

Entropy Equivalence Testing

Clément L. Canonne ✉ 

University of Sydney, Australia

Yash Pote ✉ 

National University of Singapore, Singapore

Jonathan Scarlett ✉ 

National University of Singapore, Singapore

Joy Qiping Yang ✉ 

University of Sydney, Australia

Abstract

We introduce the problem of *entropy equivalence testing* for probability distributions, a relaxation of the well-studied closeness testing problem, where the distribution testing algorithm is now only required to distinguish, given samples from two unknown distributions p, q and a parameter $\varepsilon \in (0, 1/2]$, between $p = q$ and $|H(p) - H(q)| \geq \varepsilon$ (where H denotes the Shannon entropy). We provide a time- and sample-efficient algorithm for this task, showing that the optimal sample complexity for this task can be significantly lower than that of closeness testing. As an application, we leverage this result to provide the first non-trivial testing algorithm for (standard) closeness of low-degree *Bayesian networks*, which significantly improves on either the sample or time complexity of a baseline based on full learning.

2012 ACM Subject Classification Theory of computation → Lower bounds and information complexity; Mathematics of computing → Probabilistic algorithms

Keywords and phrases Entropy, distribution testing, sublinear algorithm, Bayesian network

Digital Object Identifier 10.4230/LIPIcs.APPROX/RANDOM.2026.50

Category RANDOM

Related Version *Full Version*: <https://arxiv.org/abs/2605.23225> [14]

Funding *Jonathan Scarlett*: Supported by the National University of Singapore under the Presidential Young Professorship scheme.

Joy Qiping Yang: Supported by a JD Technology Research Scholarship in Artificial intelligence.

Acknowledgements We would like to thank Sayantan Sen, Jan Seyfried and Marco Tomamichel for the discussion on their work of testing Conditional Mutual Information. JY would also like to thank Davin Choo for the early discussions predating this project. This work is conducted while JY was interning at National Institute of Informatics, hosted by Yuichi Yoshida.

1 Introduction

Distribution testing, a subfield of property testing [35, 30] concerned with efficient algorithms for deciding statistical properties of data, has witnessed a significant interest since its systematic introduction by Batu, Fortnow, Rubinfeld, Smith, and White [5], with a flurry of works over the past two decades addressing a range of hypothesis testing tasks, under various models of distribution testing (see, e.g., [34, 10, 11] and [29, Chapter 11] for surveys and expositions).

Two of the flagship and most fundamental tasks in distribution testing are *identity testing* (and its special case, *uniformity testing*), and *closeness testing* (also known as *equivalence testing*): in the first, an algorithm is given independent samples from an (unknown) discrete



© Clément L. Canonne, Yash Pote, Jonathan Scarlett, and Joy Qiping Yang;
licensed under Creative Commons License CC-BY 4.0

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2026).

Editors: Mohit Singh and Tom Gur; Article No. 50; pp. 50:1–50:23



Leibniz International Proceedings in Informatics

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

probability distribution, and must decide whether the distribution is equal to some known reference distribution, or significantly deviates from it in a meaningful notion of distance (typically total variation distance). In the second, the algorithm has access to samples from *two* unknown distributions, and similarly must decide whether they are equal, or significantly far from each other. Here, “significantly far” is quantified by an input parameter $\varepsilon > 0$, and the algorithm is required to be correct with high probability $1 - \delta$,¹ while minimizing the *sample complexity*, that is, the total number of samples taken from the distribution(s) in the worst case as a function of the domain size n , distance parameter ε , and error probability δ .

The sample complexity of these two flagship tasks is by now well understood, with several time-efficient algorithms known to achieve the optimal sample complexity, as a function of all parameters: namely,

$$\Theta\left(\frac{\sqrt{n \log(1/\delta)} + \log(1/\delta)}{\varepsilon^2}\right)$$

for identity testing [32, 1, 28, 27], and

$$\Theta\left(\frac{n^{2/3} \log^{1/3}(1/\delta) + \log(1/\delta)}{\varepsilon^{4/3}} + \frac{\sqrt{n \log(1/\delta)} + \log(1/\delta)}{\varepsilon^2}\right)$$

for closeness testing [5, 17, 28, 26]. However, in spite of these impressive advances, these sample complexities can remain prohibitively large for many combinations of (n, ε) , notably for small ε corresponding to distributions that may be very close to one another. Given that their sample complexities are constant-factor *optimal*, there would seem to be little room for improvement, yet two possible avenues remain. The first is to *relax the guarantees*: total variation is known to be a very stringent measure of distance, and testing with respect to a more permissive distance measure may enable significant sample complexity savings, while still being “good enough” for the application at hand. This direction was partially explored in, for instance, [24]. The second avenue, particularly well suited to high-dimensional data, is to *make additional structural assumptions* on the distribution: high-dimensional data is seldom fully unstructured, and one can usually leverage domain knowledge to make modelling assumptions, for instance that the probability distributions exhibit a low-degree Bayesian network structure, or are consistent with an Ising model. This line of work, in the context of distribution testing, was initiated in [12, 23].

Our work makes progress on both of these fronts. We first define a relaxation of closeness testing, which we term *entropy equivalence testing*, where one is asked to detect if the two unknown distributions have significantly different (Shannon) entropy:²

Entropy Equivalence Testing (EET): Given sample access to distributions p, q over a known domain of size n and parameters $\varepsilon \in (0, 1/2]$ and $\delta \in (0, 1]$, distinguish, with probability at least $1 - \delta$, between the following two cases: (1) $p = q$, and (2) $|H(p) - H(q)| \geq \varepsilon$.

¹ One typically sets δ to be a small constant, e.g., $\delta = 1/10$: this is without much loss of generality, as one can amplify this by repetition to any target $1 - \delta$, at the cost of an $O(\log(1/\delta))$ factor in sample and time complexity.

² Recall that the (Shannon) entropy of a probability distribution p over a discrete domain $\mathcal{X}$ is defined as $H(p) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}$.

We note that the “single-distribution” version of this task, when q is fully known to the algorithm, was recently introduced [15], under the name Entropy Identity Testing (EIT). [15] then leveraged their algorithm for EIT to obtain an improved identity testing algorithm for “standard” identity testing (i.e., in total variation distance) of *bounded-in-degree Bayes nets*, establishing a connection between the two aforementioned avenues of research. We build upon and deepen this connection to obtain, from our algorithm for EET, the first non-trivial *closeness* testing algorithm for Bayes nets.

1.1 Our results

We now describe our results in more detail. Our first contribution is an efficient algorithm for the EET task defined above (where we implicitly set the probability of failure δ to be a small constant):

► **Theorem 1.** *There exists an algorithm for the Entropy Equivalence Testing problem with sample complexity*

$$\tilde{O}\left(\frac{n^{3/4}}{\varepsilon}\right) + O\left(\frac{\log^2(n/\varepsilon)}{\varepsilon^2}\right)$$

and running time near-linear in the number of samples.

Combining Theorem 1 and the “baseline tester” discussed below (see Section 1.2) which follows from a reduction from entropy difference to total variation distance and has sample complexity $\tilde{O}\left(\max\left(\frac{n^{2/3}}{\varepsilon^{4/3}}, \frac{\sqrt{n}}{\varepsilon^2}\right)\right)$, one gets, after simplifying the resulting sample complexity, our main result:

► **Corollary 2 (Main testing result).** *There exists an algorithm for the Entropy Equivalence Testing problem with sample complexity*

$$\tilde{O}\left(\min\left\{\frac{n^{3/4}}{\varepsilon}, \frac{n^{2/3}}{\varepsilon^{4/3}}\right\}\right) + O\left(\frac{\log^2(n/\varepsilon)}{\varepsilon^2}\right)$$

and running time near-linear in the number of samples.

We complement this positive result with an information-theoretic lower bound, showing that our algorithm has nearly optimal sample complexity:

► **Theorem 3 (Lower bound).** *Any algorithm for the Entropy Equivalence Testing problem must draw*

$$\tilde{\Omega}\left(\min\left\{\frac{n^{3/4}}{\varepsilon}, \frac{n^{2/3}}{\varepsilon^{4/3}}\right\}\right) + \Omega\left(\frac{\log^2 n}{\varepsilon^2}\right)$$

total samples from p and q .

Finally, we show how to leverage our algorithm for EET to obtain a closeness testing algorithm for bounded in-degree Bayes nets: we emphasize that, to the best of our knowledge, this is the *first* testing algorithm for this problem, besides the “naïve” (and computationally inefficient) testing-by-learning approach that fully learns the two unknown distributions with $\tilde{O}(2^d n / \varepsilon^2)$ samples, in time $2^{\Omega(n)}$:

► **Theorem 4.** *There exists an algorithm for closeness testing (in total variation distance) of in-degree d Bayesian networks with sample complexity*

$$s = \tilde{O}\left(\min\left\{\frac{2^{3d/4}n}{\varepsilon^2}, \frac{2^{2d/3}n^{4/3}}{\varepsilon^{8/3}}\right\} + \frac{n^2}{\varepsilon^4}\right),$$

and running time $\tilde{O}(n^{d+1}s)$.

Interestingly, the techniques underlying Theorem 4 can also be easily adapted to obtain an algorithm for identity testing of Bayes nets, improving upon the work of [15]: see Corollary 18.

1.2 Technical overview

We now provide an overview of the techniques and main ideas underlying our results.

