrightarrow \quad \Pr[X \geqslant k] \geqslant \frac{\Pr[n \geqslant X \geqslant k]}{\Pr[X \leqslant n]} \\ &\Leftrightarrow \quad \Pr[X \geqslant k](1 - \Pr[X > n]) \geqslant \Pr[X \geqslant k] - \Pr[X > n] \\ &\Leftrightarrow \quad \Pr[X \geqslant k]\Pr[X > n] \leqslant \Pr[X > n] \\ &\Leftrightarrow \quad \Pr[X \geqslant k] \leqslant 1\,, \end{aligned}$$

and this last inequality clearly holds. □

We also record the facts below, which follow respectively from the more general result that first-order stochastic dominance is preserved by non-decreasing mappings, and from a coupling argument.

Fact B.10. *Consider two real-valued random variables X, Y, and $n \geqslant 0$. If $X \succeq Y$, then $\min(X,n) \succeq \min(Y,n)$: for all k,*

$$\Pr[\min(X,n) \geqslant k] \geqslant \Pr[\min(Y,n) \geqslant k]\,;$$

i.e., the min *operator preserves first-order stochastic dominance relation.*

Fact B.11. *Let $X \sim \mathrm{Bin}(n,p)$ and $Y \sim \mathrm{Bin}(m,p)$, where $m \geqslant n$. Then $X \preceq Y$.*

Lemma B.12. *Let $X_1, \ldots, X_m$ be sampled from the sampling process in Definition B.8, and I, I^c be any partition of $[m]$. Define $Z_I \coloneqq \sum_{i \in I} X_i$, $Z_{I^c} \coloneqq \sum_{i \in I^c} X_i$, and $Y_I \sim \min(\mathrm{Bin}(|I|,p),n)$, $Y_{I^c} \sim \min(\mathrm{Bin}(|I^c|,p),n)$. Then*

$$Z_I \cdot Z_{I^c} \preceq Y_I \cdot Y_{I^c}\,.$$

Proof. We prove the lemma by defining a coupling $Z_I, Z_{I^c}, Y_I, Y_{I^c}$ such that $Z_I \cdot Z_{I^c} \leqslant Y_I \cdot Y_{I^c}$ with probability one. The sampling process below will generate samples $(X_i)_{1 \leqslant i \leqslant m}, Z_I, Z_{I^c}, Y_I, Y_{I^c}$ for all possible realizations of I and I^c. In other words, from a given sequence $\{X'_i\}_{i \in [m]}$, we will generate $\{X_i\}_{i \in [m]}, Y_{I_1}, Y_{I_1^c}, Y_{I_2}, Y_{I_2^c}, \ldots, Y_{I_{2^m}}, Y_{I_{2^m}^c}, Z_{I_1}, Z_{I_1^c}, Z_{I_2}, Z_{I_2^c}, \ldots, Z_{I_{2^m}}, Z_{I_{2^m}^c}$, where the (I_i, I_i^c) enumerate all partitions of $[m]$ in two sets.

1. Initialize $X_i = 0$ for all $i \in [m]$; sample $(X'_i)_{1 \leqslant i \leqslant m}$ as m i.i.d. $\mathrm{Bern}(p)$;
2. If $\sum_{i \in [m]} X'_i < n$, let $X_i = X'_i$ for all $i \in [m]$;
3. If $\sum_{i \in [m]} X'_i \geqslant n$, let $\mathcal{S}' = \{i \in [m] : X'_i = 1\}$ and let $\mathcal{S}$ be a uniformly random subset of $\mathcal{S}'$ with size n; set $X_i = X'_i$ for $i \in \mathcal{S}$.
4. For each $I \in \{I_1, \ldots, I_{2^m}\}$, denote $\mathcal{S}'_I = \mathcal{S}' \cap I$. Select a uniformly random subset of $\mathcal{S}'_I$ with at most n indices which is a superset of $\mathcal{S} \cap I$. In more detail, if $|\mathcal{S} \cap I| < n$, select $\min(|\mathcal{S}'_I|, n) - |\mathcal{S} \cap I|$ elements uniformly at random from $\mathcal{S}'_I \setminus (\mathcal{S} \cap I)$ to add to $\mathcal{S} \cap I$, which becomes $\mathcal{S}_I$; else, let $\mathcal{S}_I = \mathcal{S} \cap I$. Repeat a similar process for I^c to obtain $\mathcal{S}_{I^c}$.
5. For each $I \in \{I_1, \ldots, I_{2^m}\}$, set $Y_I = \sum_{i \in \mathcal{S}_I} X'_i$ and $Y_{I^c} = \sum_{i \in S_{I^c}} X'_i$.

