

# Near-Optimal Degree Testing for Bayes Nets

Vipul Arora<sup>\*1</sup>, Arnab Bhattacharyya<sup>†1</sup>, Clément L. Canonne<sup>2</sup>, and Joy Qiping Yang<sup>‡1,2</sup>

<sup>1</sup>National University of Singapore. {vipul, arnab, jyang}@comp.nus.edu.sg.

<sup>2</sup>University of Sydney. clement.canonne@sydney.edu.au, qyan6238@uni.sydney.edu.au

## Abstract

This paper considers the problem of testing the maximum in-degree of the Bayes net underlying an unknown probability distribution  $P$  over  $\{0, 1\}^n$ , given sample access to  $P$ . We show that the sample complexity of the problem is  $\tilde{\Theta}(2^{n/2}/\varepsilon^2)$ . Our algorithm relies on a testing-by-learning framework, previously used to obtain sample-optimal testers; in order to apply this framework, we develop new algorithms for “near-proper” learning of Bayes nets, and high-probability learning under  $\chi^2$  divergence, which are of independent interest.

## 1 Introduction

One of the most natural and widely-used ways to model high-dimensional distributions is as *Bayesian networks* (or, *Bayes nets* for short) [Pea88]. In particular, 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  $\pi$  to the parents of the  $i$ 'th node in the graph  $G$ . To generate an  $n$ -dimensional sample, one samples the nodes in a topological order of  $G$ , where the  $i$ 'th node is sampled according to  $p_{i,\pi}$  for the assignment  $\pi$  that is already fixed by the samples for the parent nodes of  $i$ . Bayes nets naturally encode causal information [Pea95], and learning a Bayes net description of a probability distribution is considered a fundamental problem in statistics and machine learning [Hec98]. For instance, Sachs et al. [SPP<sup>+</sup>05] used this approach to discover protein regulatory networks from gene expression data.

Sufficiently complex Bayes nets can encode arbitrary distributions, and so, it is infeasible in general to learn general Bayes nets. Instead, one often restricts to Bayes nets whose underlying DAGs have bounded *in-degree*. Distributions having sparse Bayes net descriptions naturally arise in machine learning, robotics, natural language processing [WJ08], medicine, and computational biology [FLNP00]. Moreover, information-theoretically, it is known that Bayes nets with in-degree bounded by  $d$  can be learned up to total variation (TV) distance  $\varepsilon$  using  $\tilde{O}(n2^d/\varepsilon^2)$  samples [CDKS17]. Hence, the maximum in-degree of a Bayes net is an important modeling parameter to consider.

In this work, we consider the problem of testing whether a distribution belongs to the concept class of degree- $d$  Bayes nets, i.e., those whose in-degree is at most  $d$ . Specifically, given sample access to a distribution  $P$  on  $\{0, 1\}^n$ , we consider the property testing question: is  $P$  described by a degree- $d$  Bayes net, or is  $P$   $\varepsilon$ -far from all such Bayes nets, in TV distance?

**THEOREM 1.1.** (MAIN THEOREM, INFORMAL, SEE **THEOREM 5.1**) *Given an unknown distribution  $P$  on  $\{0, 1\}^n$  and a degree parameter  $d \ll n$ , testing whether  $P$  is Markov with respect to any degree- $d$  Bayes net has sample complexity  $\tilde{\Theta}\left(\frac{2^{n/2}}{\varepsilon^2}\right)$ .*

<sup>\*</sup>Vipul was supported in part by NRF-AI Fellowship R-252-100-B13-281.

<sup>†</sup>A. Bhattacharyya was supported in part by an MOE Tier II award, and an Amazon faculty research award.

<sup>‡</sup>Joy was supported in part by NRF-AI Fellowship R-252-100-B13-281.

Our result requires that  $d < n/2 - \Omega(\log n)$ , which is consistent with our motivation of testing Bayes net sparsity. The main contribution of this work lies in establishing the upper bound of the theorem; we note that the (nearly) matching lower bound follows from [BCY22, Corollary B.2].

Over the course of deriving this upper bound, we obtain three new learning results which we believe to be of independent interest. (1) First, a *high-probability learning* result in  $\chi^2$  divergence (Proposition 4.1). While the sample complexity of learning an arbitrary distribution  $P$  over a discrete domain  $\Sigma$  to  $\chi^2$  divergence  $\varepsilon$  is known to be  $\Theta(|\Sigma|/\varepsilon)$  for *constant* probability of success,<sup>1</sup> boosting this success probability to  $1 - \delta$  for arbitrarily small  $\delta > 0$  was until now open. In particular, whether a *logarithmic* dependence on  $\delta$  was possible, as for total variation distance and (as recently shown) KL divergence, was unknown. We show this is indeed the case:  $O\left(\frac{|\Sigma|}{\varepsilon} \log \frac{|\Sigma|}{\delta}\right)$  i.i.d. samples suffice for high-probability learning in  $\chi^2$  divergence. Interestingly, the estimator achieving this bound is neither the empirical estimator, nor the usual Laplace add-1 estimator, but instead an add- $K$  estimator for a suitable  $K = K(\delta)$ .

(2) Second, a *near-proper* learning algorithm in  $\chi^2$ -divergence for degree- $d$  Bayes nets, with sample complexity  $\tilde{O}(2^d n^2/\varepsilon)$  (Theorem 4.1). We note that previous learning algorithms for Bayes nets either learn with respect to a *weaker distance measure* (TV or KL) or are *non-proper*, in the sense that the hypothesis they output is not a degree- $d$  Bayes net itself. In comparison, our algorithm outputs a *bona fide* degree- $d$  Bayes net  $\hat{P}$ , along with a subset  $S \subseteq \{0, 1\}^n$  of the domain such that (i)  $P, \hat{P}$  put all but  $O(\varepsilon)$  probability mass on  $S$ , and (ii)  $P$ , and  $\hat{P}$  are within  $\chi^2$  divergence  $\varepsilon$  when *restricted to this subset*  $S$ .

This hybrid guarantee, which makes our learning algorithm *near-proper* instead of proper, may seem artificial. However, our third result shows that it is indeed necessary, in a very strong sense:

(3) We prove in Proposition 4.2, a lower bound on the sample complexity of proper learning in  $\chi^2$  divergence, showing that *any* learning algorithm whose  $\chi^2$  guarantees holds on the whole domain must have sample complexity  $\Omega(2^{n/2}/\varepsilon)$ , even for degree-1 Bayes nets.

## 2 Related work

While learning graphical models in a range of settings and under various distance measures has a rich history, both in Statistics and Machine Learning, the corresponding task of *testing* properties of an unknown distribution represented as a (succinct) graphical model has only been considered much more recently. [DDK19] initiated the question of testing identity (goodness-of-fit) and independence of high-dimensional distributions with dependency structure modeled as an undirected graph (i.e., Ising models or, more generally, Markov Random Fields); this line of work was then continued in, e.g., [DDK17, GNS18, NL19, BBC<sup>+</sup>20], leading to a range of positive (algorithms), and negative (lower bounds) results. Very recently, [CDDK22] introduced the question of goodness-of-fit testing for latent Ising models, where only the leaf nodes of the tree are observable.

Focusing on another widely-studied type of graphical models, the concurrent works [CDKS17], and [DP17] studied analogous testing questions for Bayesian networks, where the underlying dependency structure is modeled as a directed graph. [ABDK18] focused on the related tasks for *causal* Bayesian networks, given the ability to perform interventions on the network. Finally, closest to our own work, [BCY22] studies the question of *independence testing* for Bayesian networks, obtaining near-optimal sample complexity bounds for the task of deciding if a given high-dimensional distribution, promised to be a sparse Bayesian network, is in fact a product distribution.

The other contribution of our work, high-probability learning in  $\chi^2$  divergence for arbitrary discrete

---

<sup>1</sup>i.e., outputting  $\hat{P}$  such that  $d_{\chi^2}(P, \hat{P}) \leq \varepsilon$  with probability at least 9/10.

distributions, follows a long line of results related to density estimation under various distance measures (see, e.g., [KOPS15, DL01, Dia16], as well as [Can20], and references within). While learning arbitrary distributions under  $\chi^2$  distance *with constant success probability* has long been well understood, even to the optimal leading constant in the minimax estimation rate [KOPS15], obtaining high-probability bounds has proven quite elusive. Even for the weaker notion of learning under Kullback–Leibler (KL) divergence, such a high-probability learning result was only obtained very recently [BGPV21].

Finally, to the best of our knowledge no result was known for learning Bayesian networks under  $\chi^2$  divergence, even for constant success probability, besides the trivial bound one gets by treating the Bayesian network as an unstructured probability distribution over  $\Sigma^n$ . This is again to contrast with the case of KL divergence, for which optimal constant-probability learning bounds, *and* high-probability learning bounds are known (cf. [BGPV21], and references within).

### 3 Overview

Our algorithm follows the testing-by-learning framework developed in [ADK15], and since used in several works (e.g., [DKW18, CDKL22]). Specifically, if the property one wants to test in TV distance is relatively easy to learn in a “harder” notion of distance, e.g.,  $\chi^2$ , then it is possible to build an efficient, and possibly *sample-optimal* tester by first learning the distribution in  $\chi^2$ , *assuming* that it has the given property, and then use a “ $\chi^2$ -TV tolerant tester” [ADK15] to test whether the hypothesis output by the learning algorithm is close to the actual distribution. The key is that these tolerant testers must not only reject distributions far in TV, but also accept those sufficiently close in  $\chi^2$  (“tolerance”): now, if the unknown distribution  $P$  does have the property, then the learning algorithm works as intended, and its output  $\hat{P}$  is close to  $P$ : so, the tester will accept. However, if  $P$  is far from the property, then either the learning algorithm fails and  $\hat{P}$  is far from  $P$  (and the tester rejects); or it still succeeds, but by the triangle inequality  $\hat{P}$  must be itself far from the property (and this can be checked and detected, as we now have an explicit description of  $\hat{P}$ ).

The key here is that the sample complexity of this “ $\chi^2$ -tolerant” tester is  $O(\sqrt{|\Sigma|}/\varepsilon^2)$  for distributions over domain  $\Sigma$  – which is optimal (up to constants) for many testing tasks, and matches the sample complexity of the “non-tolerant” testers. Thus, as long as the learning stage can be done with much fewer than  $O(\sqrt{|\Sigma|}/\varepsilon^2)$  samples, the overall approach yields a sample-efficient testing algorithm. (Moreover, this approach can be extended in many ways, e.g., for testing in Hellinger distance instead of TV [DKW18, BCY22].)

While the testing-by-learning idea seems relatively straightforward, the main technical contribution of this paper is in obtaining the required learning algorithm to apply it: namely, an efficient learning algorithm with high probability, and proper learning of Bayes nets (both in  $\chi^2$ ). Indeed, and quite surprisingly, before our work it was still unclear whether one could learn a discrete distribution in  $\chi^2$  divergence with failure probability at most  $\delta$  by paying only an  $O(\log(1/\delta))$  dependence in the sample complexity. In particular, the “obvious” approaches based on applying either McDiarmid’s inequality, or some sort of “median trick” fail for  $\chi^2$  divergence, due respectively to the high sensitivity of the estimators, and the failure of the triangle inequality. We remark that achieving an *exponentially* worse  $O(1/\delta)$  dependence is straightforward via Markov’s inequality [KOPS15]; however, this cost becomes impractical in settings when one requires an exponentially small failure probability, such as ours – as we need  $2^d \cdot n$  conditional distributions to be learned well simultaneously, and thus need to do a union bound over these many runs of a learning algorithm.

