Entropy testing and its application to testing Bayesian networks

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

Joy Qiping Yang
University of Sydney
qyan6238@uni.sydney.edu.au

Abstract

This paper studies the problem of *entropy identity testing*: given sample access to a distribution p and a fully described distribution q (both are discrete distributions over the support of size k), and the promise that either $p = q$ or $|H(p)-H(q)| \geqslant \varepsilon$, where $H(\cdot)$ denotes the Shannon entropy, a tester needs to distinguish between the two cases with high probability. We establish a near-optimal sample complexity bound of $\tilde{\Theta}(\sqrt{k}/\varepsilon + 1/\varepsilon^2)$ for this problem, and show how to apply it to the problem of identity testing for in-degree-d n-dimensional Bayesian networks, obtaining an upper bound of $\tilde{O}\big(2^{d/2}n/\varepsilon^2 + n^2/\varepsilon^4\big)$. This improves on the sample complexity bound of $\tilde{O}(2^{d/2}n^2/\varepsilon^4)$ from [CDKS20], which required an additional assumption on the structure of the (unknown) Bayesian network.

1 Introduction

Entropy is a fundamental information theory notion, which quantifies the amount of "uncertainty" a given random variable carries. Since its introduction by Shannon, this notion has found myriads of applications, and is central – among others – to compression and coding, probability, electrical engineering, and learning theory.

As a result, the task of *estimating* the Shannon entropy of a discrete random variable (or, equivalently, its probability distribution) from samples has naturally emerged, starting (in Computer Science) with the work of [BDKR02] which considered *multiplicative* approximations. *Additive* approximation of the entropy (within $\pm\varepsilon$) was then considered in a series of papers [VV11a, VV11b, VV13, HJW15a, ADOS17], culminating with the work of [WY16], which establishes the optimal sample complexity, $\Theta\big(\frac{k}{\varepsilon \log k} + \frac{\log^2 k}{\varepsilon^2}\big)$,[1] where $k \gg 1$ is the domain size.

While the resulting sample complexity is *sublinear* in the domain size k, it is only so by a mere logarithmic factor. In some settings, paying this near-linear dependence in the amount of data necessary is impractical, typically in the large-domain regime (e.g., for high-dimensional data, where k is exponential in the dimension); moreover, it may even be *unnecessary*. Specifically, one may not be concerned so much about the (approximate) value of the entropy of a distribution, but rather about whether it is above a threshold, or differs from that of a given purported model.

It is this latter task we introduce and consider in our work, which can be seen as a variant of the standard *identity testing* question from distribution testing: given a reference known hypothesis distribution q over a domain of size k, and i.i.d. samples from an unknown distribution p, what is the sample complexity of testing whether p is equal to q, or their entropies differ significantly? And, crucially, *is this testing task more sample-efficient than that of estimating $H(p)$?*

[1] All logarithms in the paper are natural (e as base).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Entropy Identity testing: Given a reference distribution q, parameter $\varepsilon > 0$, and samples from an unknown p, what is the cost of deciding (with high probability) whether $p = q$ vs. $|H(p) - H(q)| > \varepsilon$, with correct probability at least $2/3$?

Note that in the case where q is the uniform distribution over the domain, this task is equivalent to distinguishing between $H(p) = \log k$ and $H(p) < \log k - \varepsilon$.

Our main contribution is to show that the testing question can indeed be performed much more efficiently than the estimation one, at least for most parameter regimes. Specifically, we establish the following theorem:

Theorem 1.1. *The sample complexity of entropy identity testing is* $O\left(\sqrt{k \log(k/\varepsilon)}/\varepsilon + \log^2(k)/\varepsilon^2\right)$. *Moreover, this is nearly tight:* $\Omega\left(\sqrt{k}/\varepsilon + \log^2 k/\varepsilon^2\right)$ *samples are necessary in the worst case.*

Interestingly, this differs both from the *estimation* task (which, as discussed before, has a near-linear dependence on the domain size k) but also from identity testing *in total variation distance*, which has sample complexity $\Theta(\sqrt{k}/\varepsilon^2)$ (see Section 1.1).

Application: Identity testing for Bayesian networks. As an application of Theorem 1.1, we derive an efficient algorithm for identity testing (in total variation distance) for maximum in-degree d Bayesian networks (shorten as degree-d Bayes net in the remaining of the paper):[2]

Theorem 1.2 (Informal; see Theorem 3.1)**.** *There is an algorithm which, given sample access to a degree-d Bayes net p and the full description of a reference degree-d Bayes net q (both over $\{0,1\}^n$), takes* $\tilde{O}\left(\frac{2^{d/2}n}{\varepsilon^2} + \frac{n^2}{\varepsilon^4}\right)$ *samples from p, and distinguishes between $p = q$ and $d_{\mathrm{TV}}(p,q) \geq \varepsilon$.*

Prior to this, the best known sample complexity upper bound for this task [CDKS20] was quadratically worse in both n and ε, and further required an assumption on the underlying graph structure of both p and q. We emphasize that (1) our result improves on the sample complexity of the learning baseline for $d \gg \log(n/\varepsilon)$, and on its computational efficiency; and (2) compared to the previous testing results, removes strong structural assumptions which considerably limited their applicability. We elaborate on this in the next section.

1.1 Related work

As previously discussed, entropy estimation has received a considerable amount of interest from computer scientists, information theorists and statisticians [BDKR02, Pan04, HJW15a, WY16]. Entropy is also a key example of *symmetric property* (invariant to relabeling of the domain) [VV11a, VV11b, VV13, ADOS17], and has been considered in other settings as well, e.g., the quantum case [GHS21, AISW20] and the memory-limited setting [ABIS19, AMNW22]. Estimation of some generalizations of Shannon entropy, such as the family of Rényi entropies, also have been studied [AOST17].

Over the years, sample complexity of identity testing for discrete distribution has been intensively studied and essentially settled [Pan08, BFF$^+$01, VV17]. In high dimensions, however, the square root dependence of the sample complexity on the domain size means that most identity testing tasks of interest require sample complexity exponential in the dimension. Moreover, this curse of dimensionality extends to a large range of distribution testing problems [BCY22, Theorem B.1]. As such, many turn to the study of testing distributions under additional natural structural assumptions, such as graphical models: [BGKV21] look at identity testing for product distributions (degree-0 Bayes nets) and give the optimal bound of $\Theta(\sqrt{n|\Sigma|/\varepsilon^2})$, where $|\Sigma|$ is the alphabet size of each variable (rather than binary alphabet studied in our paper). [DDK19, KDDC23] study testing Ising models, obtaining sample complexity bounds that are $\mathrm{poly}(n/\varepsilon)$; [DP16], [CDKS20] give tight results to identity testing and closeness testing for a variety of constant in-degree Bayes nets, which also gives polynomial sample complexity bounds.

However, the testing algorithms provided in [CDKS20] and [DP16] are not fully satisfactory, as they require some strong assumptions on Bayes nets. Specifically, [CDKS20, Theorem 21] assumes that

[2] Our algorithm actually provides a stronger guarantee, with respect to Hellinger distance, which implies the TV result.

the topological ordering of the two Bayes nets are the same, and shows that under this assumption $O(2^{d/2}n^2/\varepsilon^4)$ samples are sufficient.[3] [CDKS20, Theorem 17] makes the further stringent restriction that the reference Bayes net has to be *balanced*, i.e., that the conditional probabilities are all bounded away from 0 and 1; moreover, it also requires every parental configuration to be bounded from 0, and that the structure of the unknown Bayes net be a subset of that of the reference one. The result of [DP16, Theorem 4.2] combined with the Hellinger tester from [DKW18, Theorem 1] implies that, under the assumption that p and q share the same factorization structure (i.e., their associated DAGs are the same or one is a subgraph of the other), then this problem is solvable in $\tilde{O}\left(2^{d/2}n/\varepsilon^2\right)$ samples. While this latter sample complexity is near-optimal (in some regime[4]), in view of the $\Omega\left(2^{d/2}n/\varepsilon^2\right)$ lower bound obtained in [BCY22, Theorem 4.1], the factorization structure requirement considerably limits the applicability of the algorithm.

One can also compare our result to the *learning* results on Bayesian networks, as any learning algorithms enables testing as well (the "testing-by-learning" baseline). It is known [CDKS20]) that learning degree-d Bayes nets can be done with $\tilde{O}(2^d n/\varepsilon^2)$ samples, without any structural assumptions. Our testing result improves on this sample complexity as long as $n^2/\varepsilon^4 \ll 2^d n/\varepsilon^2$, i.e., for $d \gg \log(n/\varepsilon)$; moreover, it is worth noting that the known learning algorithms are computationally inefficient (running in time $n^{O(dn)}$ via an enumeration of all possible underlying graph structures [CDKS20, BGMV20]), and this is believed to be inherent [CHM04]. In contrast, our algorithm runs in time $\mathrm{poly}(n^d, 1/\varepsilon)$.

1.2 Techniques overview

Testing in entropy. A first idea is to use the conversion between total variation (TV) distance and entropy difference to reduce this problem to identity testing in TV: When $d_{\mathrm{TV}}(p,q) \leqslant 1/2$, then $|H(p) - H(q)| \leqslant d_{\mathrm{TV}}(p,q)\log\frac{k}{d_{\mathrm{TV}}(p,q)}$ [CK11, Lemma 2.7]. This gives an upper bound of $O(\frac{\sqrt{k}\log^2(k/\varepsilon)}{\varepsilon^2})$, which is already better than the sample complexity of estimation: $O\left(\frac{k}{\varepsilon\log k} + \frac{\log^2 k}{\varepsilon^2}\right)$ for the parameter k. However, it is not clear whether the quadratic dependence on ε is necessary: indeed, the "hard instances" for TV testing (the Paninski construction [Pan08]), small perturbations around the uniform distribution which have TV distance ε from uniform, actually only have entropy $\log k - \Theta(\varepsilon^2)$. The $\Omega(\sqrt{k}/\varepsilon^2)$ uniformity testing lower bound from these hard instances thus only implies an $\Omega(\sqrt{k}/\varepsilon)$ entropy identity testing lower bound!

A next natural idea is to strengthen the lower bound. However, it then becomes clear that the Paninski [Pan08] construction cannot be improved: as just mentioned, when its TV distance to the uniform distribution is around $\Theta(\sqrt{\varepsilon})$ its entropy difference to it is only $\Theta(\varepsilon)$ (giving an $\Omega(\sqrt{k}/\varepsilon)$ lower bound). Moreover, this is not a coincidence: when the reference distribution q is uniform, we are able to get a matching upper bound using [DKW18, Algorithm 1], upon noticing that

$$H(p) = \log k - d_{\mathrm{KL}}(p\|u_k), \tag{1}$$

which implies $d_{\mathrm{KL}}(p\|u_k) = \log k - H(p) \geqslant \varepsilon$, where u_k is the uniform distribution on $[k]$ and d_{KL} denotes the Kullback–Leibler divergence. Interestingly, a completely different hard instance, against a very much non-uniform reference distribution, does yield the second term of our lower bound, $\Omega(\log^2 k/\varepsilon^2)$.

Inspired by these two different lower bounds, we can generalize (1) by defining $\mathcal{A}$ as the set of "not too small probability elements under q", and then observing (looking ahead, using the inequality (12)) that

$$|H(p_{\mathcal{A}}) - H(q_{\mathcal{A}})| \leqslant |d_{\mathrm{KL}}(p_{\mathcal{A}}\|q_{\mathcal{A}})| + \left|\sum_{i\in\mathcal{A}}(p_i - q_i)\log\frac{1}{q_i}\right| \tag{2}$$

where $H(p_{\mathcal{A}})$ is the "entropy" of the sub-distribution restricted to the set $\mathcal{A}$. In particular, this hints that one could solve the general problem by testing if either of the two terms on the right-hand-side is large. The name of the game now is to (i) choose the threshold for $\mathcal{A}$ (i.e., what does it mean

[3] While the sample complexity of the algorithm is not explicitly stated in their proof, inspection of their argument yields this bound.

[4] The lower bound [BCY22, Theorem 4.1] only holds under the sparse regime: $d \ll \log n$.

for an element to have "not too small probability under q"), and (ii) have algorithms to test whether these two quantities are noticeably large.

Let us focus on how to test the first term of (2). If $\min_i q_i \geqslant \Omega\left(\frac{\varepsilon}{k}\right)$, we can adapt and use an algorithm of [DKW18] to efficiently test $d_{\mathrm{KL}}(p\|q) \geqslant \varepsilon$ vs. $p = q$. In addition, if $\log(1/q_i)$ is bounded, then in fact, estimating the second term to $O(\varepsilon)$ is possible as well. Thus it is natural to wonder if we can afford to neglect the region where $q_i \leqslant \frac{\varepsilon}{k}$. Indeed, the impact on entropy is at most $O(\tau \log(k/\tau))$ if we are to remove regions with at most $O(\tau)$ as mass. Thus, by adjusting the appropriate threshold, we can still detect difference in entropy even if we only test on elements with greater than τ/k masses, where $\tau = \frac{\varepsilon}{\log(k/\varepsilon)}$.