1.2.1 Upper bound for EET

A bound of $\tilde{O}\left(\max\left\{\frac{n^{2/3}}{\varepsilon^{4/3}}, \frac{\sqrt{n}}{\varepsilon^2}\right\}\right)$ readily follows from a standard inequality, allowing one to reduce the entropy difference problem to closeness testing in TV: namely, when $d_{\text{TV}}(p, q) \leq \frac{1}{2}$, then $|H(p) - H(q)| \leq d_{\text{TV}}(p, q) \log \frac{n}{d_{\text{TV}}(p, q)}$ [21, Lemma 2.7]. Another straightforward upper bound is obtained by estimating the entropy, a harder task, which takes $\Theta\left(\frac{n}{\varepsilon \cdot \log n} + \frac{\log^2 n}{\varepsilon^2}\right)$ samples [42]. This hints at the possibility of getting a better bound of $O(n^c/\varepsilon + \log^2 n/\varepsilon^2)$ for some absolute constant $c < 1$. Surprisingly, the final upper bound of $\tilde{O}\left(\min\left\{\frac{n^{2/3}}{\varepsilon^{4/3}}, \frac{n^{3/4}}{\varepsilon}\right\} + \frac{1}{\varepsilon^2}\right)$ is related to the squared Hellinger distance (and not the total variation). This is a result of considering two testers that can solve EET independently, and choosing the one with lower sample complexity. The main contribution of our work is providing an alternative tester to the TV reduction tester above.

An $\tilde{O}(\frac{n^{3/4}}{\varepsilon} + \frac{1}{\varepsilon^2})$ tester. Our analysis starts with a reduction to testing both farness (vs. equality) in Hellinger distance and a cross-entropy difference term (with respect to the uniform mixture) via some algebraic manipulations (see Lemma 9). The problem thus becomes testing whether Hellinger distance is large (which we know costs $O(\min\{\frac{n^{3/4}}{\varepsilon}, \frac{n^{2/3}}{\varepsilon^{4/3}}\}) \leq O(\frac{n^{3/4}}{\varepsilon})$) and testing whether the cross-entropy term is large (since one of them will be large if the entropy difference is large). A natural idea is of course to employ a plug-in estimator for the cross entropy difference (see (1)). Unfortunately, this empirical estimator has an inherent³ bias of the form $\sum_{i \in [n]} \frac{|p_i - q_i|}{m(p_i + q_i)}$, where m is the (equal) number of samples we take from p and q to compute the estimate. Our argument to deal with the bias is as follows: If the expectation of the plug-in estimator has *low bias* (at most $O(\varepsilon)$), then we can rely on its estimate alone. On the other hand, if it has *high bias* (at least $\Omega(\varepsilon)$), we hope to catch it and reject (i.e., declare that $|H(p) - H(q)| > \epsilon$) early.

Low mass regime: Reduction to TV. One main challenge to designing this test is that the bias is impossible to test without a lower bound on $p_i + q_i$ (the variance of estimating the bias could be huge). The key to passing this hurdle is an idea inspired by a different algorithm from [28], whose analysis implies that for the elements $i \in [n]$ where q_i (or p_i) is small, we can perform tests more efficiently (in TV) via access to its conditional distribution (as efficiently as $\tilde{O}(n^{3/4}/\varepsilon)$), with the intuition being: The TV distance between the conditional distributions is actually larger (inverse of its mass size $\tilde{\Omega}(\varepsilon/q(T))$, where T is the set of small-mass elements). Via a reduction to TV from cross-entropy difference, we can also test farness on these elements with $\tilde{O}(n^{3/4}/\varepsilon)$ samples.

³ In the works of [42], their plug-in estimator also exhibits similar bias in the analysis, and they resort to machinery from polynomial approximation to reduce the bias (which will increase the variance the higher the degree of the polynomial).

Detecting bias in the high mass regime. Capitalizing on this, we will consider exclusively elements with large mass: $p_i \geq \tilde{\Omega}(\varepsilon/n^{3/4})$. With a bound on $m(p_i + q_i) \geq \Omega(1)$ (since we take $m = \tilde{O}(\frac{n^{3/4}}{\varepsilon})$), this reduces to testing fairness in TV distance, which can be performed with $O(\sqrt{n}/\varepsilon^2)$ samples (to bound the bias term by $O(\varepsilon)$), rather than the $O(n^c/\varepsilon)$ for some $c < 1$ that we set out for. Luckily our problem can be easier (depending on relation between n and ε) than testing in TV (see Lemma 13). Intuitively, this is related to testing in the instance-optimal setting [40, 28] – one can test with sample complexity as a function of p, q (for our case p, q has exclusively large probability masses), rather than as a function of the domain size (the worst case over all possible p, q). Interestingly, to detect the bias, we are using the same test statistic (with different threshold) first used in [17] to test in TV distance and subsequently in [24] to test in Hellinger distance.

Had we not made a “reject” decision after the bias check in the high mass regime (nor rejected due to its low mass cross-entropy difference), it would suggest that the bias on the high mass regime is at most $O(\varepsilon)$ and therefore, the estimator for $\sum_i (p_i - q_i) \log((p_i + q_i)/2)$ is reliable: A standard analysis of expectation and variance followed by an application of Chebyshev’s inequality then yields Theorem 1.

1.2.2 Lower bound for EET

The lower bound is a combination of $\tilde{\Omega}(\min\{\frac{n^{2/3}}{\varepsilon^{4/3}}, \frac{n^{3/4}}{\varepsilon}\})$ and $\Omega(\frac{\log^2 n}{\varepsilon^2})$. The former is derived via a reduction from testing mutual information to entropy difference. The task of testing mutual information (MI) is as follows: if $I(X : Y) = 0$, then accept with probability 0.99; If $I(X : Y) \geq \varepsilon$, then reject with probability 0.99. Given the close relation between mutual information and entropy: $I(X : Y) = H(X) + H(Y) - H(X, Y)$, one can reduce the problem of MI testing to EET by treating p as the joint distribution of X, Y and q as the product of X and Y (we can simulate sample access to the product of X and Y with two samples from the joint distribution). Therefore, the existing lower bound from [37, Corollary 6.2] holds for our problem. The latter lower bound of $\Omega(\frac{\log^2 n}{\varepsilon^2})$ comes from [15, Lemma 2.7], given the reduction from identity testing to equivalence testing.

1.2.3 Application to equivalence testing of Bayes nets

As laid out by [15], one can utilize testing entropy difference as a subroutine to testing Bayes nets. The key of the analysis is derived via the celebrated Chow-Liu decomposition [20]: If two Bayes nets p and q have similar local entropies, then they must share approximately similar structures (their corresponding DAGs are close in some sense; see [15, Lemma 3.4]). Once we confirm that they share a similar structure, the problem boils down to checking local (every subset of random variables of size at most $d + 1$) KL-divergences. While KL-divergences are known to be untestable [24, Theorem 7] in the extreme cases when $q(x) = 0$, there is hope if one can obtain a lower bound on $q(x)$. In fact, we can reduce testing in KL to testing in Hellinger distance with an $O\left(\max_x \log \frac{p(x)}{q(x)}\right)$ factor overhead using [37, Lemma 3.1]. To get a lower bound on $q(x)$, we use a standard technique of testing the mixture of q and the uniform distribution.

1.3 Related works

Estimating entropy. Entropy estimation is a well-studied problem in both the information theory [41, 42, 31] and property testing [39, 3] communities, with the near-optimal sample complexity $\Theta(\frac{n}{\varepsilon \log n} + \frac{\log^2 n}{\varepsilon^2})$ being eventually established by [42, 31] (for discrete distributions).

As we are considering the testing problem, our upper bound is less costly compared to its learning (estimation) variant. Moreover, other variants of entropy have been studied, such as Rényi entropy [2]. The more general problem of estimating functionals of distributions was also considered by [31].

Identity testing and closeness testing. The standard formulation of identity testing asks the following problem: Given a full description of a hypothesis q , how many i.i.d. samples are needed from p to distinguish between $p = q$ and $d_{\text{TV}}(p, q) \geq \varepsilon$, where p and q are both distributions over support of size $[n]$? This problem is well-known to have a sample complexity of $\Theta(\sqrt{n}/\varepsilon^2)$, a much lower cost compared to fully learning every parameter of p , which takes $\Theta(n/\varepsilon^2)$ samples. A similar problem can be asked if one is instead only provided sample access to p and q , both unknown. This is the closeness testing problem (in TV) and it is known that a $\Theta\left(\max\left\{\frac{\sqrt{n}}{\varepsilon^2}, \frac{n^{2/3}}{\varepsilon^{4/3}}\right\}\right)$ sample complexity is necessary and sufficient, making this a slightly harder problem than identity testing while still saving on samples compared to fully learning p and q . Our problem differs from the standard formulation by testing in the entropy difference rather than in TV distance. Most relevant to our work is the work of [15], which solves the problem of entropy identity testing (up to polylogarithmic factors) with a sample complexity of $\tilde{\Theta}\left(\frac{\sqrt{n}}{\varepsilon} + \frac{1}{\varepsilon^2}\right)$. Another relevant work is that of [24], which asks the question of testing in different distances, e.g., distinguishing between $d_{\text{KL}}(p||q) \leq \varepsilon^2/100$ and $d_{\text{H}}(p, q) \geq \varepsilon$ for identity testing (sample access to p and the full description of q). Interestingly, as shown in the same work, equivalence testing in squared Hellinger distance, i.e., distinguishing between $p = q$ and $d_{\text{H}}^2(p, q) \geq \varepsilon$ has a similar cost of $\Theta\left(\min\left\{\frac{n^{2/3}}{\varepsilon^{4/3}}, \frac{n^{3/4}}{\varepsilon}\right\}\right)$ (though perhaps not so surprising, as we are reducing to testing in squared Hellinger square in one of the cases we handle). Our work adds to their results for equivalence testing in entropy difference.