From the above definition, we can readily see that for any I, $Y_I \geqslant Z_I$ and $Y_{I^c} \geqslant Z_{I^c}$. What is left is to argue that the $Y_I \sim \min(\mathrm{Bin}(|I|,p),n)$ and $Y_{I^c} \sim \min(\mathrm{Bin}(|I^c|,p),n)$. We start by noting that for any $k < n$, $\{Y_I = k\} = \{|\mathcal{S}_I| = k\} = \{|\mathcal{S}'_I| = k\}$. The last equality comes from the fact that $|\mathcal{S}_I| < n$ can only mean that $|\mathcal{S}'_I| < n$, and the selection process in step 4 will thus add all elements from $\mathcal{S}'_I$ to $\mathcal{S}_I$. From here, we have $\Pr[Y_I = k] = \Pr[|\mathcal{S}'_I| = k] = \Pr[\mathrm{Bin}(m,p) = k]$, for $k < n$; and we have $\Pr[Y_I = n] = 1 - \Pr[Y_I < n] = \Pr[\mathrm{Bin}(m,p) \geqslant n]$. As a result, $Y_I \sim \min(\mathrm{Bin}(|I|,p),n)$. Similarly, we can argue that $Y_{I^c} \sim \min(\mathrm{Bin}(|I^c|,p),n)$. □

C Deferred Proofs

C.1 Proofs from Section 3

Proof of Lemma 3.2. We analyze Algorithm 2: first we use the algorithm of Lemma 1.3 to learn P to d_{χ^2} distance ε^2 *as if* it was a product distribution, using $O\big(\frac{n}{\varepsilon^2}\big)$ samples. Let $\hat{P}$ be the output of the learning algorithm. Note that since Lemma 1.3 guarantees *proper* learning, $\hat{P}$ is a product distribution.

We then want to check that $\hat{P}$ is indeed close to P (in Hellinger distance), as it should if P were indeed a product distribution. To do this, we use the algorithm of Theorem 1.4 on P, with reference distribution $\hat{P}$ and distance parameter ε; and reject if, and only if, this algorithm rejects.

By a union bound, since both algorithms are correct with probability at least $5/6$, both are simultaneously correct with probability at least $2/3$; we hereafter assume this is the case in our analysis. The total sample complexity is $O\big(n/\varepsilon^2\big) + O\big(\sqrt{2^n}/\varepsilon^2\big) = O\big(2^{n/2}/\varepsilon^2\big)$, as claimed. We now argue correctness.

- *Soundness.* Proof by contrapositive: if the algorithm accepts, this means, from the guarantees of Theorem 1.4, that $d_{\mathrm{H}}(P, \hat{P}) \leqslant \varepsilon$. Since $\hat{P}$ is a product distribution, we conclude that P is ε-close (in Hellinger distance) to being a product distribution.
- *Correctness.* Assume that P is a product distribution. Then, Lemma 1.3 ensures that $d_{\chi^2}(P, \hat{P}) \leqslant \varepsilon^2$; and thus, by Theorem 1.4 the second step will not reject, and the algorithm overall accepts.

Note that, by a standard amplification trick (independent repetition and majority vote), the probability of error can be reduced from $1/3$ to any $\delta \in (0, 1)$ at the price of a $O(\log(1/\delta))$ factor in the number of samples. □

Proof of Theorem 3.1. We analyze Algorithm 1, denoting as in the algorithm by P' the product of marginals of P, and setting $\delta \coloneqq \frac{1}{3\binom{n}{d+1}}$, $\varepsilon' \coloneqq \frac{\varepsilon}{\sqrt{2n}(1+\sqrt{d+1})}$, and

$$m \coloneqq O\left(\frac{2^{d/2}}{\varepsilon'^2} \log \frac{1}{\delta}\right) = O\left(\frac{2^{d/2} n}{\varepsilon^2} \cdot d^2 \log n\right).$$

The sample complexity is thus immediate; further, note that, as stated in the algorithm, given m i.i.d. samples from P and a fixed set $T \subseteq [n]$ of nodes, one can generate m i.i.d. samples from P_T by only keeping the relevant variables (those in T) of each sample from P.

- *Completeness.* Assume P is a product distribution. Then, P_T is a product distribution for every choice of T, and each of the $\binom{n}{d+1}$ performs thus accepts with probability at least $1-\delta$ by Lemma 3.2. Thus, by a union bound, all tests simultaneously accept with probability at least $1 - \binom{n}{d+1} \cdot \delta = 2/3$, and Algorithm 1 returns "accept."
- *Soundness.* Assume now that P is ε-far (in total variation distance) from every product distribution over $\{0, 1\}^n$. A fortiori, it is ε-far from the product distribution P', and thus we have
$$d_{\mathrm{H}}(P, P') \geqslant \frac{1}{\sqrt{2}} d_{\mathrm{TV}}(P, P') \geqslant \frac{\varepsilon}{\sqrt{2}}$$
By Corollary 1.2, this means there exists some node $i \in [n]$ along with the set of its (at most) d parents Π_i such that, setting $T \coloneqq \{i\} \cup \Pi_i$,
$$d_{\mathrm{H}}^2(P_T, P'_T) \geqslant \frac{\varepsilon^2}{2n}\,.$$
Now, we can invoke our localization lemma, Lemma 3.3, to conclude that P_T is not only far from P'_T, it is far from *every* product distribution on T:
$$\min_{Q \text{ product}} d_{\mathrm{H}}^2(P_T, Q) \geqslant \frac{\varepsilon^2}{2n(1+\sqrt{d+1})^2} = \varepsilon'^2\,.$$