The problem then becomes to check if p puts more than $100 \cdot \tau$ mass in $\bar{\mathcal{A}} = \{i \in [k] : q_i < \tau/k\}$, which costs $O(1/\tau) = O(\log(k/\varepsilon)/\varepsilon)$ samples. If it does, then it cannot be the case that $p = q$; we can reject. After this stage, both $p(\bar{\mathcal{A}}), q(\bar{\mathcal{A}}) \leqslant O(\tau)$. To move forward, we need to check the influence on entropy: $H(p)$ and $H(q)$. By Jensen's inequality and monotonicity of $f(x) = x \log \frac{1}{x}$ when $x < \frac{1}{e}$, we have

$$\sum_{i \in \bar{\mathcal{A}}} p_i \log \frac{1}{p_i} \leqslant p(\bar{\mathcal{A}}) \log \frac{k}{p(\bar{\mathcal{A}})} \leqslant \tau \log \frac{k}{\tau}.$$

Therefore, the impact on entropy will be at most $O\left(\tau \log \frac{k}{\tau}\right)$. Setting $\tau = \frac{\varepsilon}{\log(k/\varepsilon)}$, this becomes $O(\varepsilon)$, which gives us the room to check if $|H(p_{\mathcal{A}}) - H(p_{\mathcal{A}})| \geqslant 100\varepsilon$ or $p_{\mathcal{A}} = q_{\mathcal{A}}$.

Testing Bayesian networks. Similar to [DP16, Theorem 4.2],[5] the identity testing algorithm is straight-forward: check every possible subset of size $d + 1$ if $p_L = q_L$ or one of them is far apart, where p_L denotes the marginalization of p over subset L.[6] If $p = q$, then obviously, all such tests will accept with high probability; and by subadditivity of squared Hellinger distance, this will guarantee that for any DAG, and a projection p_G (resp. q_G) of p (resp. q) unto G (see, [DP16, Corollary 2.4]):

$$\mathrm{d}_{\mathrm{H}}(p_G, q_G) \leqslant \varepsilon.$$

While it seems unintuitive that this method does not work without the common-structure assumption, directly extending this argument does not work. Indeed, if they (p and q) do not share a common structure, then it is clear that $p \neq q$; but the tricky part is the farness case, when $\mathrm{d}_{\mathrm{H}}(p, q) \geqslant \varepsilon$ and p, q does not share a common structure; the tester could still accept if every $\mathrm{d}_{\mathrm{H}}(p_L, q_L) \leqslant \varepsilon/\sqrt{n}$, which only guarantees that $\mathrm{d}_{\mathrm{H}}(p_G, q_G) \leqslant \varepsilon$ for every G. Our fix to the problem is to additionally check if there exists a common graph $\tilde{G}$ where $\mathrm{d}_{\mathrm{H}}(p, p_{\tilde{G}})$ and $\mathrm{d}_{\mathrm{H}}(q, q_{\tilde{G}})$ are close. If it does, via triangular inequality, we can in fact show that

$$\mathrm{d}_{\mathrm{H}}(p, q) \leqslant \mathrm{d}_{\mathrm{H}}(q, q_{\tilde{G}}) + \mathrm{d}_{\mathrm{H}}(q_{\tilde{G}}, p_{\tilde{G}}) + \mathrm{d}_{\mathrm{H}}(p, p_{\tilde{G}}) \leqslant O(\varepsilon).$$

This handles the last possible issue. To this end, we will conduct local entropy identity test between p and q for subset of size $d + 1$ and d. The idea is that if all tests pass, then we can conclude that p and q are close in local entropy and thereafter, we can utilize entropy of q to learn the graphical structure [KCG+23] of p (which uses no additional samples).

Preliminaries and notation. The (Shannon) entropy H of a discrete distribution p supported on $[k]$ is given by:

$$H(p) = -\sum_{i \in [k]} p_i \log p_i.$$

The conditional entropy $H(p_X \mid p_Y)$ for X supported on $\mathcal{X}$, and Y on $\mathcal{Y}$, defined by the joint distribution $p_{X,Y}$, can be written as

$$H(p_X \mid p_Y) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(y)} = H(p_{X,Y}) - H(p_Y). \tag{3}$$

We adopt the entropy notation for a sub-probability vector $H(q_{\mathcal{A}}) = \sum_{i \in \mathcal{A}} q_i \log \frac{1}{q_i}$. Throughout this paper, we will use e as base of the log and of the entropy. We will use $\leftarrow$ for variable assignment.

[5] We note that what they refer to as "identity testing" is different from ours (and the standard) use of the term: in their setting, the reference distribution is replaced with sample access to the distribution (this is commonly referred to as "closeness testing").

[6] $L = \{X_{\sigma_1}, \ldots, X_{\sigma_{d+1}}\} \subset \{X_1, \ldots, X_n\}$; $p_L = p_{X_{\sigma_1}, \ldots, X_{\sigma_{d+1}}}$ is the marginalization of X on L.

We adopt the standard $O(\cdot)$, $\Omega(\cdot)$ and $\Theta(\cdot)$ asymptotic notation and use $\widetilde{\cdot}$ to hide any polylogarithmic factors in the argument. We will use various metrics or divergences on probability distributions: Kullback–Leibler (d_{KL}), Hellinger (d_{H}), chi-squared (d_{χ^2}), and total variation (d_{TV}). We denote $p_{\mathcal{A}}$ as restricting p onto the elements in $\mathcal{A}$, and we denote distributional distances restricting on $\mathcal{A}$ as follows: $d_{\mathrm{KL}}(p_{\mathcal{A}}, q_{\mathcal{A}}) = \sum_{i\in\mathcal{A}} p_i \log \frac{p_i}{q_i}$. $\mathrm{d}_{\mathrm{H}}(p_{\mathcal{A}}, q_{\mathcal{A}}) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i\in\mathcal{A}} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2}$. For a set $\mathcal{A}$, we write $p(\mathcal{A}) = \sum_{i\in\mathcal{A}} p_i$. We also have the following inequality [DKW18, Proposition 1]:

$$d_{\mathrm{TV}}(p_{\mathcal{A}}, q_{\mathcal{A}}) \leqslant \sqrt{2}\,\mathrm{d}_{\mathrm{H}}(p_{\mathcal{A}}, q_{\mathcal{A}}) \leqslant \sqrt{\sum_{i\in\mathcal{A}}(q_i - p_i) + d_{\mathrm{KL}}(p_{\mathcal{A}}, q_{\mathcal{A}})} \leqslant \sqrt{d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}})}. \tag{4}$$

A distribution p supported over the hypercube $\{0,1\}^n$ is a Bayesian network if its probability mass function satisfies the factorization associated with G, a directed acyclic graph (DAG):

$$p(x_1, \cdots, x_n) = \prod_{i=1}^{n} p(x_i|\Pi_i), \tag{5}$$

and Π_i is the set of parents of X_i in G; and we say that p is Markov with respect to DAG G. In section 3, slightly abusing notation, we use p_G to denote a projection of a Bayes net p to a DAG G (which it may or may not be Markov with respect to; see Definition 3.2). We work in the Poissonized setting (see, e.g., [Can22, Appendix C]) – instead of drawing N samples directly from p, we draw $Y \sim \mathrm{Poi}(N)$ samples from p, where $\mathrm{Poi}(N)$ denotes the random variable distributed as the Poisson distribution with parameter N. The Poissonized and usual sampling settings are equivalent for constant probability of failure, up to a (small) multiplicative factor in the sample complexity.

2 Near-optimal entropy testing

We prove Theorem 1.1, establishing the sample complexity upper and lower bounds separately.

2.1 An $O\big(\frac{\sqrt{k\log(k/\varepsilon)}}{\varepsilon} + \frac{\log^2(k)}{\varepsilon^2}\big)$ upper bound

We will prove the following theorem:

Theorem 2.1. *There is an algorithm (Algorithm 1) which, given n samples from a discrete distribution p, the full description of a reference distribution q, both over $[k]$, and parameter $\varepsilon > 0$, distinguishes between $p = q$ and $|H(p) - H(q)| \geqslant \varepsilon$ with probability at least $2/3$, as long as*

$$n \geq c_1 \left(\frac{\sqrt{k\log(k/\varepsilon)}}{\varepsilon} + \frac{\log^2(k)}{\varepsilon^2}\right)$$

and $c_2\varepsilon \leqslant k$, for some absolute constants $c_1, c_2 > 0$. Moreover, the algorithm runs in time linear in the number of samples n and the domain size k.

The proof will rely on the two following claims and Lemma 2.4, which is a straightforward extension of [DKW18, Lemma 2]. Their proofs are deferred to Appendix B. Throughout, we let $\tau := \frac{\varepsilon}{16\log(k/\varepsilon)}$, and $\mathcal{A} := \left\{i \in [k] \mid q_i \geqslant \frac{\tau}{k}\right\}$, as in Algorithm 1.

Claim 2.2. *Let $\mathcal{A}$ be any set such that $p(\bar{\mathcal{A}}) < \varepsilon/2$. Then, if $|H(p_{\mathcal{A}}) - H(q_{\mathcal{A}})| \geqslant \varepsilon$, we must have (i) $d_{\mathrm{KL}}(p_{\mathcal{A}}\|q_{\mathcal{A}}) \geqslant \frac{\varepsilon}{2}$ or (ii) $|\sum_{i\in\mathcal{A}}(p_i - q_i)\log(\frac{1}{q_i})| \geqslant \frac{\varepsilon}{2}$.*

Claim 2.3. *Let $\hat{p}$ be the empirical estimator for an unknown discrete distribution p supported on $[k]$, based on $\mathrm{Poi}(m)$ samples, where $m = \Theta\Big(\frac{\log^2(k)}{\varepsilon^2}\Big)$; assume that $d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) \leqslant \varepsilon/8$ and $p(\bar{\mathcal{A}}) + q(\bar{\mathcal{A}}) \leqslant 4\tau = \frac{1}{4}\frac{\varepsilon}{\log(k/\varepsilon)}$,[7] then*

$$\Pr\left[\left|\sum_{i\in\mathcal{A}}(p_i - \hat{p}_i)\log\frac{1}{q_i}\right| \geqslant \frac{1}{8}\varepsilon\right] \leqslant \frac{1}{100}.$$

[7] One can remove the assumption that $p(\bar{\mathcal{A}}) + q(\bar{\mathcal{A}}) \leqslant 4\tau$, at the cost of a slightly worse constant.

Algorithm 1 Entropy identity testing

Require: Sample access to p and full description of q, both over $[k]$; accuracy parameter ε.