Interestingly, learning with high probability with only a  $O(\log(1/\delta))$  dependence in the sample complexity, and learning a Bayes net with  $\tilde{O}(2^d \cdot n/\varepsilon)$  are both achievable for the relatively easier  $d_{\text{KL}}$  divergence (and even then, the high-probability result was only established recently [BGPV21]). Unfortunately, learning in KL is a much weaker guarantee, and trying to instantiate the aforementioned

“testing-by-learning” framework with KL-TV instead of  $\chi^2$ -TV would only yield a much looser  $\tilde{O}(2^n/(n \cdot \varepsilon^2))$  testing upper bound.

Below, we give a series of results on learning with respect to  $\chi^2$  divergence: (1) high-probability learning with  $O(|\Sigma| \cdot \log(1/\delta)/\varepsilon)$  sample; (2) a near-proper Bayes net learning algorithm with sample complexity  $\tilde{O}(2^d n^2/\varepsilon)$ ; and (3) an exponential sample complexity lower bound of  $\Omega(2^{n/2}/\varepsilon)$  for learning Bayes nets. Later in [Section 5](#), we will build on the second result as the first step of our main result, the maximum in-degree testing algorithm of [Theorem 1.1](#).

**Preliminaries.** We use the standard asymptotic notations  $O(\cdot)$ ,  $\Omega(\cdot)$ ,  $o(\cdot)$ ,  $\Theta(\cdot)$ , and the (semi)-standard  $\tilde{O}(\cdot)$ ,  $\tilde{\Omega}(\cdot)$ , and  $\tilde{\Theta}(\cdot)$  to hide polylogarithmic factors in the argument. Since we are focusing entirely on the in-degree of Bayes nets, all references to *degree* will be implicitly in-degree unless stated otherwise. We use  $A \lesssim B$  to indicate that  $A \leq C \cdot B$  for some absolute constant  $C$ .

For a multivariate random variable  $X$  supported on  $\{0, 1\}^n$ , we use  $X_i$  to denote its  $i$ -th component (coordinate); for a Bayes net, we will use  $\Pi_i$  to denote the set of parents of  $X_i$ .

## 4 Learning in $\chi^2$

**4.1 High probability learning in  $\chi^2$**  Our analysis will largely follow the analysis of [[BGPV21](#), [Theorem 6.1](#), and [Claim 4.4](#)] for KL divergence, with a few crucial differences which allow us to extend it to  $\chi^2$  divergence, and to obtain a strictly better sample complexity than prior work. Specifically, we go beyond the standard add-1 Laplace estimator, and instead analyze the more general add- $K$  estimator for a suitable, *non-constant* value of  $K$ . Recall that the add- $K$  estimator, given  $N$  i.i.d. samples from some probability distribution over a domain  $\Sigma$  (with empirical counts  $N_1, \dots, N_{|\Sigma|}$ ), is defined by

$$\hat{P}(i) = \frac{N_i + K}{N + K|\Sigma|}, \quad i \in \Sigma. \quad (4.1)$$

Intuitively, the parameter  $K$  controls the amount of smoothing for our estimator. With  $\chi^2$  divergence being a very stringent notion of distance (much more so than KL divergence, let alone TV distance), the idea to achieve high-probability learning guarantees is to increase the smoothing in order to counteract the risk of “low-probability but catastrophic” events which could make the  $\chi^2$  divergence blow up. We are then able to show that setting  $K = \Theta(\log(1/\delta))$  achieves the desired sample complexity in the high-probability regime, much better than the Laplace estimator (which corresponds to  $K = 1$ ).

**PROPOSITION 4.1.** *Fix any  $\delta \in (0, 1]$ , and let  $K = \Theta(\log(1/\delta))$ . Given  $N$  i.i.d. samples from an unknown probability distribution  $P$  over an alphabet  $\Sigma$ , with probability at least  $1 - \delta$ , the add- $K$  estimator yields a hypothesis  $\hat{P}$  such that*

$$d_{\chi^2}(P, \hat{P}) \lesssim \frac{|\Sigma| \log(|\Sigma|/\delta)}{N}.$$

*In particular,  $N = O\left(\frac{|\Sigma|}{\varepsilon} \log \frac{|\Sigma|}{\delta}\right)$  samples suffice to learn  $P$  to  $\chi^2$  divergence  $\varepsilon$  with probability  $1 - \delta$ .*

In contrast, using a similar analysis for the Laplace estimator would only yield a worse sample complexity of  $O((|\Sigma|/\varepsilon) \log^2(|\Sigma|/\delta))$ , off by a logarithmic factor in  $\frac{|\Sigma|}{\delta}$ . We will use this result in [Section 4.2](#).

*Proof.* [Analysis of the add- $K$  estimator] Let  $Q$  be the output of the add- $K$  estimator with  $N$  i.i.d. samples from  $P$ . By the proof of [[BGPV21](#), [Claim 4.4](#)], we have the following for each  $i$ , and  $T_i$  being the count of the element  $i$  from  $N$  samples, with probability at least  $1 - \delta$ ,

$$\left| P_i - \frac{T_i}{N} \right| \leq \sqrt{\frac{3P_i \log\left(\frac{2}{\delta}\right)}{N}} + \frac{3 \log\left(\frac{2}{\delta}\right)}{N}.$$

In the following, we will use  $A \lesssim B$  to indicate that  $A \leq C \cdot B$  for some absolute constant  $C$ .

$$\begin{aligned}
|P_i - Q_i| &\leq \left| P_i - \frac{T_i}{N} \right| + \left| \frac{T_i}{N} - Q_i \right| \\
&= \left| P_i - \frac{T_i}{N} \right| + \left| \frac{T_i}{N} - \frac{T_i + K}{N + K|\Sigma|} \right| \\
&= \left| P_i - \frac{T_i}{N} \right| + \left| \frac{T_i|\Sigma|/N - 1}{N + K|\Sigma|} \right| K \\
&\leq \sqrt{\frac{3P_i \log(\frac{2}{\delta})}{N}} + \frac{3 \log(\frac{2}{\delta})}{N} + \frac{T_i|\Sigma|/N}{N + K|\Sigma|} K + \frac{K}{N + K|\Sigma|} \\
&\leq \sqrt{\frac{3P_i \log(\frac{2}{\delta})}{N}} + \frac{3 \log(\frac{2}{\delta})}{N} + \frac{K|\Sigma|}{N + K|\Sigma|} \frac{T_i}{N} + \frac{K}{N} \\
&\leq \sqrt{\frac{3P_i \log(\frac{2}{\delta})}{N}} + \frac{3 \log(\frac{2}{\delta})}{N} + \frac{K}{N} \\
&\quad + \frac{K|\Sigma|}{N + K|\Sigma|} \left( \sqrt{\frac{3P_i \log(\frac{2}{\delta})}{N}} + \frac{3 \log(\frac{2}{\delta})}{N} + P_i \right) \\
&\leq 2\sqrt{\frac{3P_i \log(\frac{2}{\delta})}{N}} + \frac{6 \log(\frac{2}{\delta})}{N} + \frac{K|\Sigma|}{N + K|\Sigma|} P_i + \frac{K}{N}.
\end{aligned}$$

We wish to show that  $\Pr \left[ d_{\chi^2}(P, Q) = \sum_i \frac{(P_i - Q_i)^2}{Q_i} \leq C\varepsilon^2 \right] \geq 1 - \delta$ , when taking at most  $\frac{|\Sigma|}{\varepsilon^2} \log(|\Sigma|\delta^{-1})$  samples. We split the analysis into two cases:

If  $P_i \leq \frac{C' \log(\frac{2}{\delta})}{N}$ , then  $\sqrt{\frac{3P_i \log(\frac{2}{\delta})}{N}} \leq \frac{\sqrt{3C' \log(\frac{2}{\delta})}}{N}$  and  $\frac{K|\Sigma|}{N + K|\Sigma|} P_i \leq \frac{C' \log(\frac{2}{\delta})}{N}$  and thus  $2\sqrt{\frac{3P_i \log(\frac{2}{\delta})}{N}} + \frac{6 \log(\frac{2}{\delta})}{N} + \frac{K|\Sigma|}{N + K|\Sigma|} P_i + \frac{K}{N} \leq \left( 2\sqrt{3C'} + 6 + C' \right) \frac{\log(\frac{2}{\delta})}{N} + \frac{K}{N}$ . We then have

$$\frac{(P_i - Q_i)^2}{Q_i} \lesssim \frac{(\log(\frac{1}{\delta}))^2}{N^2 Q_i} + \frac{K^2}{N^2 Q_i} \lesssim \frac{(\log(\frac{1}{\delta}))^2}{NK} + \frac{K}{N},$$

since  $Q_i \geq \frac{K}{N + K|\Sigma|} \geq \frac{K}{2N}$ , for  $N \geq C'K \cdot |\Sigma|$ .<sup>2</sup>

If  $P_i > \frac{C' \log(\frac{1}{\delta})}{N}$ ,  $|P_i - Q_i| \lesssim P_i \left( \frac{1}{\sqrt{C'}} + \frac{1}{C'} + \frac{1}{\frac{K|\Sigma|}{N} + 1} \right) \lesssim P_i \left( \frac{1}{\sqrt{C'}} + \frac{1}{C'} + \frac{1}{C'+1} \right)$  and thus, for large enough  $C'$ ,  $Q_i \geq P_i/2$ , giving us

$$\begin{aligned}
\frac{(P_i - Q_i)^2}{Q_i} &\lesssim \frac{\left( \sqrt{\frac{P_i \log(\frac{1}{\delta})}{N}} + \frac{K|\Sigma|}{N + K|\Sigma|} P_i + \frac{K}{N} \right)^2}{P_i} \\
&\lesssim \frac{\log(\frac{1}{\delta})}{N} + \left( \frac{K|\Sigma|}{N + K|\Sigma|} \right)^2 P_i + \frac{K^2}{N^2 P_i} \\
&\lesssim \frac{\log(\frac{1}{\delta})}{N} + \left( \frac{K|\Sigma|}{N + K|\Sigma|} \right)^2 P_i + \frac{K^2}{N \log \frac{1}{\delta}}.
\end{aligned}$$

<sup>2</sup>If  $N < C'K|\Sigma|$ , then  $Q_i > \frac{1}{(1+C')|\Sigma|}$  and we can easily construct counter example for which  $|P_i - Q_i| > \varepsilon$  for small enough  $\varepsilon$ .