Testing mutual information. Testing mutual information (MI) asks the following question: Given sample access to a (discrete) distribution p over two random variables X, Y with a domain of size $|\Sigma_X|$ and $|\Sigma_Y|$ respectively, how many samples do we need from p to distinguish between $I(X : Y) = 0$ and $I(X : Y) \geq \varepsilon$? The study of this problem was initiated by [9] to solve the problem of learning a special variant of Bayes net (trees). While their MI tester is sufficient for learning trees-structured distributions near-optimally, their upper bound is loose in terms of dependence on $|\Sigma_X|$ and $|\Sigma_Y|$: $O\left(\frac{|\Sigma_X| \cdot |\Sigma_Y|}{\varepsilon}\right)$. More recently, [37] gave a near-optimal bound of $\tilde{\Theta}\left(\min\left\{\frac{|\Sigma_X|^{3/4} |\Sigma_Y|^{1/4}}{\varepsilon}, \frac{|\Sigma_X|^{2/3} |\Sigma_Y|^{1/3}}{\varepsilon^{4/3}}\right\}\right)$ (assuming $|\Sigma_X| \geq |\Sigma_Y|$). We note that this bound bears similarity (e.g., setting $|\Sigma_Y| = O(1)$) to ours due to the close relation between MI and entropy: $I(X : Y) = H(X) + H(Y) - H(X, Y)$. Indeed, MI testing can be reduced to EET naturally (as outlined above). Before these works, [4, 13] studied the problem of (conditional) independence testing in the TV distance.

Testing Bayesian networks. A Bayesian network (Bayes net) is a high-dimensional distribution over n random variables that factorizes into n conditional probabilities: $p(X_1, \dots, X_n) = \prod_{i=1}^n p(X_i | \Pi_i)$, where $\Pi_i \in [n]$ is called the parent set of X_i . This factorization is associated with a directed acyclic graph (DAG) $G = (V, E)$ ($V = \{X_1, \dots, X_n\}$), and if $X_j \in \Pi_i$, then the edge $(X_j, X_i) \in E$ (with X_j being one of X_i 's parents). A distribution p is called a Bayes net if its joint probability density function factorizes according to G and p is called Markov with respect to G . Learning Bayesian networks from samples is a fundamental problem in learning theory, and there is a plethora of evidence that this problem is hard computationally [18, 22, 7]. On the other hand, time-efficient algorithms (polynomial in n , the number of random variables) do exist in some special cases [19, 8, 9].

Identity testing and closeness testing of Bayes nets were systematically studied by [12], though their stronger upper bounds only consider relaxed version of Bayes nets where both p and q are required be *balanced* and *non-degenerate*. While they also provide a more general upper bound (without assuming balancedness and non-degeneracy for Bayes nets), they still require that the node ordering be the same between p and q , a non-trivial assumption. Our upper bound does not make any assumption other than the bounded in-degree d , which is a natural and necessary parameter to consider, since testing unbounded degree Bayes nets is exponentially hard: $\Omega(\frac{2^{n/2}}{\varepsilon^2})$ (e.g., see [6, Theorem B.1]). The strongest upper bound that we are aware of is by [25], which can solve the identity testing⁴ problem in $\tilde{O}(\frac{2^{d/2}n}{\varepsilon^2})$ samples,⁵ under the assumption that p and q have the same structure (Markov with respect to the same DAG). For equivalence testing, under the same assumption, they give a bound of $\tilde{O}(\min\{\frac{2^{3d/4}n}{\varepsilon^2}, \frac{2^{2d/3}n^{4/3}}{\varepsilon^{8/3}}\})$.⁶ Our results match their bounds if $d \gg \log(n/\varepsilon)$, and our algorithm will work without the extra assumption that p and q share the same structure. A lower bound of $\Omega(\frac{2^{d/2}n}{\varepsilon^2})$ when $d \leq \log n$ [6, Theorem 4.1] holds for identity testing of Bayes nets and thus also for its closeness testing variant (our setting).

1.4 Preliminaries

Define the factorial moment by $(x)_m := \frac{x!}{(x-m)!}$, which gives an unbiased estimator for the monomials of the Poisson mean: namely, if $X \sim \text{Poi}(\lambda)$, then $\mathbb{E}[(X)_m] = \lambda^m$ for all $m \geq 0$. More generally, we have the following fact, whose proof we provide in Appendix A for completeness:

► **Fact 5.** *Let X be a Poisson random variable with parameter λ , and $f: \mathbb{R} \rightarrow \mathbb{R}$ be any function, then, we have that $\mathbb{E}[(X)_m \cdot f(X)] = \lambda^m \cdot \mathbb{E}[f(X + m)]$. In particular, choosing $f \equiv 1$, we have that $\mathbb{E}[(X)_m] = \lambda^m$ for all $m \geq 0$.*

► **Fact 6 (Folklore).** *For any $0 < \alpha \leq 1/2$ and $\varepsilon \in (0, 1]$, testing whether the bias of a coin is at most α or at least $\alpha(1 + \varepsilon)$, with probability at least $1 - \delta$, can be done with $n = O(\frac{1}{\alpha\varepsilon^2} \log \frac{1}{\delta})$ samples.*

Distances and norms. We will use a number of standard distance measures for two distributions p, q over $[n]$, and various known inequalities between them, e.g., see [36]. The *squared Hellinger distance* is

$$d_H^2(p, q) := \frac{1}{2} \sum_{i \in [n]} (\sqrt{p_i} - \sqrt{q_i})^2$$

The *total variation distance* is $d_{\text{TV}}(p, q) := \frac{1}{2} \|p - q\|_1 = \frac{1}{2} \sum_{i \in [n]} |p_i - q_i|$. The ℓ_2 distance is $\|p - q\|_2 := (\sum_{i \in [n]} (p_i - q_i)^2)^{1/2}$, and the χ^2 -divergence of p from q is $d_{\chi^2}(p, q) := \sum_{i \in [n]} (p_i - q_i)^2 / q_i$. These satisfy the chain of inequalities

$$d_{\text{TV}}(p, q) \leq \sqrt{2} d_H(p, q) \quad \text{and} \quad d_H^2(p, q) \leq d_{\text{TV}}(p, q) \leq \sqrt{d_{\chi^2}(p, q)},$$

and the equivalence $\frac{1}{2} \sum_i (p_i - q_i)^2 / (p_i + q_i) = \Theta(d_H^2(p, q))$. The *Kullback-Leibler (KL) divergence* is $d_{\text{KL}}(p \| q) := \sum_i p_i \log(p_i / q_i)$, which satisfies $d_{\text{KL}}(p \| q) \leq d_{\chi^2}(p, q)$ and Pinsker's inequality $d_{\text{TV}}(p, q) \leq \sqrt{d_{\text{KL}}(p \| q) / 2}$.

⁴ Note that their setting of identity testing actually corresponds to equivalence testing in our setting.

⁵ Implicitly, using the Hellinger testing result of [24, Theorem 1].

⁶ Implicitly, using the optimal Hellinger equivalence test [24].

Notation. In what follows, for a given integer $n \geq 1$, we denote by $[n]$ the set $[n] = \{1, 2, \dots, n\}$, and by U_n the uniform distribution on $[n]$. When clear from context, we will omit the subscript, and write U instead of U_n . Given a probability distribution p over $[n]$ and a set $S \subseteq [n]$ such that $p(S) > 0$ we write $p(\cdot | S)$ for the conditional distribution of p on S . For the high dimensional (Bayes net) section, we use p_T to denote the marginalization of p over T , where T is a subset of $\{X_1, \dots, X_n\}$, e.g., p_{X_1, X_2} denotes the marginal distribution of (X_1, X_2) . We also use Π_i to denote the parent set of the random variable X_i in a Bayes net.

Throughout the paper, we use $\log$ for the natural logarithm, and adopt the standard asymptotic notation $O(\cdot)$, $\Omega(\cdot)$, $\Theta(\cdot)$, as well as the closely related $\tilde{O}(\cdot)$, $\tilde{\Omega}(\cdot)$, $\tilde{\Theta}(\cdot)$, which hide polylogarithmic factors in their argument.

2 An $\tilde{O}(\frac{n^{3/4}}{\varepsilon} + \frac{\log^2(n/\varepsilon)}{\varepsilon^2})$ Upper Bound for EET

In this section, we establish our main theorem, Theorem 1. We begin with a simple algorithmic “trick” that allows us to reduce the task to one where both p, q put at least some minimum probability $\eta > 0$ on every element of the domain:

► **Lemma 7.** *Let p, q be distributions over $[n]$, and set $\eta := \varepsilon/(4 \log(n/\varepsilon))$. Let $\tilde{p} := (1 - \eta)p + \eta U$ and likewise for $\tilde{q}$. Then $\tilde{p}_i, \tilde{q}_i \geq \varepsilon/(4n \log(n/\varepsilon))$ for every $i \in [n]$, and*

$$|H(\tilde{p}) - H(\tilde{q})| - |H(p) - H(q)| \leq \varepsilon/2.$$

Moreover, one sample from $\tilde{p}$ (resp. $\tilde{q}$) can be simulated from one sample from p (resp. q).