1: Set $\tau := \frac{\varepsilon}{16\log(k/\varepsilon)}$, and $\mathcal{A} := \left\{i \in [k] \mid q_i \geqslant \frac{\tau}{k}\right\}$.
2: Take $m_1 = 48/\tau$ samples from p and compute the empirical $\hat{p}'$.
3: Compute $Z_1 = \hat{p}'(\bar{\mathcal{A}})$.
4: **if** $Z_1 \geqslant 2\tau$ **then return** reject ▷ Early rejection.
▷ N_i: the empirical count among samples of the i-th element.
5: Let $m_2 = 65536\left(\frac{\sqrt{k\cdot\log(k/\varepsilon)}}{\varepsilon}\right)$. Draw $\mathrm{Poi}(m_2)$ samples from p and compute

$$Z_2 = \sum_{i\in\mathcal{A}} \frac{(N_i - Nq_i)^2 - N_i}{Nq_i}.$$

6: **if** $Z_2 \geqslant \frac{1}{16}m_2\varepsilon$ **then return** reject
7: Let $m_3 = 140800\left(\frac{\log^2(k)}{\varepsilon^2}\right)$
8: Draw $\mathrm{Poi}(m_3)$ samples from p, compute the empirical $\hat{p}$; let $Z_3 \leftarrow \left|\sum_i (\hat{p}_i - q_i)\log\left(\frac{1}{q_i}\right)\right|$.
9: **if** $Z_3 \geqslant \frac{1}{8}\varepsilon$ **then return** reject
10: **return** accept

Lemma 2.4. *Let $\mathcal{A} := \{i \in [k] \mid q_i \geqslant \alpha\}$. Let $m_2 \geqslant 16384\max\left\{\sqrt{\frac{1}{\alpha\varepsilon}}, \frac{\sqrt{k}}{\varepsilon}\right\}$ be the number of samples used to compute Z_2. Then $\mathbb{E}[Z_2] = m_2 d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}})$. Moreover, if $d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) \leqslant \frac{\varepsilon}{2}$, then $\mathrm{Var}[Z_2] \leqslant (\frac{1}{32}m_2\varepsilon)^2$. If $d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) \geqslant \varepsilon$, then $\mathrm{Var}[Z_2] \leqslant O(\mathbb{E}[Z_2]^2)$.*

Proof of Theorem 2.1. We prove the statement by analyzing Algorithm 1. First, note that excluding the set of $\bar{\mathcal{A}}$ (elements with small mass), can change the value of $H(q)$ by at most $\varepsilon/8$: indeed, by Jensen's inequality ($f(x) = \log x$ is concave) and $x\log\frac{1}{x}$ being monotonically increasing in $(0, 1/e)$,

$$H(q_{\bar{\mathcal{A}}}) = \sum_{i\in\bar{\mathcal{A}}} q_i \log\frac{1}{q_i} \leqslant q(\bar{\mathcal{A}})\log\frac{|\bar{\mathcal{A}}|}{q(\bar{\mathcal{A}})} \leqslant \tau\log\frac{k}{\tau} = \frac{\varepsilon}{16\log(k/\varepsilon)}\log\left(\frac{16k}{\varepsilon/\log(k/\varepsilon)}\right) \leqslant \frac{1}{8}\varepsilon,$$

when $\tau \leqslant 1/e$. Similarly, if $p(\bar{\mathcal{A}}) \leqslant 3\tau$, we have that $H(p_{\bar{\mathcal{A}}}) \leqslant \frac{3}{8}\varepsilon$. Therefore,

$$\begin{aligned}\varepsilon \leqslant |H(p) - H(q)| &\leqslant |H(p_{\mathcal{A}}) - H(q_{\mathcal{A}})| + |H(p_{\bar{\mathcal{A}}}) - H(q_{\bar{\mathcal{A}}})| \\ &\leqslant |H(p_{\mathcal{A}}) - H(q_{\mathcal{A}})| + |H(p_{\bar{\mathcal{A}}})| + |H(q_{\bar{\mathcal{A}}})| \\ &\leqslant |H(p_{\mathcal{A}}) - H(q_{\mathcal{A}})| + \tfrac{1}{2}\varepsilon.\end{aligned}$$

For Line 4, we prove the following: with probability at least $99/100$, if $Z_1 \geqslant 2\tau$, then $p(\bar{\mathcal{A}}) \geqslant \tau$; and if $Z_1 < 2\tau$, then $p(\bar{\mathcal{A}}) < 3\tau$ (this is a standard technique; see e.g., [Can22, Fact 2.2].) For the sake of completeness we include the full derivation in the Appendix A.

After Line 4 of Algorithm 1. We conclude from the above that

i. $\mathcal{A}$ still has sufficient entropy gap to test on: $|H(p_{\mathcal{A}}) - H(q_{\mathcal{A}})| \geqslant \frac{1}{2}\varepsilon$.

ii. With probability at least $99/100$, when $p = q$, it will not be rejected in Algorithm 4 of Line 4; and once it is pass through this stage, we have $p(\bar{\mathcal{A}}) \leqslant 3\tau$.

Completeness: when $p = q$.

- We have that $d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) = 0$, and via Lemma 2.4, we know that $\mathbb{E}[Z_2] = 0$ and $\mathrm{Var}[Z_2] \leqslant \frac{1}{32^2}m_2^2\varepsilon^2$. By Chebyshev's inequality,

$$\Pr\left[|Z_2 - \mathbb{E}[Z_2]| \geqslant 2\sqrt{\mathrm{Var}[Z_2]}\right] \leqslant \frac{1}{4}, \quad \text{and so} \quad \Pr[Z_2 \geqslant 2\cdot\frac{1}{32}m_2\varepsilon + \mathbb{E}[Z_2]] \leqslant \frac{1}{4};$$

and we have $\Pr[Z_2 \geqslant \frac{1}{16}m_2\varepsilon] \leqslant \frac{1}{4}$.

- On the other hand, by Claim 2.3, setting $m_3 = \frac{140800 \log^2(k)}{\varepsilon^2}$, we have that with probability at least $99/100$,

$$Z_3 = \left| \sum_{i \in \mathcal{A}} (\hat{p}_i - q_i) \log \frac{1}{q_i} \right| = \left| \sum_{i \in \mathcal{A}} (\hat{p}_i - p_i) \log \frac{1}{q_i} \right| \leqslant \frac{1}{8}\varepsilon.$$

Therefore, with probability at least $1 - \frac{1}{4} - \frac{2}{100} = \frac{73}{100} > \frac{2}{3}$, the tester will accept.

Soundness: when $|H(p) - H(q)| \geqslant \varepsilon$. If $p(\bar{\mathcal{A}}) \geqslant 3\tau$ then $\hat{p}(\bar{\mathcal{A}}) \geqslant 2\tau$ with probability $99/100$, and the algorithm will output Reject. We proceed assuming $p(\bar{\mathcal{A}}) \leqslant 3\tau$ and recall Item ii. from before, we have $|H(p_{\mathcal{A}}) - H(q_{\mathcal{A}})| \geqslant \frac{1}{2}\varepsilon$. By Claim 2.2, we have that either $d_{\mathrm{KL}}(p_{\mathcal{A}}, q_{\mathcal{A}}) \geqslant \frac{1}{4}\varepsilon$ or $\left|\sum_{i \in \mathcal{A}} (p_i - q_i) \log(1/q_i)\right| \geqslant \frac{1}{4}\varepsilon$. We apply Lemma 2.4, setting $\alpha = \tau/k$ and $m_2 \geqslant 65536\sqrt{k \log(k/\varepsilon)}/\varepsilon$.

- If $d_{\mathrm{KL}}(p_{\mathcal{A}} \| q_{\mathcal{A}}) \geqslant \frac{1}{4}\varepsilon$, with (4) and $\exp(3/2) \leqslant k/\varepsilon$, we have

$$\frac{1}{8}\varepsilon \leq -3\tau + d_{\mathrm{KL}}(p_{\mathcal{A}}, q_{\mathcal{A}}) \leqslant \sum_{i \in \mathcal{A}} (q_i - p_i) + d_{\mathrm{KL}}(p_{\mathcal{A}}, q_{\mathcal{A}}) \leqslant d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}),$$

  which by Lemma 2.4, and our setting of m_2 and α, implies $\mathrm{Var}[Z_2] \leqslant (\frac{1}{4}\mathbb{E}[Z_2])^2$ and $\mathbb{E}[Z_2] = m_2 \cdot d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) \geqslant \frac{1}{8} m_2 \varepsilon$. By Chebyshev,

$$\Pr\left[|Z_2 - \mathbb{E}[Z_2]| \geqslant 2\sqrt{\mathrm{Var}[Z_2]}\right] \leqslant \frac{1}{4} \text{ and so } \Pr[Z_2 \leqslant \frac{1}{16} m_2 \varepsilon] \leqslant \frac{1}{4}.$$

- On the other hand, if it is the case that $\left|\sum_{i \in \mathcal{A}} (p_i - q_i) \log(1/q_i)\right| \geqslant \frac{1}{4}\varepsilon$, by Claim 2.3, setting $m_3 = 140800 \log^2(k)/\varepsilon^2$, with probability at least $99/100$,

$$\begin{aligned} \frac{1}{4}\varepsilon &\leqslant \left| \sum_i p_i \log \frac{1}{q_i} - q_i \log \frac{1}{q_i} \right| \\ &\leqslant \left| \sum_i (p_i - \hat{p}_i) \log \frac{1}{q_i} \right| + \left| \sum_i (\hat{p}_i - q_i) \log \frac{1}{q_i} \right| \\ &\leqslant \frac{1}{8}\varepsilon + \left| \sum_i (\hat{p}_i - q_i) \log \frac{1}{q_i} \right|. \end{aligned}$$

We have that $Z_3 = \left|\sum_i (\hat{p}_i - q_i) \log \frac{1}{q_i}\right| \geqslant \frac{1}{8}\varepsilon$ and thus with probability at least $1 - \frac{1}{4} - \frac{2}{100} = \frac{73}{100}$, the following will happen, the tester will reject: either $p(\bar{\mathcal{A}}) \geqslant 3\tau$, and it is rejected at Line 4 of Algorithm 4, or it passes and $p(\bar{\mathcal{A}}) \leqslant 3\tau$ and

$$Z_2 \geqslant \tfrac{1}{8} m_2 \varepsilon \text{ or } Z_3 \geqslant \tfrac{1}{8}\varepsilon,$$

and will be rejected. This concludes the proof. □

Remark 2.5. *We note that we can slightly improve the sample complexity of Theorem 1.1 (specifically, improving on the $\sqrt{k \log(k/\varepsilon)}$ term), at the price of a more complicated algorithm, by adding thresholds $\tau' = \frac{\varepsilon}{\log\log(k/\varepsilon)}$, $\tau'' = \frac{\varepsilon}{\log\log\log(k/\varepsilon)}$, and considering separately the elements in $\mathcal{A}' = \{i : q_i \in (\tau/k, \tau'/k]\}$, $\mathcal{A}'' = \{i : q_i \in (\tau'/k, \tau''/k]\}$; specifically, by grouping them in groups, and "merging" each group to get a "new" element with larger probability. For the sake of clarity, we defer this improvement to Appendix E.*

2.2 An $\Omega(\sqrt{k}/\varepsilon + \log^2 k/\varepsilon^2)$ lower bound

The $\Omega(\sqrt{k}/\varepsilon + \log^2 k/\varepsilon^2)$ lower bound comes from the combination of Lemma 2.6 and Lemma 2.7. We obtain Lemma 2.6 through the classical hard instance used for uniformity testing [Pan08] and a simple conversion between TV distance and entropy difference gives the result. We note that distributions close to uniform distribution actually have smaller entropy difference (uniform distribution is quite special: having the highest entropy of $\log k$). Indeed, replacing the uniform distribution with a slightly biased distribution, we obtain another hard instance for Lemma 2.7, using the classical Le Cam's two-point method.

Algorithm 2 Identity testing for bounded degree Bayes nets

Require: Sample access to Bayes net p, full description of Bayes net q, accuracy parameter ε, in-degree d and dimension n.

1: $\mathcal{S}_1 \leftarrow O\left(\left(\frac{2^{d/2}n\sqrt{d\log(n/\varepsilon)}}{\varepsilon^2} + \frac{d^2n^2}{\varepsilon^4}\right) d\log n\right)$ samples from p;
2: $\mathcal{S}_2 \leftarrow O\left(\frac{2^{d/2}n}{\varepsilon^2}\sqrt{\log(1/\varepsilon)}\cdot\log n\right)$ samples from p;
3: **for all** $L \in \mathcal{N}_{d+1} \cup \mathcal{N}_d$ **do** ▷ $\mathcal{N}_\ell$ is all subsets of $\{0,1\}^n$ with size ℓ
4: Call Algorithm 1 with p_L, q_L and $\mathcal{S}_1$; ▷ Entropy test on p_L and q_L with accuracy ε^2/n.
5: **if** Entropy test rejects **then return** reject
6: $S_3 \leftarrow O\left(\frac{d\log(n)\cdot\log(1/\varepsilon)}{\varepsilon^2}\right)$ samples from p and compute its empirical distribution $\hat{p}$;
7: **for all** $i \in [n]$ **do**
8: **if** $\hat{p}_{X_i,\Pi_i^G}(\bar{\mathcal{A}}_i') \geqslant \Omega(\varepsilon^2/\log(1/\varepsilon))$ **then return** reject
9: Check $d_{\mathrm{KL}}(p_{X_i,\Pi_i^G;\mathcal{A}_i'} \| q_{X_i,\Pi_i^G;\mathcal{A}_i'}) \geqslant \frac{\varepsilon^2}{n}$ or $d_{\mathrm{KL}}(p_{X_i,\Pi_i^G;\mathcal{A}_i'} \| q_{X_i,\Pi_i^G;\mathcal{A}_i'}) = 0$.
10: **if** i-th KL test says far **then return** reject
11: **return** accept ▷ Accept if all tests pass.