Thus, when this particular set T of $d+1$ nodes is encountered by the algorithm, the corresponding independence test will reject with probability at least $1-\delta$ by Lemma 3.2, and the overall algorithm will thus reject with probability at least $1-\delta$.

This concludes the proof: the sample complexity is $O(d^2 2^{d/2} n \log(n)/\varepsilon^2)$ as claimed, and the tester is correct in both cases with probability at least $2/3$. □

C.2 Proof of Lemma 3.3

Lemma 3.3. *Let P be a distribution on $\{0,1\}^n$, and $P' = (\pi_1 P) \otimes \cdots \otimes (\pi_n P)$ be the product of marginals of P. Denoting by $\mathcal{Q}$ the set of all product distributions on $\{0,1\}^n$, we have $\min_{Q\in\mathcal{Q}} d_{\mathrm{H}}(P,Q) \geqslant \frac{1}{1+\sqrt{n}} d_{\mathrm{H}}(P,P')$.*

Proof. Since squared Hellinger distance is an f-divergence, by the data processing inequality, we have that

$$d_{\mathrm{H}}(P,Q) \geqslant d_{\mathrm{H}}(\pi P, \pi Q). \tag{39}$$

By the subadditivity of Hellinger for Bayes nets from Corollary 1.2 along with (39), we obtain the following:

$$\begin{aligned}
d_{\mathrm{H}}(P,P') &\leqslant d_{\mathrm{H}}(P,Q) + d_{\mathrm{H}}(Q,P') \\
&\leqslant d_{\mathrm{H}}(P,Q) + \Big(\sum_{i=1}^n d_{\mathrm{H}}^2(P'_{X_i}, Q_{X_i})\Big)^{1/2} \\
&= d_{\mathrm{H}}(P,Q) + \Big(\sum_{i=1}^n d_{\mathrm{H}}^2(P_{X_i}, Q_{X_i})\Big)^{1/2} \\
&\leqslant d_{\mathrm{H}}(P,Q) + \sqrt{n d_{\mathrm{H}}^2(P,Q)} \\
&= \big(1+\sqrt{n}\big)\, d_{\mathrm{H}}(P,Q).
\end{aligned}$$

□

C.3 Proof of Lemma A.3

Lemma A.3. *There exist $C_1, C_2 > 0$ such that the following holds. Let $\varepsilon \in (0,1]$ and $n \geqslant C_1$, and let a, b be two integers such that $a+b=n$ and $b \geqslant \frac{1}{4}n$. Then, for $\delta := \frac{\varepsilon}{\sqrt{n}}$, we have*

$$\frac{(1-\delta)^n}{2^n} \sum_{k_1=0}^{a} \sum_{k_2=0}^{b} \binom{a}{k_1}\binom{b}{k_2} \left| \left(\left(\frac{1+\delta}{1-\delta}\right)^{k_1+k_2} - \left(\frac{1+\delta}{1-\delta}\right)^{k_1+b-k_2} \right) \right| \geqslant C_2 \varepsilon.$$

Proof. By concentration of Binomials,

$$\begin{aligned}
&\left(\frac{1}{2}\right)^n (1-\delta)^n \sum_{k_1=0}^{a} \sum_{k_2=0}^{b} \binom{a}{k_1}\binom{b}{k_2} \left| \left(\left(\frac{1+\delta}{1-\delta}\right)^{k_1+k_2} - \left(\frac{1+\delta}{1-\delta}\right)^{k_1+b-k_2} \right) \right| \\
\geqslant\ & \left(\frac{1}{2}\right)^n (1-\delta)^n \sum_{k_1=\frac{a}{2}+\sqrt{a}}^{\frac{a}{2}+2\sqrt{a}} \sum_{k_2=\frac{b}{2}+\sqrt{b}}^{\frac{b}{2}+2\sqrt{b}} \binom{a}{k_1}\binom{b}{k_2} \left| \left(\left(\frac{1+\delta}{1-\delta}\right)^{k_1+k_2} - \left(\frac{1+\delta}{1-\delta}\right)^{k_1+b-k_2} \right) \right| \\
\geqslant\ & \left(\frac{1}{2}\right)^n \frac{C}{\sqrt{ab}} \cdot 2^{|a|+|b|} \sum_{k_1=\frac{a}{2}+\sqrt{a}}^{\frac{a}{2}+2\sqrt{a}} \sum_{k_2=\frac{b}{2}+\sqrt{b}}^{\frac{b}{2}+2\sqrt{b}} (1-\delta)^n \left| \left(\left(\frac{1+\delta}{1-\delta}\right)^{k_1+k_2} - \left(\frac{1+\delta}{1-\delta}\right)^{k_1+b-k_2} \right) \right| \\
\geqslant\ & \frac{C}{\sqrt{ab}} \sum_{k_1=\frac{a}{2}+\sqrt{a}}^{\frac{a}{2}+2\sqrt{a}} \sum_{k_2=\frac{b}{2}+\sqrt{b}}^{\frac{b}{2}+2\sqrt{b}} (1-\delta)^n \left| \left(\left(\frac{1+\delta}{1-\delta}\right)^{k_1+k_2} - \left(\frac{1+\delta}{1-\delta}\right)^{k_1+k_2+b-2k_2} \right) \right|
\end{aligned}$$