Combining both cases, and applying a union bound, we have that with probability at least  $1 - |\Sigma|\delta$ ,

$$\sum_i \frac{(P_i - Q_i)^2}{Q_i} \lesssim |\Sigma| \left( \frac{(\log(\frac{1}{\delta}))^2}{NK} + \frac{K}{N} + \frac{\log(\frac{1}{\delta})}{N} + \frac{K^2}{N \log \frac{1}{\delta}} \right) + \left( \frac{K|\Sigma|}{N + K|\Sigma|} \right)^2.$$

Analyzing all the individual terms to have the summation to be bounded by  $\varepsilon^2$ , we need  $N \geq \frac{|\Sigma|}{\varepsilon^2} \cdot \max \left\{ \frac{\log^2(1/\delta)}{K}, K, \log(1/\delta), \frac{K^2}{\log(1/\delta)}, \varepsilon \cdot K \right\} \geq \frac{|\Sigma|}{\varepsilon^2} \cdot \max \left\{ \frac{\log^2(1/\delta)}{K}, K, \log(1/\delta), \frac{K^2}{\log(1/\delta)} \right\} \geq \frac{|\Sigma| \log(1/\delta)}{\varepsilon^2}$ , and the optimal is reached when  $K = \Theta(\log(1/\delta))$ . Rescaling  $\delta$  to adjust for the  $|\Sigma|\delta$  union bound, we conclude our proof.  $\square$

**4.2 A near-proper learning algorithm for Bayes nets** For learning general distributions on the Boolean hypercube  $\{0, 1\}^n$ , it is known that  $\Theta(2^n/\varepsilon^2)$  samples are both necessary and sufficient to learn within  $\chi^2$  divergence  $\varepsilon^2$  (with constant probability). As a direct consequence, we can obtain a non-proper Bayes net learning algorithm on the entire support, costing at most  $O(2^n/\varepsilon^2)$  samples. However, this approach is not interesting in our context: it costs more samples to learn than to test, the resulting distribution is not a Bayes net, and the lack of triangle inequality in  $\chi^2$  means that we cannot find a close enough distribution in the space of Bayes net to make it proper. In fact, it is unclear (to us) whether properly learning a Bayes net in  $\chi^2$  is feasible in  $O(2^n/\varepsilon^2)$ .

While one would hope that since Bayes nets have a sparse description ( $2^d \cdot n$  vs.  $2^n$ ), it would allow us to bypass the  $\Omega(2^n/\varepsilon^2)$  lower bound presented in [KOPS15] and possibly gives us a much better upper bound. However, we will show in Section 4.6 that, in stark contrast to learning in KL, learning a sparse Bayes net in  $\chi^2$  remains exponentially hard in sample complexity.

Loosely speaking, our lower bound hides a “distinguished node” behind a very biased common parent. By construction, the learning algorithm will never observe this distinguished “rare” node unless it takes exponentially many samples, which leads to an estimation error, entirely due to this very biased, large out-degree parent node. While this estimation error would be acceptable under TV distance or even KL divergence, the brittleness of  $\chi^2$  divergence to low-probability elements leads to a very large estimation error overall. Thus, any  $\chi^2$  learning algorithm *on the entire support* cannot afford to be inaccurate even on such rare nodes, and thus must take exponentially many (in  $n$ ) samples.

Nevertheless, given that testing is our end goal here, we only need a majority of the support to be learned well to proceed, which allows us to bypass this lower bound. Specifically, for testing in TV distance, it suffices to guarantee that  $\hat{P}$  is close to  $P$  in  $\chi^2$ , on a majority of the support  $S$ :

$$d_{\chi^2}(P_S, \hat{P}_S) = d_{\chi^2}(P, \hat{P}, S) = \sum_{i \in S} \frac{(P_i - \hat{P}_i)^2}{\hat{P}_i},$$

where  $\sum_{i \in S} P_i = P(S) \geq 1 - O(\varepsilon)$ . For Hellinger, we need something slightly stronger, but in the same spirit. Such framework is already present [ADK15, DKW18], though it is only implied in the analysis of [DKW18] for Hellinger distance. Thus, the main problem that remains is the availability of such (sample-efficient) near-proper learning of Bayes nets algorithm.

Connecting back to the degree-1 lower bound, the main difficulty is the information bottleneck presented by the biased parent. By relaxing the problem into near-proper learning, we can simply give up on learning the rare set of parents altogether (when it is sufficiently small), since the sum of these masses will not exceed  $O(\varepsilon)$ . From here, a natural idea is to exclude a subset of the support, where a child and its parents’ masses are at most  $O(\varepsilon/(2^d \cdot n))$ .<sup>3</sup> This guarantee would be enough for our

<sup>3</sup>There are at most  $2^d \cdot n$  parent configurations; excluding them with this threshold, we can still have a mass of  $1 - O(\varepsilon)$  left.

main result, testing with respect to TV distance; however, to extend it to the more stringent Hellinger testing guarantee, one needs to strengthen this to removing a subset of mass at most  $O(\varepsilon^2)$ , instead of  $O(\varepsilon)$ . As this extension changes little of the proof and provides a stronger result, in the remainder of the analysis, we focus on this goal.

In the interest of space and formatting, we use the abbreviated notation below in this section:

$$P(x_i, \pi_i(x)) := P_{X_i, \Pi_i}(x_i, \pi_i(x)). \quad (4.2)$$

We define a support of interest to learn on:

$$S'_k := \left\{ x \in \{0, 1\}^k \mid \forall i \in [k], P(x_i, \pi_i(x)) \geq 4c \frac{\varepsilon^2}{2^{d+1}n} \right\}. \quad (4.3)$$

We also define a superset  $S_k$  of  $S'_k$ :

$$S_k := \left\{ x \in \{0, 1\}^k \mid \forall i \in [k], P(x_i, \pi_i(x)) \geq c \frac{\varepsilon^2}{2^{d+1}n} \right\}. \quad (4.4)$$

In what follows, we assume the algorithm is provided with a set  $\tilde{S}_k$  such that  $S'_k \subseteq \tilde{S}_k \subseteq S_k$ . Indeed, [Lemma 4.2](#), analyzed in [Section 4.3](#) via [Algorithm 1](#), guarantees that we can efficiently learn such a set with high probability.

---

**Algorithm 1:**  $\varepsilon^2$ -majority support identification

---

**Input :** Sample access to distribution  $P$ , accuracy parameter  $\varepsilon$  and a DAG  $G$ .

- 1 Draw a multiset  $S$  of  $m$  samples from  $P$ , where  $m = 3 \cdot 2^{d+1}n \log(6 \cdot 2^{d+1}n) / (c \cdot \varepsilon^2)$ .
  - 2 Let  $N_{x_i, \pi_i}$  be the number of occurrences of  $(x_i, \pi_i)$  in  $S$ .
  - 3 Let  $Z_{x_i, \pi_i} = \frac{N_{x_i, \pi_i}}{m}$  and  $\bar{S} := \emptyset$ .
  - 4 **for**  $i = 1$ ;  $i \leq n$ ;  $i = i + 1$  **do**
  - 5     **for**  $(x_i, \pi_i) \in \{0, 1\}^{|\Pi_i|+1}$  **do**
  - 6         **if**  $Z_{x_i, \pi_i} \leq 2c \frac{\varepsilon^2}{2^{d+1}n}$  **then**  $\bar{S} \leftarrow \bar{S} \cup \{i, (x_i, \pi_i)\}$
  - 7     **end**
  - 8 **end**
  - 9 Mark all pairs in  $\bar{S}$ :  $X_i = x_i, \Pi_i = \pi_i$  as excluded (from  $\{0, 1\}^n$ ) and return the remaining support.
- 

A straightforward way to approach this problem is to consider learning guarantees on all the conditionals, i.e.,  $d_{\chi^2}(P_{X_i|\Pi_i}, Q_{X_i|\Pi_i}) \leq \frac{\varepsilon^2}{n}$  for any  $\Pi_i = \pi_i$  and hence<sup>4</sup>

$$1 + d_{\chi^2}(P, Q) \leq \prod_{i=1}^n \left( 1 + \max_{\pi_i} d_{\chi^2}(P_{X_i|\Pi_i=\pi_i}, Q_{X_i|\Pi_i=\pi_i}) \right) \leq \left( 1 + \frac{\varepsilon^2}{n} \right)^n \leq 1 + \varepsilon^2.$$

Coupled this with the fact that mass on the parents is lower bounded by  $\frac{\varepsilon}{2^d n}$ , we can obtain enough samples for learning each conditional by paying an extra  $O\left(\frac{2^d n}{\varepsilon}\right)$  per  $O\left(\frac{n}{\varepsilon^2}\right)$  (and some log factors for high probability learning and a union bound), giving us a sample complexity of  $\tilde{O}\left(\frac{2^d n^2}{\varepsilon^3}\right)$  and in the case of  $1 - O(\varepsilon^2)$ , a sample complexity of  $\tilde{O}\left(\frac{2^d n^2}{\varepsilon^4}\right)$ . As we will see later, we can do something slightly better to tighten the dependence on  $\varepsilon$ .

We will defer the proof of [Lemma 4.1](#) to [Sections 4.4](#) and [4.5](#), and provide a proof sketch in-place.

---

<sup>4</sup>While  $1 + d_{\chi^2}(P, Q)$  should be  $2P(S) - Q(S) + d_{\chi^2}(P_S, Q_S)$ , it does not affect the analysis if  $P(S) \geq 1 - O(\varepsilon^2)$ .

---

**Algorithm 2:**  $\varepsilon^2$ -near proper learning in  $\chi^2$ 


---

**Input :** Sample access to distribution  $P$ , accuracy parameter  $\varepsilon$ .

- 1 /\* Obtain the majority support  $\mathcal{A}$ . \*/
- 2  $\mathcal{A} \leftarrow$  call **Algorithm 1** with  $P$  and  $\varepsilon$ .
- 3 Draw a multiset  $S$  of  $m$  samples from  $P$ , where  $m = O(2^d n^2 \log(2^d n)/\varepsilon^2)$ .
- 4  $K \leftarrow \Theta(\log(2^{d+1} \cdot n))$
- 5 Let  $N_{x_i, \pi_i}$  be the count of  $X_i = x_i, \Pi_i = \pi_i$  out of the  $m$  samples.
- 6 **for**  $i = 1; i \leq n; i = i + 1$  **do**
- 7     **for**  $x_i, \pi_i \in \{0, 1\}^{|\Pi_i|+1}$  **do**
- 8          $Q(x_i|\pi_i) \leftarrow \frac{K + N_{x_i, \pi_i}}{\sum_{x'_i \in \{0, 1\}} K + N_{x'_i, \pi_i}}$
- 9     **end**
- 10 **end**
- 11 /\* A mass shifting step in (4.8) is necessary for testing in Hellinger downstream. \*/
- 12  $Q(x_1, \dots, x_n) \leftarrow \prod_{i=1}^n Q(x_i|\pi_i)$

---

LEMMA 4.1. When  $m = O\left(\frac{2^d n^2 \log(2^d n)}{c\varepsilon^2}\right)$ , the hypothesis  $Q$  output by **Algorithm 2** satisfies the following recurrence:

$$d_{\chi^2}(P_{X_1, \dots, X_k}, Q_{X_1, \dots, X_k}, S_k) \leq \left(1 + \frac{1}{n}\right) d_{\chi^2}(P_{X_1, \dots, X_{k-1}}, Q_{X_1, \dots, X_{k-1}}, S_{k-1}) + O\left(\frac{\varepsilon^2}{n}\right),$$

which, by recursion, implies

$$d_{\chi^2}(P_{X_1, \dots, X_n}, Q_{X_1, \dots, X_n}, S_n) \leq O(\varepsilon^2).$$

*Proof.* [Sketch] By lower bounding all parents' masses, we can guarantee that the  $\chi^2$  error on all conditionals of the hypothesis will be bounded by,

$$d_{\chi^2}(P_{X_i|\Pi_i}, Q_{X_i|\Pi_i}) \leq \frac{O(1)}{m \cdot P(\Pi_i)}$$

with an application of a Chernoff bound; in contrast, the naive approach is to simply take  $P(\Pi_i) \geq \frac{\varepsilon^2}{2^d n}$ . With this and some careful rearrangement of the residual terms, we are able to show this recurrence. Such arrangement is tricky because the standard  $\chi^2$  form becomes  $-2P(S) + Q(S) + \sum_i \frac{P_i^2}{Q_i}$  rather than the standard  $-1 + \sum_i P_i^2/Q_i$  form.  $\square$

For testing in  $d_H$ , we need the following tweaks: from the obtained  $Q$ , we shift masses to get a  $\tilde{Q}$  s.t.,  $\tilde{Q}(\tilde{S}_n) \geq 1 - O(\varepsilon^2)$ , while maintaining the graphical structure of  $Q$  and the  $d_{\chi^2}$  closeness from  $P$  on  $\tilde{S}_n$ ; though this could potentially make the  $d_{\chi^2}$  on the entire support unbounded (it will be infinity since there are 0's in the denominator), it does not affect our downstream testing. As we will see later on in the analysis, this is an extra (and seemingly necessary) step for testing maximum in-degree in  $d_H$ ; but for  $d_{TV}$ , interestingly,  $Q$  and  $\tilde{S}_n$  is sufficient.