Lemma 2.6. *With fewer than $c_3 \cdot \sqrt{k}/\varepsilon$ samples from p, no tester can distinguish between $p = q$ and $|H(p) - H(q)| \geqslant \varepsilon$ with probability higher than $2/3$, where $c_3 > 0$ is an absolute constant.*

Lemma 2.7. *With fewer than $c_4 \cdot \log^2 k/\varepsilon^2$ samples from p, no tester can distinguish between $p = q$ and $|H(p) - H(q)| \geqslant \varepsilon$ with probability higher than $2/3$, where $c_4 > 0$ is an absolute constant.*

3 Application to identity testing for Bayes nets

We now provide an application of our main entropy identity testing theorem, to obtain an improved "standard" identity testing algorithm for Bayesian networks:

Theorem 3.1. *Given sample access to an in-degree d Bayes net p and full description of in-degree d Bayes net q, Algorithm 2 takes $C \cdot \left(\frac{2^{d/2}nd^{3/2}\log n\cdot\sqrt{\log(n/\varepsilon)}}{\varepsilon^2} + \frac{d^3n^2\cdot\log n}{\varepsilon^4}\right)$ samples to test between $p = q$ or $\mathrm{d_H}(p,q) \geqslant \Omega(\varepsilon)$, where $C > 0$ is an absolute constant. Moreover, the algorithm runs in time polynomial in n^d and $1/\varepsilon$.*

Before proceeding to the analysis of our algorithm, we require the following definitions.

Definition 3.2. *A* projection of a Bayes net p *on* $\{0,1\}^n$ *unto a DAG G is denoted p_G, and is defined by its probability mass function (PMF) as follows:*

$$p_G(X_1,\ldots,X_n) = \prod_{i=1}^n p(X_i \mid \Pi_i^G),$$

where Π_i^G is the set of parents of X_i in G.

Denote $\mathcal{U} := \bigcup_{i=1}^n \mathcal{A}_i$, where $\mathcal{A}_i := \left\{x \in \{0,1\}^n : q_{X_i,\Pi_i^G}(x_i(x),\pi_i^G(x)) \geqslant \Omega\left(\frac{\varepsilon^2/\log(1/\varepsilon)}{2^{d+1}n}\right)\right\}$. This gives us the property that marginalization over $X_i = x_i, \Pi_i^G = \pi_i^G$ works nicely as we include elements only based on its local property (as long as q_{X_i,Π_i} is large enough). And q is Markov w.r.t. G. We use $(x_i,\pi_i) \in \mathcal{A}_i'$, where $\mathcal{A}_i' = \left\{x' \in \{0,1\}^{|\Pi_i^G|+1} : q_{X_i,\Pi_i^G}(x') \geqslant \Omega\left(\frac{\varepsilon^2/\log(1/\varepsilon)}{2^{d+1}n}\right)\right\}$. Let $(a,B) \in \mathcal{A}_i'$, we have that as long as $(x_i(x),\pi_i(x)) = (a,B)$, then $x \in \mathcal{A}_i$ and vice versa, which means that

$$\mathcal{U} = \bigcup_{i=1}^n \mathcal{A}_i = \bigcup_{i=1}^n \{x \in \{0,1\}^n : (x_i(x),\pi_i^G(x)) \in \mathcal{A}_i'\}.$$

We will check if $p_{X_i,\Pi_i^G}(\bar{\mathcal{A}}_i') \geqslant \Omega(\varepsilon^2/\log(1/\varepsilon))$ and reject early if true; this takes $O\left(\frac{d\log(n)\cdot\log(1/\varepsilon)}{\varepsilon^2}\right)$ samples for all tests to be correct via a union bound. After passing this test, we

can conclude that

$$p(\bar{\mathcal{U}}) = \sum_{x \in \bigcap_{i=1}^{n} \bar{\mathcal{A}}_i} p(x) \leqslant \sum_{x \in \bar{\mathcal{A}}_1} p(x) \leqslant \sum_{x' \in \bar{\mathcal{A}}'_1} p_{X_1,\Pi_1^G}(x') = p_{X_1,\Pi_1^G}(\bar{\mathcal{A}}'_1) \leqslant O(\varepsilon^2/\log(1/\varepsilon)).$$

Similarly, we can upper bound $q(\bar{\mathcal{U}}) \leqslant q_{X_i,\Pi_i^G}(\bar{\mathcal{A}}'_i) \leqslant O(\varepsilon^2/\log(1/\varepsilon))$. Denote $p_{G;\mathcal{U}}$ as projecting p onto G, which gives p_G and then restricting the distribution p_G to take elements in $\mathcal{U}$.

Claim 3.3. *Suppose that $p(\bar{\mathcal{U}}) \leqslant O(\varepsilon^2/\log(1/\varepsilon))$, then we have*

$$d_{\mathrm{KL}}(p_{\bar{\mathcal{U}}} \| p_{G;\bar{\mathcal{U}}}) \geqslant -p(\bar{\mathcal{U}}) \cdot \log\left(\frac{1}{p(\bar{\mathcal{U}})}\right) \geqslant -O(\varepsilon^2).$$

We will also need the following lemma whose proof is deferred to Appendix C.

Lemma 3.4. *Suppose $d_H^2(p,q) \geqslant \Omega(\varepsilon^2)$ and $d_{\mathrm{KL}}(p\|p_G) \leqslant O(\varepsilon^2)$, then we have*

$$\sum_{i=1}^{n} d_{\mathrm{KL}}(p_{X_i,\Pi_i^G;\mathcal{A}'_i} \| q_{X_i,\Pi_i^G;\mathcal{A}'_i}) \geqslant \Omega(\varepsilon^2).$$

Therefore testing $d_{\mathrm{KL}}(p_{X_i,\Pi_i^G;\mathcal{A}'_i} \| q_{X_i,\Pi_i^G;\mathcal{A}'_i}) \geqslant \frac{\varepsilon^2}{n}$ over all i suffices to detect this case.

Proof of Theorem 3.1. We show the result by analyzing Algorithm 2. By Theorem 1.1, the sample complexity for entropy testing on any subset L of size (dimension) d or $d+1$, is $O\Big(2^{d/2} n\sqrt{d\log(n/\varepsilon)}/\varepsilon^2 + d^2n^2/\varepsilon^4\Big)$. To guarantee the success of every tests employed in the algorithm, we increase the sample complexity of each test by an extra $O(\log(n^{d+1})) = O(d\log n)$ factor to boost their success probability to $1 - \frac{1}{100n^{d+1}}$ (via a standard majority vote technique), which will allow us to use a union bound over all tests as there are at most n^{d+1} subsets with size $d+1$. For this step, the sample complexity will be

$$O\left(\left(\frac{2^{d/2} n\sqrt{d\log(n/\varepsilon)}}{\varepsilon^2} + \frac{d^2n^2}{\varepsilon^4}\right) d\log n\right).$$

With this in hand, we will proceed with the analysis under the event that every entropy test performed is correct (which by the above happens with high probability). If distribution p manages to pass all the entropy tests, it must satisfy the following:

$$|H(p_L) - H(q_L)| \leqslant \frac{\varepsilon^2}{n}, \tag{6}$$

for every subset L of size $d+1$ for the latter, and every subset L of size d or $d+1$ for the former. From here, in principle, we can perform structural learning of p through $H(q_L)$, which then gives us an approximated DAG $\hat{G}$ of p and we can check $\mathrm{d_H}(p_{\hat{G}}, q_{\hat{G}})$. Unfortunately, structural learning of Bayes nets is known to be computationally hard in many settings [Höf93, CHM04], and so this would lead to a computationally inefficient algorithm.

Instead, we argue that this (learning) step can be bypassed entirely: the intuition of the argument is to view structure learning for Bayes net as an optimization problem; and any assignment x to the two optimization problems (structure learning of p and q) satisfy $f_1(x) = f_2(x) \pm O(\varepsilon^2)$ due to their local entropy being close[8] – this means that an optima x_1 for f_1 satisfies $\min_x f_1(x) = f_1(x_1) \geqslant f_2(x_1) - \varepsilon^2 \geqslant \min_x f_2(x) - \varepsilon^2$ and vice versa (optima x_2 for f_2).

More formally, by the celebrated works of Chow and Liu [CL68] and its generalization to Bayes net, one can write the KL divergence between Bayes net and its projection to any graph G as difference between sum of n local conditional entropies (we provide a derivation for the sake of completeness in F):

$$0 \leqslant d_{\mathrm{KL}}(p\|p_G) = -\sum_{i=1}^{n} H(p_{X_i,X_{\Pi_i}}|p_{X_{\Pi_i}}) + \sum_{i=1}^{n} H(p_{X_i,X_{\Pi_i^G}}|p_{X_{\Pi_i^G}}), \tag{7}$$

[8]Here, x is the DAG's assignment of parents and child; and $f_1(x)$ (resp. $f_2(x)$) is the associated KL-divergence (also called *score* of the DAG in the literature, which measures the how well DAG models the true Bayes net) between p (resp. q) and x. Since we are in the realizable setting, the optimal is in fact 0.

$$d_{\mathrm{KL}}(q\|q_{G'}) = -\sum_{i=1}^{n} H(q_{X_i,X_{\Pi_i^G}}|q_{X_{\Pi_i^G}}) + \sum_{i=1}^{n} H(q_{X_i,X_{\Pi_i}}|q_{X_{\Pi_i}}), \tag{8}$$

where X_{Π_i} denotes the parents of X_i in Bayes net p and $X_{\Pi_i^G}$ denotes the parents of X_i in DAG G. Here we assume that q is Markov with respect to G. Since the local entropies between p and q are close by $O(\varepsilon^2/n)$ (see (6)) and its relation to conditional entropy via (3), we can conclude the following:

$$H(q_{X_i,X_{\Pi_i}}|q_{X_{\Pi_i}}) - O(\varepsilon^2/n) \leqslant H(p_{X_i,X_{\Pi_i}}|p_{X_{\Pi_i}}) \leqslant H(q_{X_i,X_{\Pi_i}}|q_{X_{\Pi_i}}) + O(\varepsilon^2/n). \tag{9}$$

$$H(q_{X_i,X_{\Pi_i^G}}|q_{X_{\Pi_i^G}}) - O(\varepsilon^2/n) \leqslant H(p_{X_i,X_{\Pi_i^G}}|p_{X_{\Pi_i^G}}) \leqslant H(q_{X_i,X_{\Pi_i^G}}|q_{X_{\Pi_i^G}}) + O(\varepsilon^2/n). \tag{10}$$

By our assumption, since q is Markov to G, we can combine (7), (9) and (10), which then give:

$$0 \leqslant d_{\mathrm{KL}}(p\|p_G) \leqslant -\sum_{i=1}^{n} H(q_{X_i,X_{\Pi_i}}|q_{X_{\Pi_i}}) + \sum_{i=1}^{n} H(q_{X_i,X_{\Pi_i^G}}|q_{X_{\Pi_i^G}}) + O(\varepsilon^2) = -d_{\mathrm{KL}}(q\|q_{G'}) + O(\varepsilon^2),$$

where p is Markov to G', as Π_i is the set of parents of X_i for p. Rearranging terms and we have

$$d_{\mathrm{KL}}(p\|p_G) \leqslant O(\varepsilon). \tag{11}$$

With this at hand, we continue onto the KL testing part. The algorithm will check if $p_{X_i,\Pi_i^G}(\bar{\mathcal{A}}_i') \geqslant \Omega(\varepsilon^2/\log(1/\varepsilon))$ and reject early if it is true (this costs $O\left(\frac{d\log(n)\cdot\log(1/\varepsilon)}{\varepsilon^2}\right)$) and then check for every $i \in [n]$,

$$d_{\mathrm{KL}}(p_{X_i,\Pi_i^G;\mathcal{A}_i'}\|q_{X_i,\Pi_i^G;\mathcal{A}_i'}) \geqslant \frac{\varepsilon^2}{n} \text{ or } d_{\mathrm{KL}}(p_{X_i,\Pi_i^G;\mathcal{A}_i'}\|q_{X_i,\Pi_i^G;\mathcal{A}_i'}) = 0.$$

Using Lemma 2.4 (after converting from KL to χ^2 noting the mass of $\mathcal{A}'$ on both distributions) and a union bound over n tests, the sample complexity is

$$O\left(\frac{2^{d/2}n}{\varepsilon^2}\sqrt{\log(1/\varepsilon)}\cdot\log n\right).$$

Following this, we look at the two cases:

- If $p = q$, then with high probability, p will pass all entropy tests, all KL local tests and the tester will accept.
- If $\mathrm{d_H}(p,q) \geqslant \varepsilon$, either it fails one of the entropy test. If it does pass the entropy test, then we must have that $d_{\mathrm{KL}}(p\|p_G) \leqslant O(\varepsilon^2)$ by (11). Then following Lemma 3.4 and Lemma 2.4, the tester will reject.

In total, the sample complexity is (dominated by entropy testing):

$$O\left(\left(\frac{2^{d/2}n\sqrt{d\log(n/\varepsilon)}}{\varepsilon^2} + \frac{d^2n^2}{\varepsilon^4}\right)d\log n\right).$$

This concludes the proof of the theorem. □

4 Conclusion and open problems

In this paper, we study a variant of distribution testing problem in terms of entropy difference; we give nearly tight upper and lower sample complexity bounds for the problem. We subsequently apply our entropy testing algorithm to identity testing of Bayes nets, which unlike prior works, makes merely the necessary assumptions (the bound on the in-degree of the Bayes nets).