where C is some constant larger than 0. For $k_1 + k_2 - \frac{n}{2} = l \in \left[\sqrt{a}+\sqrt{b}, 2\left(\sqrt{a}+\sqrt{b}\right)\right]$, $l \geqslant \sqrt{a+b} = \sqrt{n}$.

$$
\begin{aligned}
& (1-\delta)^n \left| \left(\left(\frac{1+\delta}{1-\delta}\right)^{k_1+k_2} - \left(\frac{1+\delta}{1-\delta}\right)^{k_1+k_2+b-2k_2} \right) \right| \\
= & \left| \left((1-\delta)^n \left(\frac{1+\delta}{1-\delta}\right)^{n/2+l} - (1-\delta)^n \left(\frac{1+\delta}{1-\delta}\right)^{n/2+l+b-2k_2} \right) \right| \\
= & (1-\delta^2)^{n/2} \left(\frac{1+\delta}{1-\delta}\right)^{l} \left| 1 - \left(\frac{1+\delta}{1-\delta}\right)^{b-2k_2} \right| \\
\geqslant & e^{-\delta^2 n} e^{2\delta \cdot l} \left| 1 - \left(\frac{1+\delta}{1-\delta}\right)^{b-2k_2} \right| \geqslant e^{2\varepsilon-\varepsilon^2} \left| 1 - \left(\frac{1+\delta}{1-\delta}\right)^{b-2k_2} \right| \qquad (40) \\
\geqslant & e^{\varepsilon} \left(1 - \left(\frac{1+\delta}{1-\delta}\right)^{b-2k_2} \right) \geqslant e^{\varepsilon}(1 - e^{4\delta(b-2k_2)}). \qquad (41)
\end{aligned}
$$

where (40) and (41) follows from $1 - x \geqslant e^{-2x}$, $0 < x < 0.79$ and $e^{4x} \geqslant \left(\frac{1+x}{1-x}\right) \geqslant e^{2x}$, $0 < x < 0.95$. All these inequalities hold for n larger some constant and every $\varepsilon \in (0, 1]$. Since $b \geqslant \frac{1}{4}n$, and by the summation above $b - 2k_2 \in \left[-4\sqrt{b}, -2\sqrt{b}\right]$,

$$
e^{\varepsilon}(1 - e^{4\delta(b-2k_2)}) \geqslant e^{\varepsilon}\left(1 - e^{-8\frac{\varepsilon}{\sqrt{n}}\sqrt{b}}\right) \geqslant e^{\varepsilon}(1 - e^{-4\varepsilon}) \geqslant \varepsilon.
$$

and therefore, summing up every term, we have our lower bound

$$
\frac{C}{\sqrt{ab}} \sum_{k_1=\frac{a}{2}+\sqrt{a}}^{\frac{a}{2}+2\sqrt{a}} \sum_{k_2=\frac{b}{2}+\sqrt{b}}^{\frac{b}{2}+2\sqrt{b}} (1-\delta)^n \left| \left(\left(\frac{1+\delta}{1-\delta}\right)^{k_1+k_2} - \left(\frac{1+\delta}{1-\delta}\right)^{k_1+k_2+b-2k_2} \right) \right| \geqslant C\varepsilon
$$

concluding the proof. □

C.4 Proof of (7)

Fact C.1. *For any set of cycles such that $\sum_i |\sigma_i| \leqslant n$, we have*

$$
\prod_{\substack{\sigma_i:\text{even} \\ |\sigma_i|\geqslant 4}} (1 + (-4\delta)^{|\sigma_i|}) \prod_{\substack{\sigma_i:\text{odd} \\ |\sigma_i|\geqslant 4}} (1 - (-4\delta)^{|\sigma_i|}) \leqslant e^{O\left(\varepsilon^5/n^{\frac{3}{2}}\right)} \prod_{\substack{\sigma_i:\text{even} \\ |\sigma_i|=4}} (1 + (4\delta)^4) \prod_{\substack{\sigma_i:\text{odd} \\ |\sigma_i|=4}} (1 - (4\delta)^4)
$$