THEOREM 4.1. Given  $O(2^d n^2 \log(2^d n)/\varepsilon^2)$  samples, we can obtain a proper hypothesis  $Q$  of the degree- $d$  Bayes net  $P$ , and a  $\tilde{S}_n \subset \{0, 1\}^n$  on which  $d_{\chi^2}(P, Q, \tilde{S}_n) \leq O(\varepsilon^2)$ , and  $P(\tilde{S}_n) \geq 1 - O(\varepsilon^2)$ .

Additionally, with some post-processing (without extra samples), we can obtain another proper hypothesis  $\tilde{Q}$  with guarantees subsuming the above, and:  $\tilde{Q}(\tilde{S}_n) = 1 > 1 - O(\varepsilon^2)$ .

*Proof.* By [Lemma 4.1](#), and the fact that  $\tilde{S}_n \subseteq S_n$ ,

$$d_{\chi^2}(P_{X_1, \dots, X_n}, Q_{X_1, \dots, X_n}, \tilde{S}_n) \leq d_{\chi^2}(P_{X_1, \dots, X_n}, Q_{X_1, \dots, X_n}, S_n),$$

which is  $O(\varepsilon^2)$ . This concludes the proof.  $\square$

**4.3 Efficient Estimation of  $O(\varepsilon^2)$ -effective support** We write the formal statement regarding  $\tilde{S}_n$  assumed in [Section 4.2](#) as [Lemma 4.2](#).

**LEMMA 4.2.** *There exists a routine ([Algorithm 1](#)) that takes at most  $O(2^{d+1}n \log(2^{d+1}n)/(c \cdot \varepsilon^2))$  samples from  $P$ , and return an approximation  $\tilde{S}_n$  with the following guarantees: for all  $k \in [n]$ , and with probability at least  $5/6$ ,*

$$S'_k \subseteq \tilde{S}_k \subseteq S_k,$$

where  $S'_k, S_k$  are as defined in [Equations \(4.3\) and \(4.4\)](#).

*Proof.* We prove by analyzing [Algorithm 1](#). The algorithm takes  $m = \frac{3}{c} \cdot 2^{d+1}n \log(6 \cdot 2^{d+1}n)/\varepsilon^2$  samples and checks if the ratio of each  $X_i, \Pi_i$  exceeds  $2c \frac{\varepsilon^2}{2^{d+1}n}$  – for  $x \in \{0, 1\}^n$ , if all  $i \in [n]$ ,  $(x_i, \pi_i) := (x_i(x), \pi_i(x)) \in \{0, 1\}^{|\Pi_i|+1}$ ,  $N_{x_i, \pi_i} \geq 2c \frac{\varepsilon^2 \cdot m}{2^{d+1}n}$  then  $x$  gets added to  $\tilde{S}_n$ , where  $N_{x_i, \pi_i}$  is the number of occurrences of  $X_i = x_i, \Pi_i = \pi_i$  over the  $m$  samples.

Our argument here is to ensure that the  $a \in \{0, 1\}^{|\Pi_i|+1}$  with smaller masses than  $c \frac{\varepsilon^2}{2^{d+1}n}$  won't pass the procedure, and that for each  $i \in [n]$ , at most  $O(\varepsilon^2/n)$  of masses are dropped. We do so via a Chernoff bound, a stochastic dominance argument and a union bound. First, consider the Bernoulli distribution  $T \sim \text{Bern}(m, p)$ , where  $p = c \frac{\varepsilon^2}{2^{d+1}n}$ . By Chernoff's inequality,

$$\Pr \left[ T \geq 2 \cdot c \frac{\varepsilon^2}{2^{d+1}n} \right] = \Pr[T \geq 2mp] \leq \exp(-mp/3) = \exp\left(-mc \frac{\varepsilon^2}{2^{d+1}n} / 3\right) = \frac{1}{6 \cdot 2^{d+1}n}.$$

Since any  $T' \sim \text{Bern}(m, p')$  is first-order stochastically dominated by  $T \sim \text{Bern}(m, p)$  if  $p' \leq p$ , thus for  $p' < \frac{c}{2} \frac{\varepsilon^2}{2^{d+1}n}$ ,  $\Pr \left[ T' \geq 2c \frac{\varepsilon^2}{2^{d+1}n} \right] \leq \Pr \left[ T \geq 2c \frac{\varepsilon^2}{2^{d+1}n} \right] \leq \frac{1}{2^{d+1}n}$ .

Therefore, for  $p \leq c \frac{\varepsilon^2}{2^{d+1}n}$ ,  $T \sim \text{Bern}(m, p)$ ,  $\Pr \left[ T \geq 2c \frac{\varepsilon^2}{2^{d+1}n} \right] \leq \frac{1}{2^{d+1}n}$ .

Via similar argument, all  $p \geq 4c \frac{\varepsilon^2}{2^{d+1}n}$  will pass the test with high probability: let  $T'' \sim \text{Bern}(m, p'')$ , where  $p'' \geq 4c \frac{\varepsilon^2}{2^{d+1}n}$ , and by Chernoff,

$$\Pr \left[ T'' \leq \frac{1}{2} \cdot 4c \frac{\varepsilon^2}{2^{d+1}n} \right] = \Pr \left[ T'' \leq \frac{1}{2} mp'' \right] \leq \exp(-mp''/8) \leq \exp\left(-\frac{1}{2} mc \frac{\varepsilon^2}{2^{d+1}n}\right) \leq \frac{1}{6 \cdot 2^{d+1}n}.$$

Finally, a union bound over all elements  $a \in \{0, 1\}^{|\Pi_i|+1} = \text{Support}(X_i, \Pi_i)$  implies that (w.h.p.) all  $i \in [n]$  and  $a \in \{0, 1\}^{|\Pi_i|+1}$ :  $P_{X_i, \Pi_i}(a) \geq 4c \frac{\varepsilon^2}{2^{d+1}n}$  will pass the test;  $P_{X_i, \Pi_i}(a) \leq c \frac{\varepsilon^2}{2^{d+1}n}$  will fail the test. This tells us that, for  $k \in [n]$ ,  $\left\{ x \in \{0, 1\}^k \mid \forall i \in [k], P_{X_i, \Pi_i}(x_i, \pi_i(x)) \geq 4c \frac{\varepsilon^2}{2^{d+1}n} \right\} \subset \tilde{S}_n \subset \left\{ x \in \{0, 1\}^k \mid \forall i \in [k], P_{X_i, \Pi_i}(x_i, \pi_i(x)) \geq c \frac{\varepsilon^2}{2^{d+1}n} \right\}$ .  $\square$

**4.4 Some useful high probability events for [Lemma 4.1](#)** In this subsection, we analyze some crucial high probability events in the form of [Equations \(4.7\) and \(4.9\)](#), to facilitate our main proof. But first we need to build up some technical machinery. We will use the term *configuration* to refer to

some set of binary strings that is a subset of  $\{0, 1\}^n$ . First, we define the set of configurations for parent nodes with “large enough” masses,

$$A_{S_k} := \left\{ a \in \{0, 1\}^{|\Pi_k|} \mid P_{\Pi_k}(a) \geq c \cdot \frac{\varepsilon^2}{2^{dn}} \right\}.$$

We define another closely related set of configurations: let  $C(a, S_k)$  be the set of configurations (excluding the current parent nodes  $\Pi_k$  and last node  $X_k$ ) remaining in  $S_k$ ,<sup>5</sup> given  $\Pi_k = a$  and it is independent of the value  $X_k$  takes. Formally:

$$\rightarrow C(a, S_k) = \{x \in \{0, 1\}^{k-1-|\Pi_k|} \mid \exists x' \in \{0, 1\}^{\{X_1, \dots, X_{k-1} \setminus \Pi_k\}} \{(X_1, \dots, X_{k-1} \setminus \Pi_k) = x', \Pi_k = a, X_k = x\} \in S_k\}$$

Or equivalently  $C(a, S_k) = \{x \in \{0, 1\}^{k-1-|\Pi_k|} \mid \{(X_1, \dots, X_{k-1} \setminus \Pi_k) = x, \Pi_k = a\} \in S_{k-1}\}$  – fixing  $\Pi_k$  to  $a$  out of the  $X_1, \dots, X_k$  variables and checking if  $x, a \in S_{k-1}$ . Similarly,

$$\rightarrow B(a, S_k) = \{x \in \{0, 1\} \mid \forall x' \in \{0, 1\}^{k-1-|\Pi_k|} \{(X_1, \dots, X_{k-1} \setminus \Pi_k) = x', \Pi_k = a, X_k = x\} \in S_k\}$$

$$\text{Or we may equivalently define } B(a, S_k) = \left\{ x \in \{0, 1\} \mid P(\Pi_k = a, X_k = x) \geq c \cdot \frac{\varepsilon^2}{2^{d+1}n} \right\}.$$

Other sets of configurations include:  $C_k := \{0, 1\}^{k-1-|\Pi_k|}$ , the full set of configurations for  $\{X_1, \dots, X_{k-1} \setminus \Pi_k\}$  without any restrictions; just as  $B_k := \{0, 1\}$  is related to  $X_k$ ;  $A_k := \{0, 1\}^{|\Pi_k|}$ ;  $A_{S_k^c} := A_k \setminus A_{S_k}$ ;  $B(a, S_k^c) := B_k \setminus B(a, S_k)$ ;  $C(a, S_k^c) := C_k \setminus C(a, S_k)$ . Note that  $x$  where  $(x_k, \pi_k) \in A_{S_k}$  may not imply  $x \in S_k$  (its converse is true); and knowing  $\Pi_k = a$ , in order for  $x \in S_k$ , it has to satisfy both constraints – one from  $\{X_1, \dots, X_{k-1} \setminus \Pi_k\}$  and the other  $\{X_k\}$ .