Proof. Since $d_{TV}(p, \tilde{p}) \leq \frac{\eta}{2} = \frac{\varepsilon}{8 \log(n/\varepsilon)}$, by invoking the bound on entropy difference with respect to TV distance in [21, Lemma 2.7], we get

$$|H(p) - H(\tilde{p})| \leq \frac{\varepsilon}{8 \log(n/\varepsilon)} \log \left(\frac{n \cdot 8 \log(n/\varepsilon)}{\varepsilon} \right) = \varepsilon \cdot \frac{\log(n/\varepsilon) + \log(8 \log(n/\varepsilon))}{8 \log(n/\varepsilon)} \leq \varepsilon/4.$$

Similarly, we have $|H(q) - H(\tilde{q})| \leq \varepsilon/4$. By the triangle inequality, we then have

$$|H(\tilde{p}) - H(\tilde{q})| \leq |H(p) - H(\tilde{p})| + |H(q) - H(\tilde{q})| + |H(p) - H(q)| \leq \varepsilon/2 + |H(p) - H(q)|,$$

and

$$|H(p) - H(q)| \leq |H(\tilde{p}) - H(\tilde{q})| + |H(p) - H(\tilde{p})| + |H(q) - H(\tilde{q})| \leq \varepsilon/2 + |H(\tilde{p}) - H(\tilde{q})|,$$

establishing the lemma. ◀

Of course, if $p = q$, then $\tilde{p} = \tilde{q}$; Lemma 7 guarantees that if $|H(p) - H(q)| \geq \varepsilon$, then $|H(\tilde{p}) - H(\tilde{q})| \geq \Omega(\varepsilon)$. That is, $\tilde{p}, \tilde{q}$ satisfy the promise with the same ε (up to constants) whenever p, q do, so we can henceforth assume that $\min_{i \in [n]} p_i, \min_{i \in [n]} q_i \geq \frac{\varepsilon}{4n \log(n/\varepsilon)}$. With this in hand, we are ready to establish our main theorem, restated, as follows:

► **Theorem 8 (Theorem 1, restated).** *There exists an algorithm (Algorithm 1) which, given sample access to two unknown probability distributions p, q over a known domain $[n]$, as well as parameter $\varepsilon \in (0, 1/2]$, takes*

$$O\left(\frac{n^{3/4} \log(n/\varepsilon) \cdot \log n}{\varepsilon} + \frac{\log^2(n/\varepsilon)}{\varepsilon^2}\right)$$

samples from both p and q , and has the following guarantees:

Algorithm 1 Entropy equivalence testing (EET).

Require: Sample access to p and q , both over $[n]$; accuracy parameter $\varepsilon \in (0, 1]$.

- 1: Run a Hellinger closeness tester on p, q with $O\left(\frac{n^{3/4}}{\varepsilon}\right)$ samples.
- 2: **if** $d_H^2(p, q) \geq \Omega(\varepsilon)$ **then return reject**
- 3: Set $m_1 := \tilde{O}(n^{3/4}/\varepsilon)$. Take m_1 samples each from p, q ; let X_i, Y_i be the empirical counts.
- 4: $S \leftarrow \{i \in [n] : X_i + Y_i \geq \Omega(m_1 \cdot \varepsilon / (n^{3/4} \log(n/\varepsilon)))\}$ $\triangleright$ heavy elements
- 5: Set $m_2 := \tilde{O}(1/\varepsilon)$. Using m_2 samples, test whether $p(\bar{S}) \geq \tilde{\Omega}(\varepsilon)$ and $q(\bar{S}) \geq \tilde{\Omega}(\varepsilon)$.
- 6: **if** only one of the preceding inequalities hold **then** $\triangleright p$ and q differ on $\bar{S}$.
- 7: **return reject**
- 8: **if** both preceding inequalities hold **then** $\triangleright$ low-mass regime on $\bar{S}$
- 9: Set $m_3 := O(\log^2(n/\varepsilon)/\varepsilon^2)$. Using m_3 samples, estimate $|p(\bar{S}) - q(\bar{S})|$.
- 10: **if** $|p(\bar{S}) - q(\bar{S})| \geq \tilde{\Omega}(\varepsilon)$ **then return reject**
- 11: Via rejection sampling on $\bar{S}$, run a TV closeness tester on $p_{\bar{S}}, q_{\bar{S}}$ at threshold $\tilde{\Omega}(\varepsilon/q(\bar{S}))$ using $\tilde{O}(n^{3/4}/\varepsilon)$ samples.
- 12: **if** the conditional tester rejects **then**
- 13: **return reject**
- 14: Set $m_4 := \tilde{O}(n^{3/4}/\varepsilon + \log^2(n/\varepsilon)/\varepsilon^2)$ and $s := \Omega(n^{3/4} \log(n/\varepsilon)/\varepsilon)$.
- 15: Using s samples, compute $T \leftarrow \sum_{i \in S} \frac{(X_i - Y_i)^2 - (X_i + Y_i)}{X_i + Y_i}$. $\triangleright$ bias check
- 16: **if** $T \geq \Omega(\sqrt{n})$ **then**
- 17: **return reject**
- 18: Using $O(\log^2(n/\varepsilon)/\varepsilon^2)$ samples, test both $|p(S) - q(S)| \geq \Omega(\frac{\varepsilon}{\log(n/\varepsilon)})$ and $\|p - q\|_2^2 \geq \Omega(\frac{\varepsilon^2}{\log^2(n/\varepsilon)})$.
- 19: **if** either preceding inequality holds **then**
- 20: **return reject**
- 21: Take m_4 samples; let X_i, Y_i be the empirical counts on S . Compute

$$Z \leftarrow \sum_{i \in S} \frac{X_i - Y_i}{m_4} \log \frac{1}{X_i + Y_i}.$$

- 22: **if** $|Z| \geq \Omega(\varepsilon)$ **then return reject**
 - 23: **else return accept**
-

- If $p = q$, then the algorithm outputs **accept** with probability at least $9/10$;
 - If $|H(p) - H(q)| \geq \varepsilon$, then the algorithm outputs **reject** with probability at least $9/10$.
- Further, the algorithm runs in time $O(N \log N)$, where N is the total number of samples.

The remainder of this section is dedicated to proving this theorem, which we do via a series of lemmas.

The first lemma relates a difference in entropy to a difference in (squared) Hellinger distance and a cross entropy term with respect to a mixture:

► **Lemma 9.** For any two distributions p, q over $[n]$, we have

$$|H(p) - H(q)| \leq \Theta(d_H^2(p, q)) + \left| \sum_{i=1}^n (p_i - q_i) \log \frac{1}{(p_i + q_i)/2} \right|.$$

50:10 Entropy Equivalence Testing

Proof. Letting $r := \frac{p+q}{2}$, we can write

$$\begin{aligned}
H(p) - H(q) &= \sum_i \left(p_i \log \frac{1}{p_i} - q_i \log \frac{1}{q_i} \right) \\
&= \sum_i \left(-p_i \log \frac{p_i}{(p_i + q_i)/2} + p_i \log \frac{1}{(p_i + q_i)/2} + q_i \log \frac{q_i}{(p_i + q_i)/2} - q_i \log \frac{1}{(p_i + q_i)/2} \right) \\
&= d_{\text{KL}}(q \| r) - d_{\text{KL}}(p \| r) + \sum_i \left(p_i \log \frac{1}{(p_i + q_i)/2} - q_i \log \frac{1}{(p_i + q_i)/2} \right) \\
&= d_{\text{KL}}(q \| r) - d_{\text{KL}}(p \| r) + \sum_i (p_i - q_i) \log \frac{1}{(p_i + q_i)/2}.
\end{aligned}$$

We then observe that

$$d_{\text{KL}}(q \| r) \leq d_{\chi^2}(q, r) = \sum_i \frac{\left(q_i - \frac{p_i + q_i}{2}\right)^2}{\frac{p_i + q_i}{2}} = \sum_i \frac{1}{2} \frac{(p_i - q_i)^2}{p_i + q_i} = \Theta(d_H^2(p, q)),$$

where the last equality comes from the equivalence of between Hellinger distance and Triangle distance (e.g., see [24, Proposition 3]).⁷ Similarly, we have that $d_{\text{KL}}(p \| r) \leq d_{\chi^2}(p, r) = \Theta(d_H^2(p, q))$, and we conclude that

$$\begin{aligned}
|H(p) - H(q)| &\leq |d_{\text{KL}}(q \| r)| + |d_{\text{KL}}(p \| r)| + \left| \sum_i (p_i - q_i) \log \frac{1}{(p_i + q_i)/2} \right| \\
&\leq \Theta(d_H^2(p, q)) + \left| \sum_i (p_i - q_i) \log \frac{1}{(p_i + q_i)/2} \right|,
\end{aligned}$$

as claimed. ◀

Lemma 9 allows us to split the entropy difference $|H(p) - H(q)|$ into a Hellinger part and a term $|\sum_i (p_i - q_i) \log(1/((p_i + q_i)/2))|$, so that an entropy difference ε implies at least one of these two terms must be $\varepsilon/2$: thus, it suffices to test both, and reject if, and only if, at least one of them is detected to be $\Omega(\varepsilon)$. (Clearly, if $p = q$, both the Hellinger and KL-like terms are zero.)