Future direction. We believe the *closeness* (two-sample) testing variant of the problem (testing if two unknown distribution p and q are the same or far in terms of entropy difference) could also be interesting; and, notably, has connections to other distribution testing problems: first, it should lead to a natural solution to closeness testing of Bayes nets via ideas in this paper. Second, solving the closeness entropy testing problem give another path to testing independence in terms of mutual information (studied in [BGP+23] and also covered in [CDKS18]), a notion closely related to entropy.

Acknowledgments and Disclosure of Funding

We would like to thank the reviewers for their suggestions and efforts which help improve this paper. Yang would like to acknowledge the support of the JD Technology Scholarship.

References

[ABIS19] Jayadev Acharya, Sourbh Bhadane, Piotr Indyk, and Ziteng Sun. Estimating entropy of distributions in constant space. In *NeurIPS*, pages 5163–5174, 2019.

[ADOS17] Jayadev Acharya, Hirakendu Das, Alon Orlitsky, and Ananda Theertha Suresh. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 11–21. PMLR, 2017.

[AISW20] Jayadev Acharya, Ibrahim Issa, Nirmal V. Shende, and Aaron B. Wagner. Estimating quantum entropy. *IEEE J. Sel. Areas Inf. Theory*, 1(2):454–468, 2020.

[AMNW22] Maryam Aliakbarpour, Andrew McGregor, Jelani Nelson, and Erik Waingarten. Estimation of entropy in constant space with improved sample complexity. In *NeurIPS*, 2022.

[AOST17] Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. Estimating renyi entropy of discrete distributions. *IEEE Trans. Inf. Theory*, 63(1):38–56, 2017.

[BCY22] Arnab Bhattacharyya, Clément L. Canonne, and Joy Qiping Yang. Independence testing for bounded degree bayesian network. *CoRR*, abs/2204.08690, 2022.

[BDKR02] Tŭgkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating entropy. In *STOC*, pages 678–687. ACM, 2002.

[BFF+01] Tugkan Batu, Lance Fortnow, Eldar Fischer, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *FOCS*, pages 442–451. IEEE Computer Society, 2001.

[BGKV21] Arnab Bhattacharyya, Sutanu Gayen, Saravanan Kandasamy, and NV Vinodchandran. Testing product distributions: A closer look. In *Algorithmic Learning Theory*, pages 367–396. PMLR, 2021.

[BGMV20] Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, and N. V. Vinodchandran. Efficient distance approximation for structured high-dimensional distributions via learning. In *NeurIPS*, 2020.

[BGP+23] Arnab Bhattacharyya, Sutanu Gayen, Eric Price, Vincent Y. F. Tan, and N. V. Vinodchandran. Near-optimal learning of tree-structured distributions by chow and liu. *SIAM J. Comput.*, 52(3):761–793, 2023.

[BY02] Ziv Bar-Yossef. *The Complexity of Massive Data Set Computations*. PhD thesis, UC Berkeley, 2002. Adviser: Christos Papadimitriou. Available at `http://webee.technion.ac.il/people/zivby/index_files/Page1489.html`.

[Can22] Clément L. Canonne. Topics and techniques in distribution testing: A biased but representative sample. *Found. Trends Commun. Inf. Theory*, 19(6):1032–1198, 2022.

[CDKS18] Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing conditional independence of discrete distributions. In *STOC*, pages 735–748. ACM, 2018.

[CDKS20] 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.

[CHM04] David Maxwell Chickering, David Heckerman, and Christopher Meek. Large-sample learning of bayesian networks is np-hard. *J. Mach. Learn. Res.*, 5:1287–1330, 2004.

[CK11] Imre Csiszár and János Körner. *Information Theory - Coding Theorems for Discrete Memoryless Systems, Second Edition*. Cambridge University Press, 2011.

[CL68] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theory*, 14(3):462–467, 1968.

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

[DKW18] 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.

[DP16] Constantinos Daskalakis and Qinxuan Pan. Square hellinger subadditivity for bayesian networks and its applications to identity testing. *arXiv preprint arXiv:1612.03164*, 2016. Full version of [DP17].

[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.

[GHS21] Tom Gur, Min-Hsiu Hsieh, and Sathyawageeswar Subramanian. Sublinear quantum algorithms for estimating von neumann entropy. *CoRR*, abs/2111.11139, 2021.

[HJW15a] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Adaptive estimation of shannon entropy. In *ISIT*, pages 1372–1376. IEEE, 2015.

[HJW15b] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Adaptive estimation of shannon entropy. *CoRR*, abs/1502.00326, 2015.

[Höf93] Klaus-U Höffgen. Learning and robust learning of product distributions. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 77–83, 1993.

[KCG+23] Neville Kenneth Kitson, Anthony C. Constantinou, Zhigao Guo, Yang Liu, and Kiattikun Chobtham. A survey of bayesian network structure learning. *Artif. Intell. Rev.*, 56(8):8721–8814, 2023.

[KDDC23] Anthimos Vardis Kandiros, Constantinos Daskalakis, Yuval Dagan, and Davin Choo. Learning and testing latent-tree ising models efficiently. In *COLT*, volume 195 of *Proceedings of Machine Learning Research*, pages 1666–1729. PMLR, 2023.

[Pan04] Liam Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE Trans. Inform. Theory*, 50(9):2200–2203, 2004.

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

[VV11a] Gregory Valiant and Paul Valiant. Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new clts. In *STOC*, pages 685–694. ACM, 2011.

[VV11b] Gregory Valiant and Paul Valiant. The power of linear estimators. In *FOCS*, pages 403–412. IEEE Computer Society, 2011.

[VV13] Paul Valiant and Gregory Valiant. Estimating the unseen: Improved estimators for entropy and other properties. In *NIPS*, pages 2157–2165, 2013.

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

[WY16] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inform. Theory*, 62(6):3702–3720, 2016.

A Derivation for Line 4 in Algorithm 1

We need to show the following: with probability at least $99/100$, if $Z_1 \geqslant 2\tau$, then $p(\bar{\mathcal{A}}) \geqslant \tau$; and if $Z_1 < 2\tau$, then $p(\bar{\mathcal{A}}) < 3\tau$. For the first one, we prove by contrapositive: with high probability $1 - \frac{1}{200}$, $p(\bar{\mathcal{A}}) < \tau \Rightarrow Z_1 < 2\tau$. Suppose $T = \mathrm{Binomial}(m_1, \tau)$ and setting $m_1 = \frac{48}{\tau} \geqslant \frac{3\log 200}{\tau}$, and using a Chernoff bound, we have the following

$$\Pr[T \geqslant 2\tau] \leqslant \exp(-\tau \cdot m_1/3) \leqslant \frac{1}{200}.$$

Since any $\mathrm{Binomial}(m_1, p(\bar{\mathcal{A}}))$ will be first-order stochastic dominated by $\mathrm{Binomial}(m_1, \tau)$ if $p(\bar{\mathcal{A}}) < \tau$, we can conclude the following: if $p(\bar{\mathcal{A}}) < \tau$, then $\Pr[Z_1 \geqslant 2\tau] \leqslant \Pr[T \geqslant 2\tau] \leqslant \frac{1}{200}$.

For the latter, we prove via its contrapositive: with probability $1 - \frac{1}{200}$, $p(\bar{\mathcal{A}}) \geqslant 3\tau \Rightarrow Z_1 \geqslant 2\tau$. As $p(\bar{\mathcal{A}}) \geqslant 3\tau$, take $m_1 = \frac{48}{\tau} \geqslant \frac{9\log 200}{\tau}$, by a Chernoff bound, we have

$$\Pr[\hat{p}'(\bar{\mathcal{A}}) \leqslant 2\tau] \leqslant \Pr[\hat{p}'(\bar{\mathcal{A}}) \leqslant (1-1/3)\cdot p(\bar{\mathcal{A}})] \leqslant \exp\left(-\frac{m_1 \cdot p(\bar{\mathcal{A}})}{18}\right) \leqslant \exp\left(-\frac{m_1 \cdot \tau}{9}\right) \leqslant \frac{1}{200}.$$

Combining the two with a union bound concludes the proof.

B Deferred proofs from Section 2.1

Claim 2.2. *Let $\mathcal{A}$ be any set such that $p(\bar{\mathcal{A}}) < \varepsilon/2$. Then, if $|H(p_{\mathcal{A}}) - H(q_{\mathcal{A}})| \geqslant \varepsilon$, we must have (i) $d_{\mathrm{KL}}(p_{\mathcal{A}} \| q_{\mathcal{A}}) \geqslant \frac{\varepsilon}{2}$ or (ii) $|\sum_{i\in\mathcal{A}}(p_i - q_i)\log(\frac{1}{q_i})| \geqslant \frac{\varepsilon}{2}$.*

Proof of Claim 2.2. We can bound $|H(p_{\mathcal{A}}) - H(q_{\mathcal{A}})|$ as

$$\begin{aligned}
\varepsilon \leq |H(p_{\mathcal{A}}) - H(q_{\mathcal{A}})| &= \left|\sum_{i\in\mathcal{A}}\left(p_i \log\frac{1}{p_i} - q_i\log\frac{1}{q_i}\right)\right| \\
&= \left|\sum_{i\in\mathcal{A}}\left(p_i \log\frac{q_i}{p_i} + p_i\log\frac{1}{q_i} - q_i\log\frac{1}{q_i}\right)\right| \\
&\leqslant \left|\sum_{i\in\mathcal{A}} p_i\log\frac{q_i}{p_i}\right| + \left|\sum_{i\in\mathcal{A}}\left(p_i\log\frac{1}{q_i} - q_i\log\frac{1}{q_i}\right)\right| \\
&= |d_{\mathrm{KL}}(p_{\mathcal{A}}\|q_{\mathcal{A}})| + \left|\sum_{i\in\mathcal{A}}(p_i - q_i)\log\frac{1}{q_i}\right|, \qquad (12)
\end{aligned}$$

which implies that at least one of the two terms is at least $\varepsilon/2$. If it is the second, we are done; otherwise, we know that either

$$d_{\mathrm{KL}}(p_{\mathcal{A}}\|q_{\mathcal{A}}) \geqslant \frac{1}{2}\varepsilon \text{ or } d_{\mathrm{KL}}(p_{\mathcal{A}}\|q_{\mathcal{A}}) \leqslant -\frac{1}{2}\varepsilon.$$

We will rule out the second case, using that $\log\frac{1}{x} \geqslant 1 - x$ for $x > 0$,[9]

$$d_{\mathrm{KL}}(p_{\mathcal{A}}\|q_{\mathcal{A}}) = \sum_{i\in\mathcal{A}} p_i\log\frac{p_i}{q_i} \geqslant \sum_{i\in\mathcal{A}} p_i\left(1 - \frac{q_i}{p_i}\right) = q(\bar{\mathcal{A}}) - p(\bar{\mathcal{A}}) > -\frac{1}{2}\varepsilon.$$

Thus, we cannot have $d_{\mathrm{KL}}(p_{\mathcal{A}}\|q_{\mathcal{A}}) \leqslant -\frac{1}{2}\varepsilon$ and so $d_{\mathrm{KL}}(p_{\mathcal{A}}\|q_{\mathcal{A}}) \geqslant \frac{1}{2}\varepsilon$. □

Claim 2.3. *Let $\hat{p}$ be the empirical estimator for an unknown discrete distribution p supported on $[k]$, based on $\mathrm{Poi}(m)$ samples, where $m = \Theta\left(\frac{\log^2(k)}{\varepsilon^2}\right)$; assume that $d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) \leqslant \varepsilon/8$ and $p(\bar{\mathcal{A}}) + q(\bar{\mathcal{A}}) \leqslant 4\tau = \frac{1}{4}\frac{\varepsilon}{\log(k/\varepsilon)}$,[10] then*

$$\Pr\left[\left|\sum_{i\in\mathcal{A}}(p_i - \hat{p}_i)\log\frac{1}{q_i}\right| \geqslant \frac{1}{8}\varepsilon\right] \leqslant \frac{1}{100}.$$

[9] In the case of $p_i = 0$, we still have $p_i\log\left(\frac{p_i}{q_i}\right) \geqslant p_i - q_i$.

[10] One can remove the assumption that $p(\bar{\mathcal{A}}) + q(\bar{\mathcal{A}}) \leqslant 4\tau$, at the cost of a slightly worse constant.