Proof. We have

$$
\begin{aligned}
& \prod_{\sigma_i:\text{even}} (1 + (-4\delta)^{|\sigma_i|}) \prod_{\sigma_i:\text{odd}} (1 - (-4\delta)^{|\sigma_i|}) \\
= & \prod_{\substack{\sigma_i:\text{even} \\ |\sigma_i|\geqslant 5}} (1 + (-4\delta)^{|\sigma_i|}) \prod_{\substack{\sigma_i:\text{odd} \\ |\sigma_i|\geqslant 5}} (1 - (-4\delta)^{|\sigma_i|}) \prod_{\substack{\sigma_i:\text{even} \\ |\sigma_i|=4}} (1 + (-4\delta)^{|\sigma_i|}) \prod_{\substack{\sigma_i:\text{odd} \\ |\sigma_i|=4}} (1 - (-4\delta)^{|\sigma_i|}) \\
\leqslant & \prod_{\sigma_i:|\sigma_i|>5} (1 + (4\delta)^{|\sigma_i|}) \prod_{\sigma_i:\text{even},|\sigma_i|=4} (1 + (-4\delta)^{|\sigma_i|}) \prod_{\sigma_i:\text{odd},|\sigma_i|=4} (1 - (-4\delta)^{|\sigma_i|}) \\
\leqslant & (1 + (4\delta)^5)^{\frac{n}{4}} \prod_{\sigma_i:\text{even},|\sigma_i|=4} (1 + (-4\delta)^{|\sigma_i|}) \prod_{\sigma_i:\text{odd},|\sigma_i|=4} (1 - (-4\delta)^{|\sigma_i|}) \\
\leqslant & e^{(4\delta)^5 n/4} \prod_{\sigma_i:\text{even},|\sigma_i|=4} (1 + (-4\delta)^{|\sigma_i|}) \prod_{\sigma_i:\text{odd},|\sigma_i|=4} (1 - (-4\delta)^{|\sigma_i|}) \\
\leqslant & e^{256\varepsilon^5/n^{\frac{3}{2}}} \prod_{\sigma_i:\text{even},|\sigma_i|=4} (1 + (-4\delta)^4) \prod_{\sigma_i:\text{odd},|\sigma_i|=4} (1 - (-4\delta)^4)
\end{aligned}
$$

□

D Structured Testing Lower Bound

Letting $D = 2^n$, we will rely on the construction from the "standard" lower bound of Paninski [2008] by picking a uniformly random subset S of $\{0,1\}^n$ of size $\frac{D}{2}$. Denote $\mathcal{S}$ the set of all such combinations of S, and define $\mathcal{P}_{\text{no}}$ to be $\mathcal{P}_{\text{no}} := \left\{P = \frac{1+C\varepsilon}{2}U_S + \frac{1-C\varepsilon}{2}U_{S^c} \mid S \in \mathcal{S}\right\}$, where $C > 0$ is a suitable normalizing constant. As before, U_S denotes the uniform distribution on the set of variable S and $P \in \mathcal{P}_{\text{no}}$ is a mixture of two uniform distributions on disjoint parts, with different weights.

It is known that $\Omega(2^{n/2}/\varepsilon^2)$ samples are required to distinguish between such a randomly chosen P and the uniform distribution U; further, assume we know that the uniform distribution U is in $\mathcal{C}$. What remains to show is the *distance*, that is, "most" choices of $P \in \mathcal{P}_{\text{no}}$ are ε-far from $\mathcal{C}$. To argue that last part, we will use our assumption that $\mathcal{C}$ can be learned with m samples to conclude by a counting argument: i.e., we will show that there can be at most 2^{mn} or so "relevant" elements of $\mathcal{C}$, while there are at least $2^{2^{\Omega(n)}}$ $\mathcal{P}_{\text{no}}$ that are ε-far from each other. Suitably combining the two will establish the theorem below:

Theorem D.1. *Let $\mathcal{C}$ be a class of probability distributions over $\{0,1\}^n$ such that the following holds: (1) the uniform distribution belongs to $\mathcal{C}$ (2) there exists a learning algorithm for $\mathcal{C}$ with sample complexity $m = m(n,\varepsilon)$. Then, as long as $mn \ll 2^{O(n)}$, testing whether an arbitrary distribution over $\{0,1\}^n$ belongs to $\mathcal{C}$ or is ε-far from every distribution in $\mathcal{C}$ in total variation distance requires $\Omega(2^{n/2}/\varepsilon^2)$ samples.*

Proof. As discussed above, indistinguishability follows from the literature [Paninski, 2008], and thus all we need to show now is that $\mathcal{P}_{\text{no}}$ is far from every distribution in $\mathcal{C}$. By assumption (2), there exists an algorithm $H\colon \{0,1\}^{mn} \to \mathcal{P}$ (without loss of generality, we assume H deterministic) that can output an estimated distribution given $m = m(n,\varepsilon)$ samples from $P \in \mathcal{C}$. Thus, for every $P \in \mathcal{C}$ given m *i.i.d.* samples $X \in \{0,1\}^{mn}$, $\Pr_{X\sim P^{\otimes m}}(d_{\text{TV}}(H(X),P) < \varepsilon) \geqslant 2/3$.