The Hellinger term can be tested with $\tilde{O}(n^{3/4}/\varepsilon)$ samples using a Hellinger closeness tester (e.g., [28, Proposition 2.15]): if $d_H^2(p, q) \geq \Omega(\varepsilon)$, this tester will allow us to reject with high probability. For the rest of the subsection we therefore can assume $d_H^2(p, q) \leq O(\varepsilon)$, and the task reduces to testing the second, “KL-like” term,

$$\Lambda(p, q) := \left| \sum_i (p_i - q_i) \log \frac{1}{(p_i + q_i)/2} \right|, \quad (1)$$

and specifically to distinguish $\Lambda(p, q) = 0$ from $\Lambda(p, q) \geq \frac{\varepsilon}{2}$.

To do so, the idea is to handle separately the set S^* of “heavy” elements, on which either p or q puts relatively high probability (specifically, at least some threshold $\tau := \tilde{\Omega}(\varepsilon/n^{3/4})$), and the remaining set $\bar{S}^* = [n] \setminus S^*$ of “light” elements (which have low probability under

⁷ Note that the induced KL terms have been studied in information theory (called Jensen-Shannon divergence; see e.g., [33, Chapter 7] and more systematically in [38], where they actually show that the Jensen-Shannon divergence and the Triangle distance are equivalent up to constant factors).

both p and q). Of course, we do not know this set S^* , as both p, q are unknown; however, by taking $\tilde{O}(1/\tau) = \tilde{O}(n^{3/4}/\varepsilon)$ samples (which is inconsequential, as we will pay this sample complexity for the Hellinger test anyway), we can obtain a set S which is a good proxy for S^* . To make this formal, define the following two sets:

$$S_1 := \left\{ i \in [n] \mid p_i + q_i \geq C_1 \cdot \frac{\varepsilon}{n^{3/4} \log(n/\varepsilon)} \right\},$$

$$S_2 := \left\{ i \in [n] \mid p_i + q_i \geq C_2 \cdot \frac{\varepsilon}{n^{3/4} \log(n/\varepsilon)} \right\},$$

where $C_2 \geq C_1 > 0$ are two absolute constants. Our focus is to identify a set $S \subseteq [n]$ such that $S_2 \subseteq S \subseteq S_1$.

► **Lemma 10.** *Using $O\left(\frac{n^{3/4} \log(n/\varepsilon) \log n}{\varepsilon}\right)$ samples from p and q , one can efficiently return a set S satisfying $S_2 \subseteq S \subseteq S_1$ with probability at least $\frac{99}{100}$.*

Proof. By Fact 6, choosing $2C_1 \leq C_2$, as we have sample access to the uniform mixture $\frac{p+q}{2}$, we can distinguish $\frac{p_i+q_i}{2} \geq C_2 \cdot \frac{\varepsilon}{2n^{3/4} \log(n/\varepsilon)}$ and $\frac{p_i+q_i}{2} \leq C_1 \cdot \frac{\varepsilon}{2n^{3/4} \log(n/\varepsilon)}$ with probability $1 - \frac{1}{100n}$ costing $O\left(\frac{n^{3/4} \log(n/\varepsilon) \log n}{\varepsilon}\right)$ samples from p and q . Therefore, running the test over the support $[n]$ (using the same set of samples to run this test for all i 's) and adding each $i \in [n]$ to the set S if we detect $p_i + q_i \geq C_2 \cdot \frac{\varepsilon}{n^{3/4} \log(n/\varepsilon)}$, by a union bound, the algorithm will be simultaneously correct on all $i \in [n]$ with probability at least $\frac{99}{100}$. ◀

To summarize, once we have identified such a set S , given that

$$\underbrace{\Lambda(p, q)}_{0 \text{ or } \geq \frac{\varepsilon}{2}} \leq \underbrace{\left| \sum_{i \in \bar{S}} (p_i - q_i) \cdot \log \frac{1}{p_i + q_i} \right|}_{=: \Lambda_{\bar{S}}(p, q)} + \underbrace{\left| \sum_{i \in S} (p_i - q_i) \cdot \log \frac{1}{p_i + q_i} \right|}_{=: \Lambda_S(p, q)}$$

it suffices to separately test (1) $\Lambda_{\bar{S}}(p, q) = 0$ vs. $\Lambda_{\bar{S}}(p, q) \geq \varepsilon/4$, and (2) $\Lambda_S(p, q) = 0$ vs. $\Lambda_S(p, q) \geq \varepsilon/4$; we do so below, assuming that our set S satisfies $S_2 \subseteq S$.⁸

Test 1: $\Lambda_{\bar{S}}(p, q) = 0$ vs. $\Lambda_{\bar{S}}(p, q) \geq \varepsilon$

Inspired by the ideas of [28], we can then check if the distributions are far in TV distance in the low-probability regime (i.e., elements in $\bar{S}$), which will suffice by a simple reduction from entropy difference to TV (using the lower bound $\min_i p_i, \min_i q_i \geq \tilde{\Omega}(\varepsilon/n)$). In more detail, we have

$$\left| \sum_{i \in \bar{S}} (p_i - q_i) \cdot \log \frac{1}{p_i + q_i} \right| \leq \sum_{i \in \bar{S}} |p_i - q_i| \cdot \log \frac{1}{p_i + q_i} \leq O(\log(n/\varepsilon)) \cdot \sum_{i \in \bar{S}} |p_i - q_i|.$$

While this is a task that would in general require a large number of samples, here we can leverage the additional fact that $\bar{S}$ does not contain any heavy elements.

► **Lemma 11.** *Let $T \subseteq [n]$ be a set such that $\max_{i \in T} (p_i + q_i) \leq C_2 \cdot \frac{\varepsilon}{n^{3/4} \log(n/\varepsilon)}$. Then, testing $p_T = q_T$ versus $\sum_{i \in T} |p_i - q_i| \geq \frac{\varepsilon}{\log(n/\varepsilon)}$ (with probability of success 9/10) can be done with $O(n^{3/4} \cdot \log(n/\varepsilon)/\varepsilon)$ samples from p, q .*

⁸ For simplicity of exposition, we henceforth ignore the factor 4, and consider ε instead of $\varepsilon/4$; a simple rescaling argument gives the original guarantees.

Proof. By [28], we know that $\tilde{O}\left(\frac{n^{3/4}}{\varepsilon}\right)$ samples suffice to distinguish $p_T = q_T$ from $d_{TV}(p_T, q_T) \geq \varepsilon$ with high constant probability. For the sake of completeness, we briefly explain how to adapt their argument to our setting. First, it takes $O\left(\frac{\log(n/\varepsilon)}{\varepsilon}\right)$ samples to check whether the mass of p on T is larger than $\Omega(\varepsilon/\log(n/\varepsilon))$, and similarly for q : if both have mass smaller than this, we accept, as $\sum_{i \in \bar{S}} |p_i - q_i| \leq p(T) + q(T) < \frac{\varepsilon}{\log(n/\varepsilon)}$. If only one of the two puts large enough mass on T , we reject, as this shows that $\sum_{i \in T} |p_i - q_i| \neq 0$. Otherwise, both p, q put enough mass on T , and we continue.

With $O\left(\frac{\log^2(n/\varepsilon)}{\varepsilon^2}\right)$ samples, we then confirm that their probability masses on T approximately match, i.e., that $|p(T) - q(T)| \leq O\left(\frac{\varepsilon}{\log(n/\varepsilon)}\right)$. If they do not, we can output reject. Otherwise, we have

$$\begin{aligned} \sum_{i \in T} |p_i - q_i| &= \sum_{i \in T} |p(i|i \in T)p(T) - q(i|i \in T)q(T)| \\ &\leq \sum_{i \in T} (|p(i|i \in T)p(T) - p(i|i \in T)q(T)| + |p(i|i \in T)q(T) - q(i|i \in T)q(T)|) \\ &= |p(T) - q(T)| + q(T) \cdot \|p(\cdot|T) - q(\cdot|T)\|_1. \end{aligned}$$

Thus, if $\|p_T - q_T\| \geq \Omega\left(\frac{\varepsilon}{\log(n/\varepsilon)}\right)$ and $|p(T) - q(T)| \leq O\left(\frac{\varepsilon}{\log(n/\varepsilon)}\right)$ (for suitable choice of constants in the $O(\cdot)$, $\Omega(\cdot)$), then we have that

$$d_{TV}(p_T, q_T) = \frac{1}{2} \|p(\cdot|T) - q(\cdot|T)\|_1 \geq \Omega\left(\frac{\varepsilon}{q(T) \cdot \log(n/\varepsilon)}\right).$$