Proof of Claim 2.3. We follow the same analysis as in [WY16]. Letting $Y_i := (p_i - \hat{p}_i)\log\frac{1}{q_i}$ for $i \in \mathcal{A}$, we have $\mathbb{E}[Y_i] = 0$ and

$$\mathrm{Var}[Y_i] = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])^2] = \mathbb{E}[Y_i^2] = \mathbb{E}[(p_i - \hat{p}_i)^2]\log^2\frac{1}{q_i} = \frac{1}{m^2}(mp_i)\log^2\frac{1}{q_i} = \frac{p_i}{m}\log^2\frac{1}{q_i}.$$

Let $Y := \sum_{i\in\mathcal{A}}(p_i - \hat{p}_i)\log\frac{1}{q_i}$. We will use our assumption that $d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) \leqslant \varepsilon/8$ and $p(\bar{\mathcal{A}}) + q(\bar{\mathcal{A}}) \leqslant \frac{1}{4}\frac{\varepsilon}{\log(k/\varepsilon)}$ below. Note that, our analysis is in the Poissonized setting:

$$\begin{aligned}
\mathrm{Var}[Y] &= \mathrm{Var}\left[\sum_{\mathcal{A}}(p_i - \hat{p}_i)\log\frac{1}{q_i}\right] \\
&= \sum_{i\in\mathcal{A}}\frac{p_i}{m}\log^2\left(\frac{1}{q_i}\right) \\
&= \sum_{i\in\mathcal{A}}\frac{p_i}{m}\left(\log\left(\frac{1}{p_i}\right) + \log\left(\frac{p_i}{q_i}\right)\right)^2 \\
&\leqslant \sum_{i\in\mathcal{A}}\frac{p_i}{m}\left(2\left(\log\left(\frac{1}{p_i}\right)\right)^2 + 2\left(\log\left(\frac{p_i}{q_i}\right)\right)^2\right) \\
&= \sum_{i\in\mathcal{A}}\frac{2}{m}p_i\log^2\left(\frac{1}{p_i}\right) + \sum_{i\in\mathcal{A}}\frac{2}{m}p_i\log^2\left(\frac{p_i}{q_i}\right) \\
&\leqslant \sum_{i\in\mathcal{A}}\frac{2}{m}p_i\log^2\left(\frac{1}{p_i}\right) + \sum_{\frac{p_i}{q_i}\geqslant 1, i\in\mathcal{A}}\frac{2}{m}p_i\left(\frac{p_i}{q_i} - 1\right) + \sum_{\frac{p_i}{q_i}<1, i\in\mathcal{A}}\frac{2}{m}p_i\left(\frac{q_i}{p_i} - 1\right) \qquad (13) \\
&= \sum_{i\in\mathcal{A}}\frac{2}{m}p_i\log^2\left(\frac{1}{p_i}\right) + \sum_{\frac{p_i}{q_i}\geqslant 1, i\in\mathcal{A}}\frac{2}{m}\frac{(p_i - q_i)^2}{q_i} + \sum_{\frac{p_i}{q_i}\geqslant 1, i\in\mathcal{A}}\frac{2}{m}(p_i - q_i) + \sum_{\frac{p_i}{q_i}<1, i\in\mathcal{A}}\frac{2}{m}(q_i - p_i) \\
&\leqslant \frac{4\log^2 k}{m} + \frac{6}{m} + \frac{2}{m}(d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) + d_{\mathrm{TV}}(p, q)) \qquad (14) \\
&\leqslant \frac{4\log^2 k}{m} + \frac{6}{m} + \frac{2}{m}\left(\frac{\varepsilon}{8} + \sqrt{\frac{\varepsilon}{8}} + 4\tau\right) \leqslant \frac{4\log^2 k}{m} + \frac{8}{m} \leqslant \frac{22\log^2 k}{m} \qquad (15)
\end{aligned}$$

For (13), we analyze by two cases: if $\frac{p_i}{q_i} \geqslant 1$, we have that $p_i\log^2\left(\frac{p_i}{q_i}\right) \leqslant p_i\left(\frac{p_i}{q_i} - 1\right)$; otherwise, $p_i\log^2\left(\frac{p_i}{q_i}\right) = p_i\log^2\left(\frac{q_i}{p_i}\right) < p_i\left(\frac{q_i}{p_i} - 1\right)$. And we use [HJW15b, Lemma3], $\sum_{i\in\mathcal{A}} p_i\log^2\left(\frac{1}{p_i}\right) \leqslant 2\log^2 k + 3$ in (14). We use the premise and (4) in (15) and we have that

$$d_{\mathrm{TV}}(p, q) = d_{\mathrm{TV}}(p_{\mathcal{A}}, q_{\mathcal{A}}) + d_{\mathrm{TV}}(p_{\bar{\mathcal{A}}}, q_{\bar{\mathcal{A}}}) \leqslant \sqrt{d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}})} + p(\bar{\mathcal{A}}) + q(\bar{\mathcal{A}}) \leqslant \sqrt{\frac{\varepsilon}{8}} + 4\tau;$$

and the last step is obtained by noticing that $\log(k) \geqslant \frac{2}{3}$ for $k \geqslant 2$. By Chebyshev's inequality, we then have that

$$\Pr\left[|Y| \geqslant 10\sqrt{\frac{38\log^2 k}{m}}\right] \leqslant \Pr\left[|Y - \mathbb{E}[Y]| \geqslant 10\sqrt{\mathrm{Var}[Y]}\right] \leqslant \frac{1}{100},$$

and this last inequality yields the statement as long as $m \geqslant \frac{22\times 100\times 8^2\log^2(k)}{\varepsilon^2} = \frac{140800\log^2(k)}{\varepsilon^2}$. □

Lemma 2.4. *Let $\mathcal{A} := \{i \in [k] \mid q_i \geqslant \alpha\}$. Let $m_2 \geqslant 16384\max\left\{\sqrt{\frac{1}{\alpha\varepsilon}}, \frac{\sqrt{k}}{\varepsilon}\right\}$ be the number of samples used to compute Z_2. Then $\mathbb{E}[Z_2] = m_2 d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}})$. Moreover, if $d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) \leqslant \frac{\varepsilon}{2}$, then $\mathrm{Var}[Z_2] \leqslant (\frac{1}{32}m_2\varepsilon)^2$. If $d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) \geqslant \varepsilon$, then $\mathrm{Var}[Z_2] \leqslant O(\mathbb{E}[Z_2]^2)$.*

Proof of Lemma 2.4. The proof is a relatively straightforward modification of the argument of [DKW18, Lemma 2]. We have the expectation and variance of Z_2,

$$\mathbb{E}[Z_2] = m_2 d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) \text{ and } \mathrm{Var}[Z_2] = \sum_{i\in\mathcal{A}} \left[2\frac{p_i^2}{q_i^2} + 4m_2 \frac{p_i(p_i-q_i)^2}{q_i^2} \right].$$

It boils down to bounding the following,

$$\begin{aligned} 2\sum_{i\in\mathcal{A}} \frac{p_i^2}{q_i^2} &\leqslant 4k + 4\sum_{i\in\mathcal{A}} \frac{(p_i-q_i)^2}{q_i^2} \\ &\leqslant 4k + \frac{4}{\alpha}\sum_{i\in\mathcal{A}} \frac{(p_i-q_i)^2}{q_i} \\ &\leqslant 4k + \frac{4}{\alpha m_2}\mathbb{E}[Z_2]. \end{aligned}$$

Derivation of the inequalities follow from [DKW18, proof of Lemma 2].

$$\begin{aligned} 4m_2 \sum_{i\in\mathcal{A}} \frac{p_i(p_i-q_i)^2}{q_i^2} &\leqslant 4m_2 \left(\sum_{i\in\mathcal{A}} \frac{p_i^2}{q_i^2} \right)^{1/2} \left(\sum_{i\in\mathcal{A}} \frac{(p_i-q_i)^4}{q_i^2} \right)^{1/2} \\ &\leqslant 4m_2 \left(4k + \frac{4}{\alpha m_2}\mathbb{E}[Z_2] \right)^{1/2} \left(\sum_{i\in\mathcal{A}} \frac{(p_i-q_i)^2}{q_i} \right) \\ &= 4\left(2\sqrt{k} + 2\sqrt{\frac{1}{\alpha m_2}\mathbb{E}[Z_2]} \right) \mathbb{E}[Z_2] \\ &= 8\sqrt{k}\mathbb{E}[Z_2] + 8(\alpha m_2)^{-1/2}(\mathbb{E}[Z_2])^{3/2}. \end{aligned}$$

Combing both, we have that

$$\mathrm{Var}[Z_2] \leqslant 4k + \left(\frac{4}{\alpha m_2} + 8\sqrt{k} \right) \mathbb{E}[Z_2] + 8(\alpha m_2)^{-1/2}(\mathbb{E}[Z_2])^{3/2}. \tag{16}$$

When $d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) \leqslant \varepsilon/2$, then $\mathbb{E}[Z_2] \leqslant \frac{m_2\varepsilon}{2}$; and we solve $\mathrm{Var}[Z_2] \leqslant (\frac{1}{32}m_2\varepsilon)^2$, which gives

$$4k + \left(\frac{4}{\alpha m_2} + 8\sqrt{k} \right) \frac{m_2\varepsilon}{2} + 8(\alpha m_2)^{-1/2}(\frac{m_2\varepsilon}{2})^{3/2} \leqslant (\frac{1}{32}m_2\varepsilon)^2.$$

We solve for the relaxation:

$$\text{LHS} \leqslant 4 \cdot \max\left\{ 4k, \frac{2\varepsilon}{\alpha}, 4\sqrt{k}m_2\varepsilon, 8\frac{m_2}{\sqrt{\alpha}2^{3/2}}\varepsilon^{3/2} \right\} \leqslant (\frac{1}{32}m_2\varepsilon)^2$$

In the end, we obtain:

$$\max\left\{ 128\cdot\frac{\sqrt{k}}{\varepsilon}, 64\sqrt{\frac{2}{\alpha\varepsilon}}, 32^2\cdot 16\sqrt{k}, 32^2\cdot 16\frac{1}{\sqrt{\alpha\varepsilon}\sqrt{2}} \right\} \leqslant 32^2\cdot 16\cdot\max\left\{ \frac{\sqrt{k}}{\varepsilon}, \sqrt{\frac{1}{\alpha\varepsilon}} \right\} \leqslant m_2$$

When $d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) \geqslant \varepsilon$, then $\mathbb{E}[Z_2] \geqslant m_2\varepsilon$; and we solve $\mathrm{Var}[Z_2] \leqslant (\frac{1}{4}\mathbb{E}[Z_2])^2$,

$$4k + \left(\frac{4}{\alpha m_2} + 8\sqrt{k} \right) \mathbb{E}[Z_2] + 8(\alpha m_2)^{-1/2}(\mathbb{E}[Z_2])^{3/2} \leqslant (\frac{1}{4}\mathbb{E}[Z_2])^2,$$

which is equivalent to the following

$$\frac{4k}{(\mathbb{E}[Z_2])^{3/2}} + \left(\frac{4}{\alpha m_2} + 8\sqrt{k} \right) \frac{1}{(\mathbb{E}[Z_2])^{1/2}} + 8(\alpha m_2)^{-1/2} \leqslant \frac{1}{16}(\mathbb{E}[Z_2])^{1/2}$$

Further relaxing the solution, it is enough to have

$$\begin{aligned} & \frac{4k}{(\mathbb{E}[Z_2])^{3/2}} + \left(\frac{4}{\alpha m_2} + 8\sqrt{k}\right)\frac{1}{(\mathbb{E}[Z_2])^{1/2}} + 8(\alpha m_2)^{-1/2} \\ \leqslant\ & \frac{4k}{(m_2\varepsilon)^{3/2}} + \left(\frac{4}{\alpha m_2} + 8\sqrt{k}\right)\frac{1}{(m_2\varepsilon)^{1/2}} + 8\frac{1}{\sqrt{\alpha m_2}} \\ \leqslant\ & \frac{1}{16}(m_2\varepsilon)^{1/2} \leqslant \frac{1}{16}(\mathbb{E}[Z_2])^{1/2}, \end{aligned}$$

as long as the following holds,

$$m_2 \geqslant 64\max\left\{\frac{2\sqrt{k}}{\varepsilon}, 2\sqrt{\frac{1}{\alpha\varepsilon}}, 8\frac{\sqrt{k}}{\varepsilon}, 8\sqrt{\frac{\alpha}{\varepsilon}}\right\} = \max\left\{128\sqrt{\frac{1}{\alpha\varepsilon}}, 512\frac{\sqrt{k}}{\varepsilon}\right\}. \tag{17}$$

Letting $m_2 \geqslant 512\max\left\{\sqrt{\frac{1}{\alpha\varepsilon}}, \frac{\sqrt{k}}{\varepsilon}\right\}$, we have that both statements. □

C Proofs of testing Bayesian networks

Proof of Claim 3.3.