Let  $m_{\Pi_k=a}$  be the random variable counting the number of samples with  $\Pi_k = a$ . For any  $x \in S_n$ , every  $k \in [n]$ , and  $a \in \{0, 1\}^{|\Pi_k|}$  satisfies  $P(\Pi_k = a) = \sum_{x \in X_k} P(x, \Pi_k = a) \geq 2 \cdot c \frac{\varepsilon^2}{2^{d+1}n} = c \frac{\varepsilon^2}{2^{dn}}$ , and thus by Chernoff,  $m_{\Pi_k=a} \geq \frac{1}{2} m P(\Pi_k = a)$ , with v.h.p. We condition on this event, and with the learning result in [Proposition 4.1](#), by setting  $K = \log(6 \cdot 2^{d+1}n)$ , we can derive: for all  $a \in A_{S_k}$ ,

$$\Pr \left[ -1 + \sum_{b \in B} \frac{P_{X_k|\Pi_k}^2(b|a)}{Q_{X_k|\Pi_k}(b|a)} \geq \frac{c' \log(2^{d+1}n)}{m_{\Pi_k=a}} \right] \leq \frac{1}{6} \frac{1}{2^{d+1}n}, \quad (4.5)$$

for some constant  $c'$ .

And thus, with high probability, for all  $a \in A_{S_k}$ ,

$$-1 + \sum_{b \in B} \frac{P_{X_k|\Pi_k}^2(b|a)}{Q_{X_k|\Pi_k}(b|a)} \leq \frac{c' \log(2^{d+1}n)}{m_{\Pi_k=a}} \leq \frac{2c' \cdot \log(2^{d+1}n)}{mP(\Pi_k = a)}. \quad (4.6)$$

We then again condition on this event happening; since

$$\sum_{b \in B(a, S_k)} \frac{P_{X_k|\Pi_k}^2(b|a)}{Q_{X_k|\Pi_k}(b|a)} \leq \sum_{b \in B} \frac{P_{X_k|\Pi_k}^2(b|a)}{Q_{X_k|\Pi_k}(b|a)},$$

giving us,

$$-1 + \sum_{b \in B(a, S_k)} \frac{P_{X_k|\Pi_k}^2(b|a)}{Q_{X_k|\Pi_k}(b|a)} \leq \frac{2c' \cdot \log(2^{d+1}n)}{mP(\Pi_k = a)}. \quad (4.7)$$

<sup>5</sup>This term is used to analyze (independently) the two terms generated by conditioning on the parent nodes  $\Pi_k$  in this Markov chain:  $\{X_1, \dots, X_{k-1}\} \setminus \Pi_k \rightarrow \Pi_k \rightarrow X_k$ .

In the later part of the proof of [Lemma 4.1](#), we will need a stronger condition on our density estimate (a slightly different  $Q$  by moving at most  $O(\varepsilon^2)$  mass around) – we will call it  $\tilde{Q}$ , and we can obtain it directly from  $Q$  and  $\tilde{S}_n$ :

$$\tilde{Q}_{X_k|\Pi_k}(b|a) := \begin{cases} \frac{Q_{X_k|\Pi_k}(b|a)}{\sum_{b \in B(a, \tilde{S}_k)} Q_{X_k|\Pi_k}(b|a)} & \text{if } b \in B(a, \tilde{S}_k), \\ 0 & \text{otherwise.} \end{cases} \quad (4.8)$$

Furthermore,  $\tilde{Q}$  on  $\tilde{S}_n$  shares very similar guarantees as  $Q$  on  $S_n$ . For every  $a \in A_{\tilde{S}_k} \subseteq A_{S_k}$ , since  $B(a, \tilde{S}_k) \subseteq B(a, S_k)$  (conditioning on [Equation 4.6](#)),

$$\sum_{b \in B(a, \tilde{S}_k)} \frac{P_{X_k|\Pi_k}^2(b|a)}{Q_{X_k|\Pi_k}(b|a)} \leq \sum_{b \in B(a, S_k)} \frac{P_{X_k|\Pi_k}^2(b|a)}{Q_{X_k|\Pi_k}(b|a)} \leq 1 + \frac{2c' \cdot \log(2^{d+1}n)}{mP(\Pi_k = a)}.$$

As moving masses from  $B(a, \tilde{S}_k^c)$  to  $B(a, \tilde{S}_k)$  will only reduce the quantity  $\frac{P_{X_k|\Pi_k}^2(b|a)}{Q_{X_k|\Pi_k}(b|a)}$  further, we have

$$-1 + \sum_{b \in B(a, \tilde{S}_k)} \frac{P_{X_k|\Pi_k}^2(b|a)}{\tilde{Q}_{X_k|\Pi_k}(b|a)} \leq \frac{2c' \cdot \log(2^{d+1}n)}{mP(\Pi_k = a)}. \quad (4.9)$$

Finally, by a union bound, we have that with probability at least  $5/6$ , for all  $k \in [n]$  and  $\Pi_k = a$ , both statements in [Equations \(4.7\)](#) and [\(4.9\)](#) will hold. Throughout the rest of the appendix, we will condition on these two statements.

**4.5 Proof of [Lemma 4.1](#)** We write the partial sum of  $\chi^2$  between  $P_{X_1, \dots, X_k}$  and  $Q_{X_1, \dots, X_k}$  on the subset  $S_k \subset \{0, 1\}^k$  as:

$$\begin{aligned} d_{\chi^2}(P_{X_1, \dots, X_k}, Q_{X_1, \dots, X_k}, S_k) &= \sum_{x \in S_k} \frac{(P_{X_1, \dots, X_k}(x) - Q_{X_1, \dots, X_k}(x))^2}{Q_{X_1, \dots, X_k}(x)} \\ &= \sum_{x \in S_k} -2P_{X_1, \dots, X_k}(x) + Q_{X_1, \dots, X_k}(x) + \frac{P_{X_1, \dots, X_k}(x)^2}{Q_{X_1, \dots, X_k}(x)} \\ &= -2P_{X_1, \dots, X_k}(S_k) + Q_{X_1, \dots, X_k}(S_k) + \sum_{x \in S_k} \frac{P_{X_1, \dots, X_k}(x)^2}{Q_{X_1, \dots, X_k}(x)}, \end{aligned}$$

By definition of  $S_k$ , we can write the sum over  $x \in S_k$  as  $a \in A_{S_k}$ ,  $b \in B(a, S_k)$  and  $g \in C(a, S_k)$ :

$$\begin{aligned} &d_{\chi^2}(P_{X_1, \dots, X_k}, Q_{X_1, \dots, X_k}, S_k) \\ &= \left\{ \sum_{a \in A_{S_k}} \frac{P_{\Pi_k}^2(a)}{Q_{\Pi_k}(a)} \cdot \sum_{g \in C(a, S_k)} \frac{P_{X_1, \dots, X_{k-1} \setminus \Pi_k | \Pi_k}^2(g|a)}{Q_{X_1, \dots, X_{k-1} \setminus \Pi_k | \Pi_k}(g|a)} \cdot \sum_{b \in B(a, S_k)} \frac{P_{X_k | \Pi_k}^2(b|a)}{Q_{X_k | \Pi_k}(b|a)} \right\} \\ &\quad - (2P_{X_1, \dots, X_k}(S_k) - Q_{X_1, \dots, X_k}(S_k)) \\ &= \left\{ \sum_{a \in A_{S_k}} \frac{P_{\Pi_k=a}^2}{Q_{\Pi_k=a}} \cdot \sum_{g \in C(a, S_k)} \frac{P_{X_1, \dots, X_{k-1} \setminus \Pi_k | \Pi_k}^2(g|a)}{Q_{X_1, \dots, X_{k-1} \setminus \Pi_k | \Pi_k}(g|a)} \cdot \left( -1 + \sum_{b \in B(a, S_k)} \frac{P_{X_k | \Pi_k}^2(b|a)}{Q_{X_k | \Pi_k}(b|a)} \right) \right\} \\ &\quad - (2P_{X_1, \dots, X_{k-1}}(S_{k-1}) - Q_{X_1, \dots, X_{k-1}}(S_{k-1})) \end{aligned}$$

$$\begin{aligned}
& + \sum_{a \in A_{S_k}} \sum_{g \in C(a, S_k)} \frac{P_{\Pi_k=a}^2}{Q_{\Pi_k=a}} \cdot \frac{P_{X_1, \dots, X_{k-1} \setminus \Pi_k | \Pi_k}^2(g | a)}{Q_{X_1, \dots, X_{k-1} \setminus \Pi_k | \Pi_k}(g | a)} \\
& + (2P_{X_1, \dots, X_{k-1}}(S_{k-1}) - Q_{X_1, \dots, X_{k-1}}(S_{k-1})) \\
& - (2P_{X_1, \dots, X_k}(S_k) - Q_{X_1, \dots, X_k}(S_k)) \tag{4.10} \\
= & \left\{ \sum_{a \in A_{S_k}} \frac{P_{\Pi_k}^2(a)}{Q_{\Pi_k}(a)} \sum_{g \in C(a, S_k)} \frac{P_{X_1, \dots, X_{k-1} \setminus \Pi_k | \Pi_k}^2(g | a)}{Q_{X_1, \dots, X_{k-1} \setminus \Pi_k | \Pi_k}(g | a)} \cdot \left( -1 + \sum_{b \in B(a, S_k)} \frac{P_{X_k | \Pi_k}^2(b | a)}{Q_{X_k | \Pi_k}(b | a)} \right) \right\} \\
& + d_{\chi^2}(P_{X_1, \dots, X_{k-1}}, Q_{X_1, \dots, X_{k-1}}, S_{k-1}) \\
& + (2P_{X_1, \dots, X_{k-1}}(S_{k-1}) - Q_{X_1, \dots, X_{k-1}}(S_{k-1})) \\
& - (2P_{X_1, \dots, X_k}(S_k) - Q_{X_1, \dots, X_k}(S_k)) \tag{4.11} \\
\leq & 2c \frac{\varepsilon^2}{n} + 4c' \frac{2^d \log(2^d n)}{m} + \left( 1 + \frac{2c' 2^d n \log(2^d n)}{cm \varepsilon^2} \right) \times d_{\chi^2}(P_{X_1, \dots, X_{k-1}}, Q_{X_1, \dots, X_{k-1}}, S_{k-1}). \tag{4.12}
\end{aligned}$$

We obtain (4.10) by adding and subtracting the same terms, giving us a recurrence in  $S_{k-1}$  in (4.11). From this, we get (4.12) by applying a couple of technical results: [Lemmas 4.3](#) and [4.4](#). Finally, by setting  $m = \frac{4c2^d n^2 \log(2^d n)}{\varepsilon^2}$ , we get the desired recursive formulation for  $d_{\chi^2}(P_{X_1, \dots, X_k}, Q_{X_1, \dots, X_k}, S_k)$ , concluding our proof of [Lemma 4.1](#).

We now prove the aforementioned technical lemmata.