Recalling that $q(T), p(T) \geq \frac{\varepsilon}{\log(n/\varepsilon)}$ and $q(T) = \Theta(p(T))$, we can sample from p_T, q_T via rejection sampling at the cost of an expected $O(1/q(T))$, $O(1/p(T))$ factor in the sample complexity. Thus, we can use an off-the-shelf TV equivalence tester (e.g., from [17, Theorem 1]) to test $d_{TV}(p(\cdot|T), q(\cdot|T)) \geq \Omega\left(\frac{\varepsilon}{q(T) \cdot \log(n/\varepsilon)}\right)$ vs. $p(\cdot|T) = q(\cdot|T)$ at the cost of

$$\max \left\{ \frac{n^{2/3} q^{4/3}(T) \log^{4/3}(n/\varepsilon)}{\varepsilon^{4/3}}, \frac{n^{1/2} q^2(T) \log^2(n/\varepsilon)}{\varepsilon^2} \right\}$$

samples from $p(\cdot|T), q(\cdot|T)$; which, by the above discussion, requires an expected

$$\max \left\{ \frac{n^{2/3} q^{1/3}(T) \log^{4/3}(n/\varepsilon)}{\varepsilon^{4/3}}, \frac{n^{1/2} q(T) \log^2(n/\varepsilon)}{\varepsilon^2} \right\}$$

samples from p, q (which can be made high-probability via Markov's inequality, paying a constant factor in the sample complexity and a small additional probability of failure). Since $q(T) \leq \frac{\varepsilon}{n^{3/4} \log(n/\varepsilon)} \cdot n = \frac{\varepsilon \cdot n^{1/4}}{\log(n/\varepsilon)}$, we finally conclude that taking $O\left(\frac{n^{3/4} \cdot \log(n/\varepsilon)}{\varepsilon}\right)$ samples suffices. $\blacktriangleleft$

Test 2: $\Lambda_S(p, q) = 0$ vs. $\Lambda_S(p, q) \geq \varepsilon$

We now look at the second – and more involved – case, namely, detecting when the entropy difference is far in the high mass regime. That is, we are testing $\Lambda_S(p, q) = 0$ vs. $\Lambda_S(p, q) \geq \varepsilon$.

We will define a suitable test statistic, Z , and analyze its expectation (with the goal of bounding its bias and relate it to $\Lambda_S(p, q)$), before bounding its variance to conclude with an application of Chebyshev's inequality. At a high level, our strategy is as follows: either the

empirical entropy difference Z has *low bias*, which means we can simply rely on its output for the test, or it has *high bias*, in which case this can be detected by performing additional checks to detect this bias.

Let $m = \tilde{O}\left(\frac{n^{3/4}}{\varepsilon} + \frac{\log(n/\varepsilon)}{\varepsilon^2}\right)$ be the number of samples taken from p, q , and assume the Poissonized setting: that is, instead of exactly m samples, we take a random $\text{Poi}(m)$ number of samples.⁹ Recall that by construction of the set $S \subseteq S_1$, we can assume $p_i + q_i \geq \tilde{\Omega}\left(\frac{\varepsilon}{n^{3/4}}\right)$ for all $i \in S$. Focusing on S , we run two further checks, each at cost $O(\log^2 m / \varepsilon^2)$:

1. $|p(S) - q(S)| \leq O(\varepsilon / \log m)$, which implies $|\sum_{i \in S} (p_i - q_i)| \log m \leq O(\varepsilon)$.
2. $\|p - q\|_2^2 \leq O(\varepsilon^2 / \log^2 m)$, which is checked via testing at distance $\varepsilon / \log m$ using an ℓ_2 -closeness tester (e.g., via learning p, q in ℓ_2 distance, see for instance [16]).

If either check rejects, we reject; otherwise we proceed assuming both guarantees hold.

Letting X_i (resp. Y_i) be the number of occurrences of element $i \in S$ among the samples from p (resp. from q), we define our entropy difference estimator to be

$$Z := \sum_{i \in S} \frac{X_i - Y_i}{m} \log \frac{1}{X_i + Y_i},$$

and we write $Z_i := \frac{X_i - Y_i}{m} \log \frac{1}{X_i + Y_i}$ for the i -th summand and let $Z_i = 0$ when $X_i + Y_i = 0$. By the above Poissonization procedure, we have $X_i \sim \text{Poi}(m \cdot p_i)$ and $Y_i \sim \text{Poi}(m \cdot q_i)$ for all i , and the $(X_i)_i$ (resp. $(Y_i)_i$) are independent. We first bound the bias of this estimator Z :

► **Lemma 12.** *For any set $S \subseteq [n]$,*

$$\left| \sum_{i \in S} (p_i - q_i) \log \frac{1}{m(p_i + q_i)} - \mathbb{E}[Z] \right| \leq \sum_{i \in S} \frac{|p_i - q_i|}{m(p_i + q_i)}.$$

Proof. Let $J_i := X_i + Y_i \sim \text{Poi}(\lambda_i)$ with $\lambda_i := m(p_i + q_i)$. Conditional on $J_i = j$, the count X_i is $\text{Bin}(j, p_i/(p_i + q_i))$ -distributed, so $\mathbb{E}[X_i - Y_i \mid J_i] = J_i \cdot (p_i - q_i)/(p_i + q_i)$. By the law of total expectation, pulling the factor $\log(1/J_i)$ outside the inner expectation,

$$\mathbb{E}[Z_i] = \frac{1}{m} \mathbb{E} \left[(X_i - Y_i) \log \frac{1}{J_i} \right] = \frac{1}{m} \mathbb{E} \left[\log \frac{1}{J_i} \cdot \mathbb{E}[X_i - Y_i \mid J_i] \right] = \frac{p_i - q_i}{m(p_i + q_i)} \mathbb{E}[J_i \log(1/J_i)].$$

We then apply the Poisson identity (Fact 5) $\mathbb{E}[J_i f(J_i)] = \lambda_i \mathbb{E}[f(J_i + 1)]$, taking $f(j) = \log(1/j)$:

$$\mathbb{E}[Z_i] = \frac{p_i - q_i}{m(p_i + q_i)} \cdot \lambda_i \mathbb{E}[\log(1/(J_i + 1))] = -(p_i - q_i) \mathbb{E}[\log(J_i + 1)].$$

Subtracting the target $(p_i - q_i) \log(1/\lambda_i)$,

$$(p_i - q_i) \log \frac{1}{\lambda_i} - \mathbb{E}[Z_i] = (p_i - q_i) \mathbb{E} \left[\log \frac{J_i + 1}{\lambda_i} \right].$$

We proceed to bound $\mathbb{E}[\log((J_i + 1)/\lambda_i)]$ from above and below. By Jensen's inequality applied to the concave function $\log$,

$$\mathbb{E} \left[\log \frac{J_i + 1}{\lambda_i} \right] \leq \log \left(1 + \frac{1}{\lambda_i} \right) \leq \frac{1}{\lambda_i}.$$

⁹ This standard technique is valid as it only affects the sample complexity by constant terms and an arbitrarily small error probability adjustment, since a Poisson random variable tightly concentrates around its mean. Its advantage is to provide additional independence between the empirical counts of each element, which is useful in the analysis.

50:14 Entropy Equivalence Testing

For the lower bound, the inequality $\log x \geq 1 - 1/x$ applied to $x = (J_i + 1)/\lambda_i$, together with $\mathbb{E}[1/(J_i + 1)] = (1 - e^{-\lambda_i})/\lambda_i$ for $J_i \sim \text{Poi}(\lambda_i)$, gives

$$\mathbb{E}\left[\log \frac{J_i + 1}{\lambda_i}\right] \geq 1 - \lambda_i \mathbb{E}\left[\frac{1}{J_i + 1}\right] = e^{-\lambda_i} \geq 0.$$

Combining these findings gives $|\mathbb{E}[\log((J_i + 1)/\lambda_i)]| \leq 1/\lambda_i$ (using $e^{-\lambda_i} \leq 1/\lambda_i$ for all $\lambda_i > 0$), so

$$|(p_i - q_i) \log \frac{1}{\lambda_i} - \mathbb{E}[Z_i]| \leq \frac{|p_i - q_i|}{\lambda_i}.$$

Summing over $i \in S$ yields the claim. $\blacktriangleleft$

The bias of our estimator is determined by the RHS of Lemma 12; note that if $p = q$, the RHS simplifies to 0. As a result, a natural idea would be to use this RHS as a proxy, and to check if the RHS is noticeably large using another test: if it is larger than $\geq \Omega(\varepsilon)$, we can immediately reject, as this is clear evidence that $p \neq q$. We do so in the following lemma:

► **Lemma 13.** *Suppose that $d_H^2(p, q) \leq O(\varepsilon)$. Then, with $s = O(n^{3/4} \log(n/\varepsilon)/\varepsilon)$ samples from p, q , one can distinguish between the following two cases with high constant probability:*

1. $p = q$ (which implies $\sum_{i \in S} \frac{|p_i - q_i|}{m(p_i + q_i)} = 0$).
2. $\sum_{i \in S} \frac{|p_i - q_i|}{(p_i + q_i)} \geq \Omega(m\varepsilon)$.

Proof. We use the test statistic

$$T := \sum_{i \in S} \frac{(X_i - Y_i)^2 - (X_i + Y_i)}{X_i + Y_i},$$

where $X_i \sim \text{Poi}(s \cdot p_i)$ and $Y_i \sim \text{Poi}(s \cdot q_i)$. We show:

- In the first case, $\mathbb{E}[T] = 0$, while in the second case $\mathbb{E}[T] \geq \Omega(\sqrt{n})$.
- $\text{Var}[T] \leq O(n)$ in both cases.

Chebyshev's inequality then will enable us to separate the two cases by thresholding T at a value scaling as $\Theta(\sqrt{n})$.