$$\begin{aligned} d_{\mathrm{KL}}(p_{\bar{\mathcal{U}}}\|p_{G;\bar{\mathcal{U}}}) &= \sum_{x\in\bar{\mathcal{U}}} p(x)\log\frac{p(x)}{p_G(x)} \\ &= -\sum_{x\in\bar{\mathcal{U}}} p(x)\log\frac{p_G(x)}{p(x)} \\ &\geqslant -\left(\sum_{x\in\bar{\mathcal{U}}} p(x)\right)\cdot\log\left(\frac{\sum_{x\in\bar{\mathcal{U}}} p(x)\cdot\frac{p_G(x)}{p(x)}}{\sum_{x\in\bar{\mathcal{U}}} p(x)}\right) \\ &= -p(\bar{\mathcal{U}})\cdot\log\left(\frac{p_G(\bar{\mathcal{U}})}{p(\bar{\mathcal{U}})}\right) \\ &\geqslant -O(\varepsilon^2/\log(1/\varepsilon))\cdot\log\left(\frac{1}{\varepsilon^2/\log(1/\varepsilon)}\right) \\ &\geqslant -O(\varepsilon^2), \end{aligned}$$

where we use monotonicity of $-x\log\frac{1}{x}$ and the fact that $p_G(\bar{\mathcal{U}}) \leqslant 1$ in the second last inequality. □

Proof of Lemma 3.4. By Claim 3.3 and the assumption that $d_{\mathrm{KL}}(p\|p_G) \leqslant O(\varepsilon^2)$, we have that

$$d_{\mathrm{KL}}(p_{\mathcal{U}}\|p_{G;\mathcal{U}}) = d_{\mathrm{KL}}(p\|p_G) - d_{\mathrm{KL}}(p_{\bar{\mathcal{U}}}\|p_{G;\bar{\mathcal{U}}}) \leqslant O(\varepsilon^2).$$

$$\Omega(\varepsilon^2) \leqslant d_H^2(p,q) = d_H^2(p_{\mathcal{U}},q_{\mathcal{U}}) + d_H^2(p_{\bar{\mathcal{U}}},q_{\bar{\mathcal{U}}}) \leqslant d_H^2(p_{\mathcal{U}},q_{\mathcal{U}}) + \frac{1}{2}(p(\bar{\mathcal{U}})+q(\bar{\mathcal{U}})) \leqslant d_H^2(p_{\mathcal{U}},q_{\mathcal{U}}) + O(\varepsilon^2).$$

$$\Omega(\varepsilon^2) \leqslant d_H^2(p_{\mathcal{U}},q_{\mathcal{U}}) \leqslant d_{\mathrm{KL}}(p_{\mathcal{U}}\|q_{\mathcal{U}}) + q(\mathcal{U}) - p(\mathcal{U}) \Rightarrow d_{\mathrm{KL}}(p_{\mathcal{U}}\|q_{\mathcal{U}}) \geqslant \Omega(\varepsilon^2). \tag{18}$$

By (18), we write

$$\begin{aligned}
\Omega(\varepsilon^2) - d_{\mathrm{KL}}(p_{\mathcal{U}} \| p_{G;\mathcal{U}}) &\leqslant d_{\mathrm{KL}}(p_{\mathcal{U}} \| q_{\mathcal{U}}) - d_{\mathrm{KL}}(p_{\mathcal{U}} \| p_{G;\mathcal{U}}) \\
&= \sum_{x \in \mathcal{U}} p(x) \sum_{i=1}^{n} \log \frac{p(x_i|\pi_i^G)}{q(x_i|\pi_i^G)} \\
&= \sum_{x \in \mathcal{U}} p(x) \sum_{i=1}^{n} \log \frac{p(x_i, \pi_i^G)}{q(x_i, \pi_i^G)} \frac{q(\pi_i^G)}{p(\pi_i^G)} \\
&\leqslant \sum_{x \in \mathcal{U}} p(x) \left(\sum_{i=1}^{n} \log \frac{p(x_i, \pi_i^G)}{q(x_i, \pi_i^G)} - \sum_{i=1}^{n} \log \frac{q(\pi_i^G)}{p(\pi_i^G)} \right) \\
&= \sum_{x \in \bigcup_{i=1}^{n} \mathcal{A}_i} p(x) \left(\sum_{i=1}^{n} \log \frac{p(x_i, \pi_i^G)}{q(x_i, \pi_i^G)} - \sum_{i=1}^{n} \log \frac{q(\pi_i^G)}{p(\pi_i^G)} \right) \\
&= \sum_{i=1}^{n} \sum_{x : x_i(x), \pi_i^G(x) \in \mathcal{A}'_i} p(x) \left(\log \frac{p(x_i, \pi_i^G)}{q(x_i, \pi_i^G)} - \log \frac{q(\pi_i^G)}{p(\pi_i^G)} \right) \\
&= \sum_{i=1}^{n} \sum_{(x_i, \pi_i^G) \in \mathcal{A}'_i} p(x_i, \pi_i^G) \left(\log \frac{p(x_i, \pi_i^G)}{q(x_i, \pi_i^G)} - \log \frac{q(\pi_i^G)}{p(\pi_i^G)} \right) \\
&= \sum_{i=1}^{n} d_{\mathrm{KL}}(p_{X_i, \Pi_i^G; \mathcal{A}'_i} \| q_{X_i, \Pi_i^G; \mathcal{A}'_i}) - d_{\mathrm{KL}}(p_{\Pi_i^G; \mathcal{A}'_i} \| q_{\Pi_i^G; \mathcal{A}'_i})
\end{aligned}$$

$$\sum_{i=1}^{n} d_{\mathrm{KL}}(p_{X_i, \Pi_i^G; \mathcal{A}'_i} \| q_{X_i, \Pi_i^G; \mathcal{A}'_i}) \geqslant \Omega(\varepsilon^2) + \sum_{i=1}^{n} d_{\mathrm{KL}}(p_{\Pi_i^G; \mathcal{A}'_i} \| q_{\Pi_i^G; \mathcal{A}'_i}).$$

$$d_{\mathrm{KL}}(p_{\Pi_i^G; \bar{\mathcal{A}}'_i} \| q_{\Pi_i^G; \bar{\mathcal{A}}'_i}) \geqslant -p_{\Pi_i^G}(\bar{\mathcal{A}}'_i) \cdot \log \left(\frac{q_{\Pi_i^G}(\bar{\mathcal{A}}'_i)}{p_{\Pi_i^G}(\bar{\mathcal{A}}'_i)} \right) \geqslant O(\varepsilon^2/n).$$

And we can conclude by rearranging:

$$\sum_{i=1}^{n} d_{\mathrm{KL}}(p_{X_i, \Pi_i^G; \mathcal{A}'_i} \| q_{X_i, \Pi_i^G; \mathcal{A}'_i}) \geqslant \Omega(\varepsilon^2)$$

□

D Proofs of entropy testing lower bounds

Lemma 2.6. *With fewer than $c_3 \cdot \sqrt{k}/\varepsilon$ samples from p, no tester can distinguish between $p = q$ and $|H(p) - H(q)| \geqslant \varepsilon$ with probability higher than $2/3$, where $c_3 > 0$ is an absolute constant.*

Proof of Lemma 2.6. This follows from the standard uniformity testing lower bound [Pan08], which provides a lower bound of $\Omega(\sqrt{k}/\eta^2)$: there exists a family of distributions that are hard to distinguish from uniform u_k, using fewer than $c_1 \cdot \sqrt{k}/\eta$ samples. Let k be an even number; the construction is by taking $\theta = \{-1, 1\}^{k/2}$ uniformly at random, and letting, for every $i \in [k/2]$,

$$p_\theta^{\mathrm{no}}(2i) = \frac{1 + \theta_i \cdot \eta}{k}, \qquad p_\theta^{\mathrm{no}}(2i+1) = \frac{1 - \theta_i \cdot \eta}{k}.$$

We can verify that for any θ:

$$|H(p_\theta^{\mathrm{no}}) - H(u_k)| = \log k - \frac{k}{2} \left(\frac{1+\eta}{k} \log \left(\frac{1+\eta}{k} \right) + \frac{1-\eta}{k} \log \left(\frac{1-\eta}{k} \right) \right) = \Theta(\eta^2)$$

Setting $\eta = \varepsilon^2$ yields the lower bound of $\Omega\left(\frac{\sqrt{k}}{\varepsilon}\right)$. □

Lemma 2.7. *With fewer than $c_4 \cdot \log^2 k/\varepsilon^2$ samples from p, no tester can distinguish between $p = q$ and $|H(p) - H(q)| \geqslant \varepsilon$ with probability higher than $2/3$, where $c_4 > 0$ is an absolute constant.*

Proof of Lemma 2.7. Following [WY16, B.2 Proof of Proposition 2], we look at the same construction but with different parameters $\varepsilon' = \frac{\varepsilon}{\log(2(k-1))}$:

$$p = \left(\frac{1}{3(k-1)}, \ldots, \frac{1}{3(k-1)}, \frac{2}{3}\right), \quad q = \left(\frac{1+\varepsilon'}{3(k-1)}, \ldots, \frac{1+\varepsilon'}{3(k-1)}, \frac{2-\varepsilon'}{3}\right).$$

One can check that

$$H(q) - H(p) \geqslant \frac{1}{3}\log(2(k-1))\varepsilon' - \varepsilon'^2 = \Omega(\varepsilon).$$

Moreover, direct calculation of the (squared) Hellinger distance shows that

$$\mathrm{d_H}(p,q)^2 = \Theta(\varepsilon'^2) = \Theta\left(\frac{\varepsilon^2}{\log^2 k}\right)$$

which implies that p and q cannot be distinguished with fewer than $c_4 \frac{\log^2 k}{\varepsilon^2}$ samples [BY02, Theorem 4.7]. □

E Sketch proof of $O\left(\frac{\sqrt{k \log\log\log(k/\varepsilon)}}{\varepsilon} + \frac{\log^2(k)}{\varepsilon^2}\right)$ upper bound.

We will rely the following inequality for compression, both via Jensen's inequality,

$$\left(\sum_{i\in\Delta} p_i\right) \log\left(\frac{1}{\sum_{i\in\Delta} p_i}\right) \leqslant \sum_{i\in\Delta} p_i \log\frac{1}{p_i} \leqslant \left(\sum_{i\in\Delta} p_i\right) \log\left(\frac{|\Delta|}{\sum_{i\in\Delta} p_i}\right), \tag{19}$$

as $\log(x)$ is concave and $\log\left(\frac{1}{x}\right)$ is convex.

$$\sum_{i\in\Delta} p_i \log\frac{1}{p_i} \geqslant \left(\sum_{i\in\Delta} p_i\right) \log\left(\frac{\sum_{i\in\Delta} p_i}{\sum_{i\in\Delta} p_i^2}\right) \geqslant \left(\sum_{i\in\Delta} p_i\right) \log\left(\frac{1}{\sum_{i\in\Delta} p_i}\right),$$

suggesting that if we merge elements of Δ into one, then we will lose a $\log(|\Delta|)$ factor of the entropy. By merging enough elements, we can then reduce this problem into the first case, where elements have large enough mass in each location.