In particular, this implies the weaker statement that, for every $P \in \mathcal{C}$, there exists *some* x in $\{0,1\}^{mn}$ s.t. $P \in B(H(x),\varepsilon)$ (where $B(x,r)$ denotes the TV distance ball of radius r centered at x). By enumerating all possible values in $\{0,1\}^{mn}$, we then can obtain an ε-cover $\{H(x_1),\dots,H(x_{2^{mn}})\}$ of $\mathcal{C}$, that is, such that $\mathcal{C} \subseteq \bigcup_{i=1}^{2^{mn}} B(H(x_i),\varepsilon)$. The ε-covering number of $\mathcal{C}$ is thus upper bounded by $2^{O(mn)}$.

Next, we lower bound the size of $\mathcal{P}_{\text{no}}$ by constructing an ε-packing P_ε, where $P_\varepsilon = \{P_i \in \mathcal{P}_{\text{no}}, i \in \mathbb{N} : d_{\text{TV}}(P_i,P_j) > \varepsilon, i \neq j\}$. For $P,Q \in \mathcal{P}_{\text{no}}$ corresponding to two sets S,S', each of size $\frac{D}{2} = 2^{n-1}$, we have

$$d_{\text{TV}}(P,Q) \quad = \quad \frac{1}{2}|S\triangle S'| \cdot \left|\frac{1+C\varepsilon}{2} - \frac{1-C\varepsilon}{2}\right| \cdot \frac{2}{D} = C\varepsilon \cdot \frac{|S\triangle S'|}{D} > \varepsilon$$

For this to be at least ε, the pairrwise symmetric difference of (the sets corresponding to the) distributions in P_ε should be at least $\frac{D}{C} = \Omega(2^n)$. We know, by, e.g., Blais et al. [2019, Proposition 3.3] that there exist such families of balanced subsets of $\{0,1\}^n$ of cardinality at least $\Omega(2^{2^{\rho n}})$, where $\rho > 0$ is a constant that only depends on C.

Thus, the size of $\mathcal{P}_{\text{no}}$ is itself $\Omega(2^{2^{\rho n}})$; combining this lower bound with the upper bound on the covering number of $\mathcal{C}$ concludes the proof. □

As a corollary, instantiating the above to the class $\mathcal{C}$ of degree-d Bayes nets over n nodes readily yields the following:

Corollary D.2. *For large enough n, testing whether an arbitrary distribution over $\{0,1\}^n$ is a degree-d Bayes net or is ε-far from every such Bayes net requires $\Omega(2^{n/2}/\varepsilon^2)$ samples, for any $d = o(n)$ and $\varepsilon \geqslant 2^{-O(n)}$.*

Proof. We can obtain a learning upper bound of $m = O(2^d n \log(2^{d+1}n)\log(n^{dn})/\varepsilon^2)$ for degree-d Bayes nets by combining the known-structure case (proven in Bhattacharyya et al. [2020]) with the reduction from known-structure to unknown-structure (via hypothesis selection/tournament [Canonne

et al., 2020]). We have $mn \leqslant O(2^d n^6/\varepsilon^2)$. To have $2^{mn} \ll 2^{2^{\rho n}}$, where ρ is some constant, we need $mn < 2^{O(n)}$, which requires $d = o(n)$ and $\varepsilon \geqslant 2^{-O(n)}$ for large enough n. □

D.1 An $\Omega(2^{d/2}\sqrt{n}/\varepsilon^2)$ Lower Bound

In this section, we state and prove a simpler, but quantitatively weaker lower bound than Theorem 2.1 for independence testing, Theorem D.3. This simpler lower bound is adapted from Canonne et al. [2020, Theorem 13] – the "mixture-of-products" construction. Their analysis readily provides indistinguishability, and distance *from the uniform distribution*. Thus, all we need here is to show that most of these hard instances (i.e., "mixtures of products") are far from every product distribution (Lemma D.4), not just the uniform distribution. While the $\Omega(2^{d/2}\sqrt{n}/\varepsilon^2)$ lower bound this yields is not as tight in terms of sample complexity, with a $\sqrt{n}$ dependence instead of n (at a high level, this is because we fix the Bayesian structure, and thus the algorithms have additional information they can leverage), the restriction on d is much milder than the one in Theorem 2.1, allowing up to $d = n/2$.