LEMMA 4.3.

$$(2P_{X_1, \dots, X_{k-1}}(S_{k-1}) - Q_{X_1, \dots, X_{k-1}}(S_{k-1})) - (2P_{X_1, \dots, X_k}(S_k) - Q_{X_1, \dots, X_k}(S_k)) \leq 2c \frac{\varepsilon^2}{n}.$$

*Proof.*

$$\begin{aligned}
& (2P_{X_1, \dots, X_{k-1}}(S_{k-1}) - Q_{X_1, \dots, X_{k-1}}(S_{k-1})) - (2P_{X_1, \dots, X_k}(S_k) - Q_{X_1, \dots, X_k}(S_k)) \\
= & \sum_{x_1, \dots, x_{k-1} \in S_{k-1}} \sum_{x_k \in \{0,1\}} 2P_{X_1, \dots, X_k}(x_1, \dots, x_k) - \sum_{x_1, \dots, x_k \in S_k} 2P_{X_1, \dots, X_k}(x_1, \dots, x_k) \\
& - (Q_{X_1, \dots, X_{k-1}}(S_{k-1}) - Q_{X_1, \dots, X_k}(S_k)) \\
= & \sum_{x_1, \dots, x_{k-1} \in S_{k-1}, x_k \in S^c(x_1, \dots, x_{k-1})} (2P(x_1, \dots, x_k) - Q(x_1, \dots, x_k)) \\
\leq & \sum_{x_1, \dots, x_{k-1} \in \{0,1\}^{k-1}} \sum_{x_k \in S^c(x_1, \dots, x_{k-1})} 2P(x_1, \dots, x_k) \\
= & 2 \sum_{\{x_1, \dots, x_{k-1} \setminus \pi_k\} \in \{0,1\}^{k-1-|\Pi_k|}} \sum_{\pi_k} \sum_{x_k \in S^c(\pi_k)} P(x_1, \dots, x_{k-1} \setminus \pi_k | \pi_k) P(x_k, \pi_k) \\
\leq & 2 \sum_{x_k, \pi_k \in S^c(X_k, \Pi_k)} P_{X_k, \Pi_k}(x_k, \pi_k) \leq 2c \frac{\varepsilon^2}{n}.
\end{aligned}$$

□

LEMMA 4.4.

$$\sum_{\substack{a \in A_{S_k} \\ g \in C(a, S_k)}} \frac{P_{\Pi_k=a}^2}{Q_{\Pi_k=a}} \frac{P_{X_1, \dots, X_{k-1} \setminus \Pi_k | \Pi_k}^2(g | a)}{Q_{X_1, \dots, X_{k-1} \setminus \Pi_k | \Pi_k}(g | a)} \left( -1 + \sum_{b \in B(a, S_k)} \frac{P_{X_k | \Pi_k}^2(b | a)}{Q_{X_k | \Pi_k}(b | a)} \right)$$

$$\leq \frac{2c' \cdot 2^d n \log(2^d n)}{cm\varepsilon^2} d_{\chi^2}(P_{X_1, \dots, X_{k-1}}, Q_{X_1, \dots, X_{k-1}}, S_{k-1}) + \frac{4c \cdot 2^d \log(2^d n)}{m}.$$

*Proof.*

$$\begin{aligned} & \sum_{\substack{a \in A_{S_k} \\ g \in C(a, S_k)}} \frac{P_{\Pi_k=a}^2}{Q_{\Pi_k=a}} \frac{P_{X_1, \dots, X_{k-1} \setminus \Pi_k | \Pi_k}^2(g|a)}{Q_{X_1, \dots, X_{k-1} \setminus \Pi_k | \Pi_k}(g|a)} \left( -1 + \sum_{b \in B(a, S_k)} \frac{P_{X_k | \Pi_k}^2(b|a)}{Q_{X_k | \Pi_k}(b|a)} \right) & (4.13) \\ \leq & \sum_{\substack{a \in A_S \\ g \in C(a, S_k)}} \frac{P_{\Pi_k=a}^2}{Q_{\Pi_k=a}} \frac{P_{X_1, \dots, X_{k-1} \setminus \Pi_k | \Pi_k}^2(g|a)}{Q_{X_1, \dots, X_{k-1} \setminus \Pi_k | \Pi_k}(g|a)} \frac{2c' \cdot \log(2^d n)}{m P_{\Pi_k}(a)} \\ = & \sum_{\substack{a \in A_S \\ g \in C(a, S_k)}} \left( -2P_{X_1, \dots, X_{k-1}}(a, g) + Q_{X_1, \dots, X_{k-1}}(a, g) + \frac{P_{X_1, \dots, X_{k-1}}^2(a, g)}{Q_{X_1, \dots, X_{k-1}}(a, g)} \right) \cdot \frac{2c' \cdot \log(2^d n)}{m P_{\Pi_k}(a)} \\ & + \sum_{\substack{a \in A_S \\ g \in C(a, S_k)}} (2P_{X_1, \dots, X_{k-1}}(a, g) - Q_{X_1, \dots, X_{k-1}}(a, g)) \cdot \frac{2c' \cdot \log(2^d n)}{m P_{\Pi_k}(a)} \\ = & \sum_{\substack{a \in A_S \\ g \in C(a, S_k)}} \frac{(P_{X_1, \dots, X_{k-1}}(a, g) - Q_{X_1, \dots, X_{k-1}}(a, g))^2}{Q_{X_1, \dots, X_{k-1}}(a, g)} \cdot \frac{2c' \cdot \log(2^d n)}{m P_{\Pi_k}(a)} \\ & + \sum_{a \in A_S} \sum_{g \in C(a, S_k)} (2P_{X_1, \dots, X_{k-1}}(a, g) - Q_{X_1, \dots, X_{k-1}}(a, g)) \cdot \frac{2c' \cdot \log(2^d n)}{m P_{\Pi_k}(a)} & (4.14) \\ \leq & \frac{2c' 2^d n \log(2^d n)}{cm\varepsilon^2} \cdot \sum_{\substack{a \in A_S \\ g \in C(a, S_k)}} \frac{(P_{X_1, \dots, X_{k-1}}(a, g) - Q_{X_1, \dots, X_{k-1}}(a, g))^2}{Q_{X_1, \dots, X_{k-1}}(a, g)} \\ & + \sum_{a \in A_S} \sum_{g \in C(a, S_k)} P_{X_1, \dots, X_{k-1}}(a, g) \cdot \frac{4c' \cdot \log(2^d n)}{m P_{\Pi_k}(a)} \\ = & \frac{2c' 2^d n \log(2^d n)}{cm\varepsilon^2} (P_{X_1, \dots, X_{k-1}}, Q_{X_1, \dots, X_{k-1}}, S_{k-1}) \\ & + \sum_{a \in A_S} P_{\Pi_k}(a) \cdot \frac{4c' \cdot \log(2^d n)}{m P_{\Pi_k}(a)} \underbrace{\sum_{g \in C(a, S_k)} P_{X_1, \dots, X_{k-1} \setminus \Pi_k | \Pi_k}(g|a)}_{\leq 1} \\ \leq & \frac{2c' 2^d n \log(2^d n)}{cm\varepsilon^2} d_{\chi^2}(P_{X_1, \dots, X_{k-1}}, Q_{X_1, \dots, X_{k-1}}, S_k) + \frac{4c' \cdot 2^d \log(2^d n)}{m}. & (4.15) \end{aligned}$$

□

**4.6 A lower bound for learning a Bayes net in  $\chi^2$**  Our lower bound relies on a family of degree-1 Bayes nets, with all  $n - 1$  nodes sharing the same common 1-node parent. We will set the probability of the parent to be so imbalanced that by taking even  $\exp(O(n))$  number of samples, we still cannot obtain one sample from the rare side. In this case, it would be impossible to observe any sample from one side of the  $n - 1$  conditionals and thus, it is information theoretically hard to obtain good estimates of these conditional densities. Due to the multiplicative accumulation of error in these “hidden” conditional

densities, the  $\chi^2$  distance remains large despite the small parent probability. Note that, this is not a problem for KL: the error expressed in terms of KL is linearly accumulated as compared to  $\chi^2$ 's multiplicative (and hence exponential) accumulation.

DEFINITION 4.1. We define a process for drawing our hard instances to analyze in [Proposition 4.2](#):

1. Let the prior  $P \sim \pi$  be distributions such that  $P = P_{X_1} \cdot P_{X_2|X_1} \cdots P_{X_n|X_1}$  and  $P(X_1 = 1) = \varepsilon_0$ , a parameter of our choosing.
2. Draw  $x^h$  uniformly at random from  $\{0, 1\}^{n-1}$ .
3. Based on  $x^h$ , set  $P_{X_i|X_1=1} = \delta_{x_i^h}$  and as a consequence  $P_{X_2, \dots, X_n|X_1=1} = \prod_{i=2}^n P_{X_i|X_1=1} = \delta_{x^h}$ , where

$$\delta_y(x) = \begin{cases} 1, & x = y \\ 0, & \text{otherwise} \end{cases}.$$

4. Let  $P_{X_2, \dots, X_n|X_1=0} = U_{n-1}$  be uniform on  $\{0, 1\}^{n-1}$  in the other case.

Intuitively, we are merely hiding a particular point  $x^h$  from the learners; each distribution in  $\pi$ , when conditioned on  $X_1 = 1$  will concentrate their mass on one point in the simplex (deterministic). While it takes only one sample (with  $x_1 = 1$ ) to learn, no learner can get one with less than  $O(1/\varepsilon_0)$  number of samples. We state the main result below.

PROPOSITION 4.2. *Minimax risk of estimating a degree-1 Bayes net  $P$  in  $\chi^2$  is at least  $\Omega(\varepsilon)$  when the number of samples  $m \leq O(2^{n/2}/\varepsilon)$ . In particular, the family of distributions in [Definition 4.1](#) takes at least  $\Omega(2^{n/2}/\varepsilon)$  to learn to  $\varepsilon$ .*

By a standard Lagrange multiplier calculation, as inspired by the proof of [\[KOPS15, Lemma 5\]](#), we have these useful facts. We will use them to prove [Proposition 4.2](#).