Claim E.1. *Let $\mathcal{S} \subseteq [k]$, if $p(\mathcal{S}) - q(\mathcal{S}) > -\eta$, then*

$$|d_{\mathrm{KL}}(p_{\mathcal{S}} \| q_{\mathcal{S}})| \geqslant \eta \Rightarrow d_{\mathrm{KL}}(p_{\mathcal{S}} \| q_{\mathcal{S}}) \geqslant \eta.$$

Proof. If $|d_{\mathrm{KL}}(p_{\mathcal{S}} \| q_{\mathcal{S}})| \geqslant \eta$, then

$$d_{\mathrm{KL}}(p_{\mathcal{S}} \| q_{\mathcal{S}}) \geqslant \eta \text{ or } d_{\mathrm{KL}}(p_{\mathcal{S}} \| q_{\mathcal{S}}) \leqslant -\eta.$$

We will rule out the second case, using that $\log\frac{1}{x} \geqslant 1 - x$ for $x > 0$, [11]

$$d_{\mathrm{KL}}(p_{\mathcal{S}} \| q_{\mathcal{S}}) = \sum_{i\in\mathcal{S}} p_i \log\frac{p_i}{q_i} \geqslant \sum_{i\in\mathcal{S}} p_i \left(1 - \frac{q_i}{p_i}\right) = p(\mathcal{S}) - q(\mathcal{S}) > -\eta. \tag{20}$$

Thus, we cannot have $d_{\mathrm{KL}}(p_{\mathcal{S}} \| q_{\mathcal{S}}) \leqslant -\eta$ and so $d_{\mathrm{KL}}(p_{\mathcal{S}} \| q_{\mathcal{S}}) \geqslant \eta$. □

We use the same idea as our first upper bound but choose a series of thresholds. Let

$$\mathcal{S}_3 := \left\{ i \in [k] | q_i \geqslant \Omega\left(\frac{\varepsilon}{k \log\log\log(k/\varepsilon)}\right)\right\};$$

$$\mathcal{S}_2 = \left\{ i \in [k] | \Omega\left(\frac{\varepsilon}{k \log\log(k/\varepsilon)}\right) \leqslant q_i \leqslant O\left(\frac{\varepsilon}{k \log\log\log(k/\varepsilon)}\right)\right\};$$

[11] In the case of $p_i = 0$, we still have $p_i \log\left(\frac{p_i}{q_i}\right) \geqslant p_i - q_i$.

$$\mathcal{S}_1 = \left\{ i \in [k] | \Omega\left(\frac{\varepsilon}{k \log(k/\varepsilon)}\right) \leqslant q_i \leqslant O\left(\frac{\varepsilon}{k \log\log(k/\varepsilon)}\right) \right\}.$$

The following calculation ensues

$$\begin{aligned}
\Omega(\varepsilon) &\leqslant |H(p_{\mathcal{A}}) - H(q_{\mathcal{A}})| \\
&\leqslant \left| \sum_{i=1}^{3} d_{\mathrm{KL}}(p_{\mathcal{S}_i}, q_{\mathcal{S}_i}) \right| + \left| \sum_{i \in \mathcal{A}} (p_i - q_i) \log \frac{1}{q_i} \right| \\
&\leqslant \sum_{i=1}^{3} |d_{\mathrm{KL}}(p_{\mathcal{S}_i}, q_{\mathcal{S}_i})| + \left| \sum_{i \in \mathcal{A}} (p_i - q_i) \log \frac{1}{q_i} \right|.
\end{aligned}$$

We have that one of the four terms will be at least $\Omega(\varepsilon/4)$. If it is

$$\left| \sum_{i \in \mathcal{A}} (p_i - q_i) \log \frac{1}{q_i} \right| \geqslant \Omega(\varepsilon),$$

which is testable with $O\left(\frac{\log^2(k)}{\varepsilon^2}\right)$ samples using arguments from proof of Theorem 2.1. If it is $|d_{\mathrm{KL}}(p_{\mathcal{S}_i}, q_{\mathcal{S}_i})| \geqslant \Omega(\varepsilon)$, for $i = 1, 2, 3$. We have the following:

Case $\mathcal{S}_3$. Suppose $|d_{\mathrm{KL}}(p_{\mathcal{S}_3}, q_{\mathcal{S}_3})| \geqslant \Omega(\varepsilon)$. We will check whether $p(\mathcal{S}_3) \geqslant \Omega\left(\frac{\varepsilon}{\log\log\log(k/\varepsilon)}\right)$, if not, we can reject. We proceed assuming the inequality holds. Note that

$$|p(\mathcal{S}_3) - q(\mathcal{S}_3)| = |p(\overline{\mathcal{S}_3}) - q(\overline{\mathcal{S}_3})| \leqslant O\left(\frac{\varepsilon}{\log\log\log(k/\varepsilon)}\right).$$

Thus, $p(\overline{\mathcal{S}_3}) - q(\overline{\mathcal{S}_3}) > -O\left(\frac{\varepsilon}{\log\log\log(k/\varepsilon)}\right)$, and by Claim E.1, we have that $d_{\mathrm{KL}}(p_{\mathcal{S}_3}, q_{\mathcal{S}_3}) \geqslant \Omega(\varepsilon)$. Using (4), we then have that $d_{\chi^2}(p_{\mathcal{S}_3}, q_{\mathcal{S}_3}) \geqslant \Omega(\varepsilon)$. Using Lemma 2.4 (setting $\alpha = O\left(\frac{\varepsilon}{k \log\log\log(k/\varepsilon)}\right)$), and similar argument from the proof of Theorem 2.1, we have that $O\left(\frac{\sqrt{k \log\log\log(k/\varepsilon)}}{\varepsilon}\right)$ suffices to check between the case that $d_{\chi^2}(p_{\mathcal{S}_3}, q_{\mathcal{S}_3}) \geqslant \Omega(\varepsilon)$ and $p_{\mathcal{S}_3} = q_{\mathcal{S}_3}$.

Case $\mathcal{S}_2$. Suppose $|d_{\mathrm{KL}}(p_{\mathcal{S}_2}, q_{\mathcal{S}_2})| \geqslant \Omega(\varepsilon)$. We will check whether

$$\Omega\left(\frac{\varepsilon}{\log\log(k/\varepsilon)}\right) \leqslant p(\mathcal{S}_2) \leqslant O\left(\frac{\varepsilon}{\log\log\log(k/\varepsilon)}\right),$$

if not, we will reject. We proceed assuming the inequality holds. Now, recall that the main bottleneck of the χ^2 tester analyzed in Lemma 2.4 is due to the minimum probability $\alpha = \min_{i \in \mathcal{S}_2} q_i$ (increasing this would decrease the sample complexity). Our main idea here is to increase α by merging a suitable number ($\log\log(k/\varepsilon)$ in this case) of elements into one single bin to form a new distribution to test. Denote Δ_j where $j \in \left[\frac{|\mathcal{S}_2|}{\log\log(k/\varepsilon)}\right]$ and $\bigcup_j \Delta_j = \mathcal{S}_2$. We will subsequently treat every elements in Δ_j as 1 bin in the new distribution, calling it p_Δ, q_Δ and denote $p(\Delta_j), q(\Delta_j)$ as mass on Δ_j, where $p(\Delta_j) = \sum_{i \in \Delta_j} p_i$. This gives us the following:

i. $q(\Delta_j) \geqslant \Omega\left(\frac{\varepsilon}{k \log\log(k/\varepsilon)}\right) \cdot |\Delta_j| \geqslant \Omega\left(\frac{\varepsilon}{k}\right)$; the domain size is $\frac{|\mathcal{S}_2|}{\min_j |\Delta_j|} \leqslant O\left(\frac{k}{\log\log(k/\varepsilon)}\right)$.

ii. $\sum_j p(\Delta_j) = p(\mathcal{S}_2) \leqslant O\left(\frac{\varepsilon}{\log\log\log(k/\varepsilon)}\right)$ and $\sum_j q(\Delta_j) = q(\mathcal{S}_2) \leqslant O\left(\frac{\varepsilon}{\log\log\log(k/\varepsilon)}\right)$.

iii. Their entropy difference is preserved, which we will prove next:

$$\left| \sum_j p(\Delta_j) \log \frac{1}{p(\Delta_j)} - \sum_j q(\Delta_j) \log \frac{1}{q(\Delta_j)} \right| \geqslant \Omega(\varepsilon).$$

Note that these are better conditions compared to i. and ii. in the proof of Theorem 2.1 (in this analysis, using ii., it is sufficient to prove that $d_{\mathrm{KL}}(p_\Delta, q_\Delta) \geqslant \Omega(\varepsilon)$ in view of Claim E.1). The gain comes from the fact that we can apply Lemma 2.4 with better $\alpha = \min_j q(\Delta_j) \geqslant \Omega\left(\frac{\varepsilon}{k}\right)$ and thus

$$O\left(\sqrt{\frac{1}{\alpha\varepsilon}} + \sqrt{\frac{k'}{\varepsilon}}\right) = O\left(\sqrt{\frac{k}{\varepsilon^2}}\right) = O\left(\frac{\sqrt{k}}{\varepsilon}\right).$$

However, the gain only affect Claim 2.3 by constant factors. The soundness and completeness then follows similarly to the proof of Theorem 2.1. We prove (iii.) next:

Suppose $H(p_{\mathcal{S}_2}) - H(q_{\mathcal{S}_2}) \geqslant \varepsilon$, then,

$$\begin{aligned}
\Omega(\varepsilon) &\leqslant \sum_{l\in\mathcal{S}_2} p_l \log\frac{1}{p_l} - \sum_{l\in\mathcal{S}_2} q_l \log\frac{1}{q_l} \\
&= \sum_j \sum_{i\in\Delta_j} p_{i,j} \log\frac{1}{p_{i,j}} - \sum_j \sum_{i\in\Delta_j} q_{i,j} \log\frac{1}{q_{i,j}} \\
&\leqslant \sum_j p(\Delta_j) \log\frac{|\Delta_j|}{p(\Delta_j)} - \sum_j q(\Delta_j) \log\frac{1}{q(\Delta_j)} \qquad (21) \\
&= \sum_j p(\Delta_j) \log|\Delta_j| + \sum_j p(\Delta_j) \log\frac{1}{p(\Delta_j)} - \sum_j q(\Delta_j) \log\frac{1}{q(\Delta_j)}. \\
&\leqslant O\left(\frac{\varepsilon}{\log\log\log(k/\varepsilon)}\right) \max_j \log|\Delta_j| + \sum_j p(\Delta_j) \log\frac{1}{p(\Delta_j)} - \sum_j q(\Delta_j) \log\frac{1}{q(\Delta_j)} \qquad (22)
\end{aligned}$$

where the (21) is due to (19) and for (22), recall that $\sum_j p(\Delta_j) = p(\mathcal{S}_2) \leqslant O\left(\frac{\varepsilon}{\log\log\log(k/\varepsilon)}\right)$.

Suppose $H(q_{\mathcal{S}_2}) - H(p_{\mathcal{S}_2}) \geqslant \Omega(\varepsilon)$, the same goes below:

$$\begin{aligned}
\Omega(\varepsilon) &\leqslant \sum_l q_l \log\frac{1}{q_l} - \sum_l p_l \log\frac{1}{p_l} \\
&= \sum_j \sum_{i\in\Delta_j} q_{i,j} \log\frac{1}{q_{i,j}} - \sum_j \sum_{i\in\Delta_j} p_{i,j} \log\frac{1}{p_{i,j}} \\
&\leqslant \sum_j q(\Delta_j) \log\frac{|\Delta_j|}{q(\Delta_j)} - \sum_j p(\Delta_j) \log\frac{1}{p(\Delta_j)} \\
&\leqslant O(\varepsilon) + \sum_j q(\Delta_j) \log\frac{1}{q(\Delta_j)} - \sum_j p(\Delta_j) \log\frac{1}{p(\Delta_j)}.
\end{aligned}$$

Therefore, we have proved (iii.).

Case $\mathcal{S}_1$. The proof follow similar to Case $\mathcal{S}_2$, but by merging $\log(k/\varepsilon)$ elements.

F Derivation for KL decomposition

$$
\begin{aligned}
& d_{\mathrm{KL}}(p, p_G) \\
= & \sum_{x \in \{0,1\}^n} p(x) \log \frac{p(x)}{p_G(x)} \\
= & \sum_{x \in \{0,1\}^n} p(x) \log \frac{\prod_{i=1}^n p(x_i|\pi_i)}{\prod_{i=1}^n p_G(x_i|\pi_i^G)} \\
= & \sum_{x \in \{0,1\}^n} p(x) \log \left(\prod_{i=1}^n p(x_i|\pi_i) \right) - p(x) \log \left(\prod_{i=1}^n p_G(x_i|\pi_i^G) \right) \\
= & \sum_{x \in \{0,1\}^n} \sum_{i=1}^n p(x) \log(p(x_i|\pi_i)) - p(x) \log(p_G(x_i|\pi_i^G)) \\
= & \sum_{i=1}^n \sum_{x \in \{0,1\}^n} p(x) \log(p(x_i|\pi_i)) - p(x) \log(p_G(x_i|\pi_i^G)) \\
= & \sum_{i=1}^n \left(\sum_{x_i, \pi_i \in X_i, \Pi_i} p(x_i, \pi_i) \log(p(x_i|\pi_i)) \right) - \left(\sum_{x_i, \pi_i \in X_i, \Pi_i^G} p(x_i, \pi_i^G) \log(p_G(x_i|\pi_i^G)) \right) \\
= & \sum_{i=1}^n H(p_{X_i, \Pi_i^G} | p_{\Pi_i^G}) - H(p_{X_i, \Pi_i} | p_{\Pi_i}),
\end{aligned}
$$

where π_i, Π_i denote the parents of x_i, X_i in Bayes net p (a set of random variables or their domain); and π_i^G, Π_i^G as the parents defined by G. p_G is the projection of p unto G as defined by Definition 3.2. It is not hard to see that the derivation extends beyond the case of hypercube, $\{0,1\}^n$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The proofs are provided either in the main paper or the appendix, both part of the submission. Assumptions are fully stated in the theorem and lemma statements.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper does not contain experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

Justification:

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: this work is theoretical in nature; it is hard to predict its societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the

technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.