Theorem D.3. *Let $1 \leqslant d \leqslant n/2$. Testing whether an arbitrary degree-d Bayes net over $\{0,1\}^n$ is a product distribution or is ε-far from every product distribution requires $\Omega(2^{d/2}\sqrt{n}/\varepsilon^2)$ samples. This holds even if the structure of the degree-d Bayes net is known.*

Proof. As discussed above, we will use the same "mixture-of-products" construction as in Canonne et al. [2020, Theorem 13], which established a lower bound of $\Omega(2^{d/2}\sqrt{n}/\varepsilon^2)$ samples to distinguish it from the uniform distribution. We first recall the definition of this "mixture-of-products" construction.

Letting $N \coloneqq n - d \geqslant n/2$, we define, for $z \in \{\pm 1\}^N$ the product distribution p_Z over $\{0,1\}^N$ by

$$p_z(x) = \prod_{i=1}^{N} \left(\frac{1}{2} + z_i (-1)^{x_i} \delta \right), \qquad x \in \{0,1\}^N. \tag{42}$$

where $\delta \coloneqq \frac{\varepsilon}{\sqrt{N}} = \Theta\left(\frac{\varepsilon}{\sqrt{n}}\right)$. A mixture-of-products distribution is then defined by choosing 2^d i.i.d. $Z_1, \ldots, Z_{2^d} \in \{\pm 1\}^N$ uniformly at random, and setting $\mathbf{p}_{Z_1,\ldots,Z_{2^d}}$ to be the distribution over $\{0,1\}^n$ which is uniform on the first d bits, and where the first d bits of x are seen as the binary representation (i.e., a "pointer") for which $\mathbf{p}_{Z_i}$ will be used for the last N bits of x. That is,

$$\mathbf{p}_{Z_1,\ldots,Z_{2^d}}(x) = \frac{1}{2^d} p_{Z_{\iota(x_1,\ldots,x_d)}}(x_{d+1}, \ldots, x_n), \qquad x \in \{0,1\}^n \tag{43}$$

where, analogously to Definition 2.2, $\iota\colon \{0,1\}^d \to [2^d]$ is the indexing function, mapping the binary representation (here on d bits) to the corresponding number.

As mentioned in the preceding discussion, this construction was already used in Canonne et al. [2020, Theorem 13], where the authors show an $\Omega(2^{d/2}\sqrt{n}/\varepsilon^2)$ sample complexity lower bound to distinguish a uniformly randomly chosen mixture-of-products distribution (which is a degree-d Bayes net) from the uniform distribution (which is a product distribution). For their theorem (a lower bound on testing *uniformity*), they then conclude from the easy fact that every such mixture-of-products distribution is ε-far from the uniform distribution. This is not enough for us, as, to obtain the lower bound stated in Theorem D.3, what we need is to show that every such mixture-of-products distribution (or at least *most* of them) is far from *every* product distribution, not just the uniform one. This is the only missing part towards proving Theorem D.3, and is established in our next lemma:

Lemma D.4 (Distance from Product distributions)**.** *For p uniformly sampled from the mixture-of-products construction,*

$$\Pr\left[\min_{q_1,q_2} d_{\mathrm{TV}}(p, q_1 \otimes q_2) \geqslant \frac{\varepsilon}{750}\right] \geqslant \frac{9}{10}$$

as long as $n \geqslant d + C_1$, for some constants $C_1 > 0$ and $n/2 \geqslant d$.

This lemma will directly follow from Claim D.5 (below) and Lemma 1.5; the rest of this appendix is thus dedicated to proving the former, which states that most mixture-of-products distributions are far from the product of their marginals.

Claim D.5. *Given a mixture-of-products distribution p as in* (43)*, let p_1 be the marginal of $\mathbf{p}$ on the first d variables (parent nodes) and p_2 the marginal on the N last variables (child nodes). Note that $p_1 \otimes p_2$ is then a product distribution on $\{0,1\}^n$. Then, we have*

$$\Pr\left[d_{\mathrm{TV}}(p, p_1 \otimes p_2) \geqslant \frac{\varepsilon}{250}\right] \geqslant \frac{9}{10}$$

as long as $n \geqslant C_1 + d$, for some constants $C_1 \geqslant 0$ and $n \geqslant 2d$.

Proof. Fix any mixture-of-products distribution p. From Lemma A.2 and the structure of p as given in (43), one can show that

$$d_{\mathrm{TV}}(p, p_1 \otimes p_2) \geqslant \frac{1}{2^{d-1}} \sum_{x_2,\ldots,x_d} d_{\mathrm{TV}}(p(\cdot \mid 0, x_2, \ldots, x_d), p(\cdot \mid 1, x_2, \ldots, x_d)).$$