FACT 4.1. *Let  $q_i, a_i \geq 0, i \in [k]$ , such that  $\sum_{i=1}^k q_i \leq 1$ . Then the quantity  $\sum_{i=1}^k \frac{a_i}{q_i}$  is minimized when  $q_i \propto \sqrt{a_i}$ .*

FACT 4.2. *Optima of the following form is obtained when  $Q_i^2 = \frac{1}{2^{n-1}}, i = 1, \dots, n-1$ :*

$$\min_{Q \in \mathbb{R}^{n-1}} \sum_{i=1}^{2^{n-1}} \frac{1}{Q_i} \cdot \frac{1}{2^{n-1}}, \text{ s.t. } \sum_{i=1}^{2^{n-1}} Q_i^2 = 1.$$

*Proof.* To see this, we simply verify the K.K.T. condition [\[BV14\]](#) of the constrained optimization problem:

$$\sum_{i=1}^{2^{n-1}} \frac{1}{Q_i} \cdot \frac{1}{2^{n-1}} + \lambda \left( \sum_{i=1}^{2^{n-1}} Q_i^2 - 1 \right).$$

Necessary conditions:

$$-\frac{1}{Q_i^2} \cdot \frac{1}{2^{n-1}} + 2\lambda Q_i = 0, \forall i; \quad \sum_{i=1}^{2^{n-1}} Q_i^2 - 1 = 0.$$

Solving the first equation gives  $Q_i = \frac{1}{2^{n/3}\lambda^{1/3}}$ ; and since all  $Q_i$ s are equal,  $Q^*$  ought to be uniform. Since  $\frac{1}{Q_i}$  and  $Q_i^2$  are both convex, we have that the necessary conditions are also sufficient for global minima.  $\square$

*Proof.* [Proof of [Proposition 4.2](#)] Let  $s_i = (X_{1,i}, \dots, X_{n,i})$  be the  $i^{\text{th}}$  sample, denote event  $S = \{X_{1,i} = 0, i \in [m]\}$ , the probability is thus  $\Pr[S] = (1 - \varepsilon_0)^m \geq e^{-\frac{1}{2}\varepsilon_0 m} \geq 1 - \frac{1}{2}\varepsilon_0 m$ . We will condition on  $S$  being true during the computation. Let  $\mathcal{A}_m$  denote the set of deterministic algorithms taking  $m$  samples from  $P$ . Then, for  $R_{\chi^2}(m)$  being the minimax risk over  $\mathcal{A}_m$ ,

$$\begin{aligned}
R_{\chi^2}(m) &= \inf_{Q \in \mathcal{A}_m} \sup_{P \in \mathcal{P}} \mathbb{E}_{s_1, \dots, s_m \sim P} [d_{\chi^2}(P, Q_s)] \\
&\geq \inf_{Q \in \mathcal{A}_m} \mathbb{E}_{P \sim \pi} \mathbb{E}_{s \sim P^{\otimes m}} [d_{\chi^2}(P, Q_s)] \\
&\geq \inf_{Q \in \mathcal{A}_m} \mathbb{E}_{P \sim \pi} \mathbb{E}_{s \sim P^{\otimes m}} [d_{\chi^2}(P, Q_s) | S] \Pr[S] \\
&= \inf_{Q \in \mathcal{A}_m} \mathbb{E}_{P \sim \pi} \mathbb{E}_{s \sim P^{\otimes m}} \left[ -1 + \sum_{x_1, \dots, x_n} \frac{P^2(x_1, \dots, x_n)}{Q_s(x_1, \dots, x_n)} | S \right] \Pr[S]. \tag{4.16}
\end{aligned}$$

Denote for convenience,  $P_0 = P_{X_1, \dots, X_{n-1} | X_n=0}$ ;  $P_1 = P_{X_1, \dots, X_{n-1} | X_n=1}$ ;  $Q_{0,s} = Q_s(X_1, \dots, X_{n-1} | X_n=0)$ ;  $Q_{1,s} = Q_s(X_1, \dots, X_{n-1} | X_n=1)$ . We focus on the inner summation and lower bound them separately; and before that, we need a separate auxiliary tool – using [Fact 4.1](#) above, we can show that, for any fixed  $P$  and  $Q_s$ ,

$$\begin{aligned}
&\frac{P^2(x_1=0)}{Q_s(x_1=0)} + \frac{P^2(x_1=1)}{Q_s(x_1=1)} (1 + d_{\chi^2}(P_1, Q_{1,s})) \\
&\geq \frac{P^2(x_1=0)}{P(x_1=0) + P(x_1=1)\sqrt{1+d_{\chi^2}(P_1, Q_1)}} + \frac{P^2(x_1=1)(1+d_{\chi^2}(P_1, Q_1))}{P(x_1=1)\sqrt{1+d_{\chi^2}(P_1, Q_1)}} \\
&\geq \left( P(x_1=0) + P(x_1=1)\sqrt{1+d_{\chi^2}(P_1, Q_1)} \right)^2 \tag{4.17}
\end{aligned}$$

$$\begin{aligned}
&\implies -1 + \sum_{x_1, \dots, x_n} \frac{P^2(x_1, \dots, x_{n-1}, x_n)}{Q_s(x_1, \dots, x_{n-1}, x_n)} \\
&= -1 + \sum_{x_1, \dots, x_{n-1}} \frac{P^2(x_n=0)P^2(x_1, \dots, x_{n-1} | x_n=0)}{Q_s(x_n=0)Q_s(x_1, \dots, x_{n-1} | x_n=0)} \\
&\quad + \sum_{x_1, \dots, x_{n-1}} \frac{P^2(x_n=1)P^2(x_1, \dots, x_{n-1} | x_n=1)}{Q_s(x_n=1)Q_s(x_1, \dots, x_{n-1} | x_n=1)} \\
&= -1 + \frac{P^2(x_n=0)}{Q_s(x_n=0)} (1 + d_{\chi^2}(P_0, Q_{0,s})) + \frac{P^2(x_n=1)}{Q_s(x_n=1)} (1 + d_{\chi^2}(P_1, Q_{1,s})) \\
&\geq -1 + \frac{P^2(x_n=0)}{Q_s(x_n=0)} + \frac{P^2(x_n=1)}{Q_s(x_n=1)} (1 + d_{\chi^2}(P_1, Q_{1,s})) \\
&\geq -1 + \left( P(x_n=0) + P(x_n=1)\sqrt{1+d_{\chi^2}(P_1, Q_{1,s})} \right)^2 \tag{4.18} \\
&= -1 + \left( (1 - \varepsilon_0) + \varepsilon_0\sqrt{1+d_{\chi^2}(P_1, Q_{1,s})} \right)^2 \\
&= -1 + \left( 1 + \varepsilon_0 \left( \sqrt{1+d_{\chi^2}(P_1, Q_{1,s})} - 1 \right) \right)^2 \\
&= 2\varepsilon_0 \left( \sqrt{1+d_{\chi^2}(P_1, Q_{1,s})} - 1 \right) + \left( \varepsilon_0 \left( \sqrt{1+d_{\chi^2}(P_1, Q_{1,s})} - 1 \right) \right)^2
\end{aligned}$$

$$\geq 2\varepsilon_0 \left( \sqrt{1 + d_{\chi^2}(P_1, Q_{1,\mathbf{s}})} - 1 \right). \quad (4.19)$$

For any fixed  $Q_{\mathbf{s}}(x_1, \dots, x_n)$ , we can compute  $Q_{\mathbf{s}}(x_n = 0)$ ,  $Q_{\mathbf{s}}(x_n = 1)$ , and  $d_{\chi^2}(P_1, Q_{1,\mathbf{s}})$ ; in other words, they are also fixed, given  $\mathbf{s}$ . Then a lower bound can be obtained via a variation argument in (4.18) via (4.17). Connecting (4.16), and (4.19), we continue with the following expression,

$$\begin{aligned} \frac{R_{\chi^2}(m)}{\Pr[S]} &\geq \inf_{Q \in \mathcal{Q}} \mathbb{E}_{\substack{P \sim \pi \\ \mathbf{s} \sim P^{\otimes m}}} \left[ 2\varepsilon_0 (\sqrt{1 + d_{\chi^2}(P_1, Q_{1,\mathbf{s}})} - 1) | S \right] \\ &= \inf_{Q \in \mathcal{Q}} \mathbb{E}_{P_1 \sim \pi} \mathbb{E}_{\mathbf{s} \sim P^{\otimes m} | S} \left[ 2\varepsilon_0 \left( \sqrt{1 + d_{\chi^2}(P_1, Q_{1,\mathbf{s}})} - 1 \right) \right] \\ &= \inf_{Q \in \mathcal{Q}} \mathbb{E}_{x^h \sim U_{n-1}} \mathbb{E}_{\mathbf{s} \sim \tilde{U}_{n-1}^{\otimes m}} \left[ 2\varepsilon_0 \left( \sqrt{1 + d_{\chi^2}(P_1, Q_{1,\mathbf{s}})} - 1 \right) \right] \end{aligned} \quad (4.20)$$

$$= \inf_{Q \in \mathcal{Q}} \mathbb{E}_{\mathbf{s} \sim \tilde{U}_{n-1}^{\otimes m}} \mathbb{E}_{x^h \sim U_{n-1}} \left[ 2\varepsilon_0 \left( \sqrt{1 + d_{\chi^2}(P_1, Q_{1,\mathbf{s}})} - 1 \right) \right] \quad (4.21)$$

$$\geq \left( \mathbb{E}_{\mathbf{s} \sim \tilde{U}_{n-1}^{\otimes m}} \inf_{Q \in \mathcal{Q}} \mathbb{E}_{x^h \sim U_{n-1}} \left[ 2\varepsilon_0 \left( \sqrt{\sum_x \frac{P_1^2(x)}{Q_{1,\mathbf{s}}(x)}} - 1 \right) \right] \right) \quad (4.22)$$

$$\geq \left( \inf_{Q^* \in \Delta^{2^{n-1}}} \mathbb{E}_{x^h \sim U_{n-1}} \left[ 2\varepsilon_0 \left( \sqrt{\sum_x \frac{P_1^2(x)}{Q^*(x)}} - 1 \right) \right] \right) \quad (4.23)$$

$$\begin{aligned} &= \inf_{Q^* \in \Delta^{2^{n-1}}} \mathbb{E}_{\alpha \triangleq (\alpha_1, \dots, \alpha_{n-1}) \sim U_{n-1}} \left[ 2\varepsilon_0 \left( \sqrt{\frac{1}{Q^*(\alpha)}} - 1 \right) \right] \\ &= 2\varepsilon_0 \left( 2^{\frac{n-1}{2}} - 1 \right). \end{aligned} \quad (4.24)$$

Since we are only changing  $P_1$  (or  $x^h$ ) in the construction, we can replace  $P \sim \pi$  with  $P_1 \sim \pi$  (or  $x^h \sim U_{n-1}$ ); note that there is no sample with  $X_1 = 1$  in  $\mathbf{s}$ , and thus  $\mathbf{s}$  is merely samples drawn from uniform distribution with all their corresponding  $X_1 = 0$  and this is what we mean by  $\mathbf{s} \sim \tilde{U}_{n-1}^{\otimes m}$  in (4.20). Therefore,  $P_1$  or  $x^h$  is independent with  $\mathbf{s}$ , and we can swap the expectation in (4.21); and in (4.22), we lower bound the expectation as the learner can first observe the samples  $\mathbf{s}$  before choosing the algorithm from  $\mathcal{Q}$ . But in any case, it is fixed before the last expectation, and hence (4.23) follows. As we assume the learning algorithm  $Q$  is deterministic, for a fixed  $\mathbf{s}$ ,  $Q_{1,\mathbf{s}}$  is also fixed. We obtain (4.24) through Fact 4.2.