Analyzing $\mathbb{E}[T]$. The first part is similar to the analysis from [17, Lemma 4], which relies on the same statistic T (for a different purpose): write $\delta_i := (p_i - q_i)/(p_i + q_i)$ and $J_i := X_i + Y_i$. Conditional on $J_i = j$, we have $Y_i = j - X_i$ so $X_i - Y_i = 2X_i - j$, and $X_i \sim \text{Bin}(j, p_i/(p_i + q_i))$. Hence,

$$\mathbb{E}[X_i - Y_i \mid J_i = j] = j\delta_i,$$

and the binomial variance $\text{Var}[X_i \mid J_i = j] = j \cdot \frac{p_i q_i}{(p_i + q_i)^2}$ gives

$$\text{Var}[X_i - Y_i \mid J_i = j] = 4j \cdot \frac{p_i q_i}{(p_i + q_i)^2} = j(1 - \delta_i^2).$$

Combining, $\mathbb{E}[(X_i - Y_i)^2 \mid J_i = j] = \text{Var}[X_i - Y_i \mid J_i = j] + \mathbb{E}[X_i - Y_i \mid J_i = j]^2 = j(1 - \delta_i^2) + j^2\delta_i^2$. The i -th summand of T is $\frac{(X_i - Y_i)^2 - J_i}{J_i}$, which we set to 0 when $J_i = 0$ (the 0/0 case). For $J_i = j \geq 1$,

$$\mathbb{E}\left[\frac{(X_i - Y_i)^2 - J_i}{J_i} \mid J_i = j\right] = \frac{(j^2 - j)\delta_i^2}{j} = (j - 1)\delta_i^2.$$

Taking the outer expectation, with the indicator $\mathbf{1}\{J_i \geq 1\}$ enforcing the 0/0 convention,

$$\mathbb{E}[T] = \sum_{i \in S} \delta_i^2 \cdot \mathbb{E}[(J_i - 1)\mathbf{1}\{J_i \geq 1\}] = \sum_{i \in S} \delta_i^2 (\lambda_i - 1 + e^{-\lambda_i}),$$

where $\mathbb{E}[(J_i - 1)\mathbf{1}\{J_i \geq 1\}] = \mathbb{E}[J_i] - \Pr[J_i \geq 1] = \lambda_i - (1 - e^{-\lambda_i})$.

In particular, in the first case (when $p = q$), $\delta_i = 0$ for every $i \in S$, and so $\mathbb{E}[T] = 0$. We therefore now focus on the second case, for which we depart from [17]. By the construction of S (Lemma 10), $p_i + q_i \geq \Omega(\frac{\varepsilon}{n^{3/4} \log(n/\varepsilon)})$ for every $i \in S$, so choosing $s = \Omega(n^{3/4} \cdot \log(n/\varepsilon)/\varepsilon)$ with a sufficiently large constant gives $\lambda_i = s(p_i + q_i) \geq 2$ for all $i \in S$. The bound $\lambda - 1 + e^{-\lambda} \geq \lambda - 1 \geq \lambda/2$ for $\lambda \geq 2$ then yields

$$\mathbb{E}[T] \geq \frac{1}{2} \sum_{i \in S} \lambda_i \delta_i^2 = \frac{s}{2} \sum_{i \in S} \frac{(p_i - q_i)^2}{p_i + q_i}.$$

By Cauchy–Schwarz,

$$\sum_{i \in S} \frac{(p_i - q_i)^2}{p_i + q_i} \geq \frac{1}{n} \left(\sum_{i \in S} \frac{|p_i - q_i|}{\sqrt{p_i + q_i}} \right)^2.$$

Since $m(p_i + q_i) \geq 1$ (in view of $i \in S$ and $m = \tilde{\Omega}(\frac{n^{3/4}}{\varepsilon})$), we have $\frac{1}{\sqrt{p_i + q_i}} \geq \frac{1}{\sqrt{m(p_i + q_i)}}$. Applying this termwise and using the definition of the second case,

$$\sum_{i \in S} \frac{|p_i - q_i|}{\sqrt{p_i + q_i}} \geq \frac{1}{\sqrt{m}} \sum_{i \in S} \frac{|p_i - q_i|}{p_i + q_i} \geq \Omega(\sqrt{m\varepsilon}).$$

Combining,

$$\mathbb{E}[T] \geq \frac{s}{2n} \cdot \Omega(m\varepsilon^2) = \Omega\left(\frac{sm\varepsilon^2}{n}\right),$$

which is $\Omega(\sqrt{n} + \frac{\log^2(n/\varepsilon)}{n^{1/4}\varepsilon}) \geq \Omega(\sqrt{n})$ given the settings of $s = \tilde{\Theta}(n^{3/4}/\varepsilon)$ and $m = \tilde{\Theta}(n^{3/4}/\varepsilon + \frac{\log^2(n/\varepsilon)}{\varepsilon^2})$.

Upper bound on $\text{Var}[T]$. The variance bound from [17, Lemma 5] gives

$$\text{Var}[T] \leq 2 \min\{n, s\} + 5s \sum_{i \in S} \frac{(p_i - q_i)^2}{p_i + q_i}.$$

Under the lemma's assumption on the Hellinger distance, $\sum_{i \in S} (p_i - q_i)^2 / (p_i + q_i) = \Theta(d_H^2(p, q)) \leq O(\varepsilon)$, so the second term is $\tilde{O}(s \cdot \varepsilon)$, and $\text{Var}[T] \leq O(n + s \cdot \varepsilon)$, where $s = \tilde{O}(n^{3/4}/\varepsilon)$. This scaling of s leads to the simplified bound $\text{Var}[T] \leq O(n)$.

Concluding via Chebyshev. In the first case, $\mathbb{E}[T] = 0$ and $\text{Var}[T] \leq O(n)$, so $|T| \leq O(\sqrt{n})$ with probability $\geq 9/10$. In the second case, $\mathbb{E}[T] \geq C\sqrt{n}$ and $\text{Var}[T] \leq O(n) \leq \mathbb{E}[T]^2/C'$, so $T \geq C''\sqrt{n}$ with probability $\geq 9/10$. Thresholding at $\Theta(\sqrt{n})$ thus distinguishes the two cases. $\blacktriangleleft$

Having used the RHS of Lemma 12 as a proxy for the bias, and checked that it is small (and rejected otherwise), we are not yet done: we could still have both small bias for Z (so that $\mathbb{E}[Z] \approx \sum_{i \in S} (p_i - q_i) \log \frac{1}{m(p_i + q_i)}$) and large $\Lambda_S(p, q) = \left| \sum_{i \in S} (p_i - q_i) \log \frac{1}{m(p_i + q_i)} \right|$. However, now, since we know that $|\mathbb{E}[Z]| \approx \Lambda_S(p, q)$, to conclude it suffices to run one last check: namely, that Z is small enough.

50:16 Entropy Equivalence Testing

Indeed, we know now that Z has small bias: if we can also show, via a variance argument, that Z concentrates around its expectation, then we will have $|Z| \approx_{\text{w.h.p.}} |\mathbb{E}[Z]| \approx \Lambda_S(p, q)$, and checking that $|Z|$ is small will ensure (with high probability) that we can guarantee that $\Lambda_S(p, q)$ is small. Thus, our last missing piece is an upper bound on $\text{Var}[Z]$, which will let us show that Z concentrates around its expectation.

► **Lemma 14.** *The variance of the entropy difference estimator Z satisfies*

$$\text{Var}[Z] \leq O(\log^2 m) \|p - q\|_2^2 + \frac{\log^2 m}{m} \leq O(\varepsilon^2).$$

Proof. Under Poissonization, the counts $\{X_i\}_{i \in S}$ and $\{Y_i\}_{i \in S}$ are mutually independent across coordinates. As a result, the Z_i 's are also independent, and

$$\text{Var}[Z] = \sum_{i \in S} \text{Var}[Z_i] \leq \sum_{i \in S} \mathbb{E}[Z_i^2].$$

Fix any $i \in S$, letting $J_i := X_i + Y_i \sim \text{Poi}(m(p_i + q_i))$, and defining the event $\mathcal{E}_i := \{J_i \leq 3m\}$. Since $m(p_i + q_i) \leq m$, the Chernoff bound for Poisson random variables gives $\Pr[\mathcal{E}_i^c] \leq e^{-\Omega(m)}$. Then:

- On $\mathcal{E}_i$: $|\log(1/J_i)| \leq \log(3m)$ for $J_i \neq 0$, and noting from before that $Z_i = 0$ when $J_i = 0$, we have $|Z_i| \mathbf{1}_{\mathcal{E}_i} \leq \frac{\log(3m)}{m} |X_i - Y_i|$. Squaring and taking expectations,

$$\mathbb{E}[Z_i^2 \mathbf{1}_{\mathcal{E}_i}] \leq \frac{\log^2(3m)}{m^2} \mathbb{E}[(X_i - Y_i)^2] = \frac{\log^2(3m)}{m^2} (m(p_i + q_i) + m^2(p_i - q_i)^2),$$

using $\text{Var}[X_i - Y_i] = m(p_i + q_i)$ and $\mathbb{E}[X_i - Y_i] = m(p_i - q_i)$ (moments of Poisson r.v.'s).