Denoting $p(\cdot \mid 0, x_2, \ldots, x_d)$ by $p_{\iota(x_2,\ldots,x_d)}$ and $p(\cdot \mid 0, x_2, \ldots, x_d)$ by $q_{\iota(x_2,\ldots,x_d)}$ (where $\iota(x_2, \ldots, x_d) \in [2^{d-1}]$, abusing slightly the definition of the indexing function to extend it to $d-1$ bits), we can rewrite this as

$$d_{\mathrm{TV}}(p, p_1 \otimes p_2) \geqslant \frac{1}{2^{d-1}} \sum_{t=1}^{2^{d-1}} d_{\mathrm{TV}}(p_t, q_t) =: \overline{d_{\mathrm{TV}}}\,.$$

Now, since $d \leqslant n/2$, one can show that, for every fixed t,

$$\Pr[d_{\mathrm{TV}}(p_t, q_t) < \varepsilon/25] < e^{-C \cdot N} \tag{44}$$

where the probability is taken over the choice of p (i.e., its 2^d parameters $Z_1, \ldots, Z_{2^d}$), and $C > 0$ is an absolute constant. We defer the proof of this inequality to the end of the appendix, and for now observe that the RHS is less than $1/10$ for N greater than some (related) absolute constant $C_1 > 0$. We can then write, letting $D := 2^{d-1}$ and $X_t := \mathbb{1}\{d_{\mathrm{TV}}(p_t, q_t)\}$

$$\overline{d_{\mathrm{TV}}} \geqslant \frac{\varepsilon}{25} \cdot \frac{1}{2^{d-1}} \sum_{t=1}^{2^{d-1}} \mathbb{1}\left\{d_{\mathrm{TV}}(p_t, q_t) \geqslant \frac{\varepsilon}{25}\right\} = \frac{\varepsilon}{25} \cdot \frac{1}{D} \sum_{t=1}^{D} X_t,$$

where the X_t's are i.i.d. Bernoullis with, by the above analysis, parameter $\alpha \geqslant 1 - e^{-C \cdot N} \geqslant 9/10$. We then have

$$\Pr[d_{\mathrm{TV}}(p, p_1 \otimes p_2) < \varepsilon/250] \leqslant \Pr[\overline{d_{\mathrm{TV}}} < \varepsilon/250] \leqslant \Pr\left[\sum_{t=1}^{D} X_t < \frac{D}{10}\right],$$

so it remains to show that the RHS is less than $1/10$. Since $\mathbb{E}[X_t] \geqslant 9/10$ for all t, this readily follows from a Hoeffding bound, for $d \geqslant 1$. □

To conclude, we only need to prove (44), which (slightly rephrasing it) tells us that two independent parameterizations p_Z, p'_Z will be at total variation distance at least $\Omega(\varepsilon)$ far with overwhelming probability.

Proof of (44). Let distribution $p_Z, p_{Z'}$ be defined as in (42), and Z, Z' be i.i.d. and uniform on $\{\pm 1\}^N$. The statement to show is then

$$\Pr\left[d_{\mathrm{TV}}(p_Z, p_{Z'}) \geqslant \frac{\varepsilon}{25}\right] \geqslant 1 - e^{-N/18}, \tag{45}$$

We know (see, e.g., Kamath et al. [2019, Lemma 6.4]) that as long as $\delta \leqslant 1/6$ (which holds for $n \geqslant 36$), then the TV distance is related to the ℓ_2 distance between mean vectors $\mu_Z, \mu_{Z'} \in [0,1]^N$ as

$$d_{\mathrm{TV}}(p_Z, p_{Z'}) \geqslant \frac{1}{20} \sqrt{\sum_{i=1}^{N} (\mu_{Z,i} - \mu_{Z',i})^2}. \tag{46}$$

Relating this ℓ_2 distance between mean vectors the Hamming distance between Z and Z', we have

$$\sqrt{\sum_{i=1}^{N}(\mu_{Z,i}-\mu_{Z',i})^2} = \sqrt{\sum_{i=1}^{N}\left(\left(\frac{1}{2}-Z_i\delta\right)-\left(\frac{1}{2}-Z_i'\delta\right)\right)^2} = 2\delta\sqrt{\mathrm{Hamming}(Z,Z')}, \quad (47)$$

where $\mathrm{Hamming}(Z,Z') = \sum_{i=1}^{N}\mathbb{1}[Z_i \neq Z_i']$. Noting that Hamming$(Z,Z') \sim \mathrm{Bin}(N,1/2)$, we have the following via Hoeffding's inequality along with (46) and (47),

$$\Pr\left[d_{\mathrm{TV}}(p_Z,p_{Z'}) \geqslant \frac{1}{20}\sqrt{\frac{4\delta^2 N}{3}}\right] \geqslant \Pr[\mathrm{Hamming}(Z,Z') \geqslant N/3] \geqslant 1-e^{-N/18}.$$

Since $n/2 \geqslant d$ and thus, $\sqrt{\frac{4\delta^2 N}{3}} \geqslant \frac{4}{5}\varepsilon$, we get (45). □

This concludes the proof of Lemma D.4, and with it of Theorem D.3. □