In the end, we have that

$$R_{\chi^2}(m) \geq 2\varepsilon_0 \left( 2^{\frac{n-1}{2}} - 1 \right) \Pr[S] \geq \varepsilon_0 2^{\frac{n}{2}} (1 - \varepsilon_0 m).$$

By setting  $\varepsilon_0 = \frac{2\varepsilon}{2^{n/2}}$ , we can see that if  $m \leq \frac{1}{4\varepsilon} 2^{\frac{n}{2}}$ , then  $R_{\chi^2}(m) \geq 2\varepsilon - \frac{4\varepsilon^2}{2^{n/2}} m \geq \varepsilon$ .  $\square$

## 5 Testing maximum in-degree of Bayes nets

**THEOREM 5.1.** *Given an unknown distribution  $P$ , and a maximum degree- $d$  graph  $G$  supported on  $\{0, 1\}^n$ , it takes at most  $O\left(\max\left(\frac{2^{n/2}}{\varepsilon^2}, \frac{2^d n^2 d \log(n)}{\varepsilon^2}\right)\right)$  i.i.d. samples to test whether  $d_H(P, G) = 0$  or  $d_H(P, G) \geq \varepsilon$ , with probability  $\geq 2/3$ .*

*Furthermore, testing whether  $P$  is Markov w.r.t. any max degree- $d$  graphs with success probability at least  $2/3$ , takes at most  $O\left(\max\left(\frac{2^{n/2}}{\varepsilon^2}, \frac{2^d n^2 d \log(n)}{\varepsilon^2}\right) \cdot \log(n^{dn})\right)$  samples.*

---

**Algorithm 3:** Testing  $P$  is a degree- $d$  DAG  $G$ 

---

**Input :** Sample access to distribution  $P$ , accuracy parameter  $\varepsilon$  and a degree- $d$  DAG  $G$ .

- 1 Learn  $P$  with  $\varepsilon$ ,  $G$  and  $\frac{2^d n^2 \log(2^d n)}{\varepsilon^2}$  samples via [Algorithm 2](#): obtaining an estimate  $\tilde{Q}$ , and an  $O(\varepsilon^2)$ -effective support set  $\mathcal{A}$  via [Algorithm 1](#).
  - 2 Draw a multiset  $S$  of  $\text{Poisson}(m)$  samples from  $P$ , where  $m = \frac{2^{n/2}}{\varepsilon^2}$ .
  - 3 Call [[DKW18](#), Algorithm 1] and return  $P_{\mathcal{A}}$ ,  $\tilde{Q}_{\mathcal{A}}$ ,  $S$ ,  $\varepsilon$ .
- 

*Proof.* We prove by analyzing [Algorithm 3](#). By the guarantee of the underlying tester in [[DKW18](#), Algorithm 1], it suffices to verify the following:

- *Soundness:* If  $d_H(P, G) \geq \varepsilon$  (it is far from any Bayes nets of graph  $G$ ), then  $d_H(P_{\mathcal{A}}, \tilde{Q}_{\mathcal{A}}) \geq \Omega(\varepsilon)$ ;
- *Correctness:* If  $d_H(P, G) = 0$ , then  $d_{\chi^2}(P_{\mathcal{A}}, \tilde{Q}_{\mathcal{A}}) \leq O(\varepsilon^2)$ ; and we have this from [Theorem 4.1](#).

Roughly speaking, we can pretend  $P$  and  $\tilde{Q}$  are supported only on  $\mathcal{A}$ , and since  $|\mathcal{A}| \leq 2^n$ ,  $O(\sqrt{2^n})/\varepsilon^2$  samples suffice for testing. For soundness, by [Theorem 4.1](#), we have

$$\tilde{Q}(\tilde{S}) \geq 1 - O(\varepsilon^2), \text{ and } P(\tilde{S}) \geq 1 - O(\varepsilon^2).$$

Since

$$d_H^2(P, Q) = d_H^2(P_{\mathcal{A}}, Q_{\mathcal{A}}) + d_H^2(P_{\bar{\mathcal{A}}}, Q_{\bar{\mathcal{A}}}), \text{ and } d_H^2(P_{\bar{\mathcal{A}}}, Q_{\bar{\mathcal{A}}}) \leq d_{\text{TV}}(P_{\bar{\mathcal{A}}}, Q_{\bar{\mathcal{A}}}) \leq \frac{1}{2}(P(\bar{\mathcal{A}}) + Q(\bar{\mathcal{A}})) \leq O(\varepsilon^2),$$

we have that  $d_H^2(P_{\mathcal{A}}, Q_{\mathcal{A}}) \geq \Omega(\varepsilon^2)$ .

Since it costs an extra  $O(\log(1/\delta))$  to amplify the success probability to  $1 - \delta$  for each test, we will run amplified accurate tests on all  $n^{O(dn)}$  possible maximum in-degree- $d$  graphs and follow up with a union bound of  $1 - \delta \cdot n^{O(dn)}$ . In particular, we set  $\delta = \frac{1}{n^{dn}}$ , which brings an additional  $O(\log(n^{dn}))$  factor to the overall sample complexity, and thus, it gives us a tester for maximum in-degree- $d$  graphs with sample complexity  $O\left(\max\left(\frac{2^{n/2}}{\varepsilon^2}, \frac{2^d n^2 d \log(n)}{\varepsilon^2}\right) \cdot \log(n^{dn})\right)$ .  $\square$

**5.1 Extending [Theorem 5.1](#) to TV distance** While our result in [Theorem 5.1](#) already implies a tester in  $d_{\text{TV}}$ , with our near-proper learner in  $d_{\chi^2}$  for bounded degree Bayes net, it also implies a similar graphical tester in  $d_{\text{TV}}$  analogous to [Theorem 5.1](#), where the shifting of masses is unnecessary (see [[ADK15](#), Remark 1]), i.e., the additional requirement of  $Q(\tilde{S}) \geq 1 - O(\varepsilon^2)$  is no longer necessary in the case of TV; and we can also weaken requirement on  $P(\tilde{S})$ :  $P(\tilde{S}) \geq 1 - O(\varepsilon)$ .

To see this, we only need to verify that  $d_{\text{TV}}(P_{\mathcal{A}}, Q_{\mathcal{A}}) \geq \Omega(\varepsilon)$  in the case of soundness. Assuming that  $d_{\text{TV}}(P, Q) > 10\varepsilon$  and  $P(S) > 1 - \varepsilon$ , we analyze the two cases,

- When  $Q(S) < 1 - 2\varepsilon$ , we have  $P(S^c) \leq \varepsilon$ ,  $Q(S^c) > 2\varepsilon$ , and thus

$$\frac{1}{2} \sum_{i \in S} |P_i - Q_i| \geq \frac{1}{2} \left| \sum_{i \in S} (P_i - Q_i) \right| = \frac{1}{2} (P(S) - Q(S)) > \frac{1}{2} (1 - \varepsilon - (1 - 2\varepsilon)) = \frac{\varepsilon}{2}.$$

- When  $Q(S) < 1 - 2\varepsilon$ , similarly,

$$\frac{1}{2} \sum_{i \in S} |P_i - Q_i| = \frac{1}{2} \sum_{i \in \Omega} |P_i - Q_i| - \frac{1}{2} \sum_{i \notin S} |P_i - Q_i| > \frac{1}{2} (10\varepsilon - (\varepsilon + 2\varepsilon)) = \frac{7}{2}\varepsilon.$$

In both cases, we have  $d_{\text{TV}}(P_S, Q_S) \geq \Omega(\varepsilon)$ . Nevertheless, we note that the same technique does not work for Hellinger, and thus requires a slightly stronger guarantee.

## 6 Conclusion and future directions

In this paper, we provided (nearly) tight sample complexity bounds for testing the maximum in-degree of an unknown Bayesian network. Along the way, we obtained several results of independent interest, including a near-proper learner for Bayesian networks under  $\chi^2$  divergence, and a high-probability  $\chi^2$  learning algorithm (for arbitrary discrete distributions).

Our results raise two interesting future directions. The first is to generalize our testing result to the more general question of maximum degree- $d$  testing *under maximum degree- $k$  assumption*, where  $k > d$  are both input parameters; in particular, our results correspond to  $k = n$ . The second is to either strengthen our high-probability  $\chi^2$  learning bound to obtain an *additive*  $\log(1/\delta)$  dependence (as is known for total variation distance learning), instead of a multiplicative one; or to show that such a multiplicative dependence on  $\log(1/\delta)$  is necessary. We note that such a result is not known even for the weaker KL divergence learning.

## Acknowledgment

Yang would like to thank Philips George John for the helpful discussions, and for suggesting the  $\delta$  function notation in the lower bound analysis.

## References

- [ABDK18] Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions. In *NeurIPS*, pages 9469–9481, 2018. 2
- [ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *NIPS*, pages 3591–3599, 2015. 3, 6, 17
- [BBC<sup>+</sup>20] Ivona Bezáková, Antonio Blanca, Zongchen Chen, Daniel Stefankovic, and Eric Vigoda. Lower bounds for testing graphical models: Colorings and antiferromagnetic ising models. *J. Mach. Learn. Res.*, 21:25:1–25:62, 2020. 2
- [BCY22] Arnab Bhattacharyya, Clément L. Canonne, and Joy Qiping Yang. Independence testing for bounded degree bayesian network. *CoRR*, abs/2204.08690, 2022. 2, 3
- [BGPV21] Arnab Bhattacharyya, Sutanu Gayen, Eric Price, and N. V. Vinodchandran. Near-optimal learning of tree-structured distributions by Chow–Liu. In *STOC*, pages 147–160. ACM, 2021. 3, 4
- [BV14] Stephen P. Boyd and Lieven Vandenbergh. *Convex Optimization*. Cambridge University Press, 2014. 14
- [Can20] Clément L. Canonne. A short note on learning discrete distributions, 2020. 3
- [CDDK22] Davin Choo, Yuval Dagan, Constantinos Daskalakis, and Anthimos Vardis Kandiros. Learning and testing latent-tree ising models efficiently. *CoRR*, abs/2211.13291, 2022. 2
- [CDKL22] Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Sihan Liu. Near-optimal bounds for testing histogram distributions. *CoRR*, abs/2207.06596, 2022. 3
- [CDKS17] Clément L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Testing bayesian networks. In *Conference on Learning Theory*, pages 370–448. PMLR, 2017. 1, 2
- [DDK17] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Concentration of multilinear functions of the ising model with applications to network data. In *NIPS*, pages 12–23, 2017. 2
- [DDK19] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing ising models. *IEEE Trans. Inf. Theory*, 65(11):6829–6852, 2019. 2
- [Dia16] Ilias Diakonikolas. Learning structured distributions. In *Handbook of big data*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pages 267–283. CRC Press, Boca Raton, FL, 2016. 3
- [DKW18] Constantinos Daskalakis, Gautam Kamath, and John Wright. Which distribution distances are sublinearly testable? In *SODA*, pages 2747–2764. SIAM, 2018. 3, 6, 17
- [DL01] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2001. 3
- [DP17] 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. [2](#)
- [FLNP00] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe’er. Using bayesian networks to analyze expression data. In *Proceedings of the fourth annual international conference on Computational molecular biology*, pages 127–135, 2000. [1](#)
- [GNS18] Aditya Gangrade, Bobak Nazer, and Venkatesh Saligrama. Two-sample testing can be as hard as structure learning in ising models: Minimax lower bounds. In *ICASSP*, pages 6931–6935. IEEE, 2018. [2](#)
- [Hec98] David Heckerman. *A tutorial on learning with Bayesian networks*. Springer, 1998. [1](#)
- [KOPS15] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 1066–1100. JMLR.org, 2015. [3](#), [6](#), [14](#)
- [NL19] Matey Neykov and Han Liu. Property testing in high-dimensional Ising models. *Ann. Statist.*, 47(5):2472–2503, 2019. [2](#)
- [Pea88] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988. [1](#)
- [Pea95] Judea Pearl. From bayesian networks to causal networks. *Mathematical models for handling partial knowledge in artificial intelligence*, pages 157–182, 1995. [1](#)
- [SPP+05] Karen Sachs, Omar Perez, Dana Pe’er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005. [1](#)
- [WJ08] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008. [1](#)