- On $\mathcal{E}_i^c$: from $|X_i - Y_i| \leq J_i$ and $|\log(1/J_i)| = \log J_i \leq J_i$ (valid since $J_i > 3m \geq 1$), we get $|Z_i| \mathbf{1}_{\mathcal{E}_i^c} \leq J_i^2/m$. Combining Cauchy–Schwarz with the moment bound $\mathbb{E}[J_i^8] = O(m^8)$ (since Poisson moments are polynomials in the mean, by Fact 5) and the tail bound $\Pr[\mathcal{E}_i^c] \leq e^{-\Omega(m)}$,

$$\mathbb{E}[Z_i^2 \mathbf{1}_{\mathcal{E}_i^c}] \leq \frac{1}{m^2} \mathbb{E}[J_i^4 \mathbf{1}_{\mathcal{E}_i^c}] \leq \frac{1}{m^2} \sqrt{\mathbb{E}[J_i^8] \Pr[\mathcal{E}_i^c]} \leq e^{-\Omega(m)}.$$

Summing over $i \in S$, using $\log^2(3m) \leq 4 \log^2 m$ (for $m \geq 3$) and $\sum_i (p_i + q_i) \leq 2$,

$$\text{Var}[Z] \leq O(\log^2 m) \|p - q\|_2^2 + \frac{O(\log^2 m)}{m} + n \cdot e^{-\Omega(m)}.$$

The last term is absorbed into $O(\frac{\log^2 m}{m})$ since $m \geq \log n$ makes $n \cdot e^{-\Omega(m)} \leq 1/m$. ◀

Combining Lemma 12 and Lemma 13, and conditioning on all tests passing, the expectation of Z has small bias:

$$\left| \mathbb{E}[Z] - \sum_i (p_i - q_i) \log \frac{1}{(p_i + q_i)} \right| \leq O(\varepsilon).$$

By Lemma 14, $\text{Var}[Z] \leq O(\varepsilon^2)$, so Chebyshev gives $|Z - \mathbb{E}[Z]| \leq O(\varepsilon)$ with probability at least 9/10. Combining via the triangle inequality,

$$\Pr \left[\left| Z - \sum_i (p_i - q_i) \log \frac{1}{(p_i + q_i)} \right| \leq O(\varepsilon) \right] \geq \frac{9}{10}.$$

To conclude this subcase. We choose a threshold for $|Z|$ at a suitable constant multiple of ε that distinguishes the two cases. If $\Lambda(p, q) = \left| \sum_{i \in S} (p_i - q_i) \log \frac{1}{(p_i + q_i)} \right| = 0$, then with probability at least $9/10$ the test on T from Lemma 13 will not reject and $|Z| \leq O(\varepsilon)$, and so this overall test accepts. If $d_H^2(p, q) \leq O(\varepsilon)$ and $\left| \sum_{i \in S} (p_i - q_i) \log \frac{1}{(p_i + q_i)} \right| \geq \Omega(\varepsilon)$, then with probability at least $9/10$ either the test on T from Lemma 13 rejects, or it does not and then $|Z| \geq \Omega(\varepsilon)$, and this overall test rejects.

Wrapping up

Putting the above ingredients together:

- If $p = q$, then with high constant probability, (1) the Hellinger test on $d_H(p, q)$ accepts, and (2) (a) the set S is correctly identified, (b) the test on $\Lambda_{\bar{S}}(p, q)$ accepts, (c) (i) the test on T (checking for low bias for Z) accepts, and (ii) the test on Z (test on $\Lambda_S(p, q)$ assuming low bias for Z) accepts. Overall, the algorithm accepts with probability at least $9/10$ by a union bound.
- If $|H(p) - H(q)| \geq \varepsilon$, then with high constant probability (1) either the Hellinger test on $d_H(p, q)$ rejects; or it accepts because $d_H(p, q)$ is small, in which case $\Lambda(p, q)$ must be large; then, with high constant probability (2) (a) the set S is correctly identified, (b) the test on $\Lambda_{\bar{S}}(p, q)$ rejects, or it accepts because $\Lambda_{\bar{S}}(p, q)$ is small, which means $\Lambda_S(p, q)$ must be large; in which case (c) (i) the test on T (checking for low bias for Z assuming $d_H(p, q)$ is small) rejects, or it accepts because the bias is small, in which case (ii) the test on Z (test on $\Lambda_S(p, q)$ assuming low bias for Z) rejects. Overall, the algorithm rejects with probability at least $9/10$ by a union bound.

Overall, each subtest in Algorithm 1 uses no more than $O\left(\frac{n^{3/4} \log(n/\varepsilon) \log n}{\varepsilon}\right)$ or $O\left(\frac{\log^2(n/\varepsilon)}{\varepsilon^2}\right)$ samples, and hence the total sample complexity of our algorithm is

$$O\left(\frac{n^{3/4} \log(n/\varepsilon) \log n}{\varepsilon} + \frac{\log^2(n/\varepsilon)}{\varepsilon^2}\right).$$

All our test statistics can be computed in $O(n)$ time once all X_i, Y_i are collected.¹⁰ To compute X_i, Y_i , one can first sort over all samples which takes $O(s \log s)$ time, where s is the number of samples in one of the subtests, and then sum them up. Thus, the time complexity of our algorithm is $O(N \log N)$, where N is the total number of samples. This concludes the proof of Theorem 8.

3 Bayesian Network Equivalence Testing

In this section, we show how to leverage our Entropy Equivalence Testing (EET) algorithm (Algorithm 2) to obtain an *equivalence testing* (closeness testing) algorithm for bounded-degree Bayesian networks (Bayes nets). Specifically, the problem we consider is the following:

¹⁰The statistics for testing Hellinger square is the same as T (albeit with different threshold; see [24, Theorem 5]). T can be computed in $O(n)$ time once all X_i, Y_i are known. Computing $S \leftarrow \{i \in [n] : X_i + Y_i \geq \Omega(m_1 \cdot \varepsilon / (n^{3/4} \log(n/\varepsilon)))\}$ can be done in $O(n)$ time and calculating the empirical of S can be implemented in $O(m_2)$ time. As for the conditional TV tester, the rejection sampling will cost as many samples as we take in total, and for TV closeness testing we use the same statistic T (see, e.g., [24, Theorem 4]). For ℓ_2 tester, the statistic is also computable in $O(n)$ time [17, Section 3]. Assuming that we can compute $\log(x)$ in constant time, one can compute the statistic Z in $O(n)$ time.

Bayes-Net Equivalence Testing: Given sample access to distributions p, q over $\{0, 1\}^n$, each Markov with respect to a DAG of in-degree at most d (the DAG is unknown, and possibly distinct between p and q), and parameters $\varepsilon \in (0, 1/2]$ and $\delta \in (0, 1]$, distinguish, with probability at least $1 - \delta$, between the following two cases: (1) $p = q$, and (2) $d_{\text{TV}}(p, q) \geq \varepsilon$.

► **Definition 15.** A projection of a Bayes net p on $\{0, 1\}^n$ onto a DAG G is denoted by p_G , and is defined by its probability mass function (PMF) as follows:

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

where Π_i^G is the set of parents of X_i in G . Moreover, with a slight abuse of notation, we write p_{X_i, Π_i} or $p_{X_i, \Pi_i}(x_i, \pi_i)$ for the marginal distribution of p on the subset $\{X_i, \Pi_i\}$.

The baseline here is the testing-by-learning approach, whereby one learns the two unknown in-degree d Bayes nets to TV distance $\frac{\varepsilon}{3}$ and checks the distance between the two learned hypotheses. It is known that learning an in-degree d Bayes net in total variation distance can be done in $\tilde{O}(2^d n / \varepsilon^2)$ samples (see [12, Appendix A]): to the best of our knowledge, the time complexity of any learning algorithm achieving this bound is exponential, namely, $2^{\tilde{\Omega}(n)}$. In contrast, we obtain a testing algorithm with drastically better running time (for all $d = o(n)$), and significantly better sample complexity whenever $d \gg \log(n/\varepsilon^2)$.

► **Theorem 16** (Theorem 4, restated). *There exists an algorithm (Algorithm 2) that, given sample access to two Bayesian networks p and q over $\{0, 1\}^n$ with in-degree at most d , as well as a parameter $\varepsilon \in (0, 1]$, takes*

$$s = \tilde{O} \left(\min \left\{ \frac{2^{3d/4} n}{\varepsilon^2}, \frac{2^{2d/3} n^{4/3}}{\varepsilon^{8/3}} \right\} + \frac{n^2}{\varepsilon^4} \right),$$

samples from p, q , and distinguishes with probability at least $2/3$ between (1) $p = q$ and (2) $d_{\text{TV}}(p, q) \geq \varepsilon$. Moreover, the algorithm runs in time $\tilde{O}(n^{d+1} s)$.

Proof. The pseudocode is provided in Algorithm 2. Due to space constraints, we defer the proof of correctness to the full version [14]. ◀

► **Remark 17.** We have an interesting dichotomy: our algorithm is efficient (polynomial in n) only if d is constant. However, it is worse than its learning counterpart when d is constant: n^2 (our test) vs. n (learning). On the other hand, if $d = \Omega(\log n)$, then our tester will be more efficient than learning sample-wise, but its runtime becomes $n^{\Omega(\log n)}$, and is thus quasi-polynomial.

As identity testing is a special case of closeness testing, our above analysis can straightforwardly be ported to yield an identity testing algorithm (using an algorithm for identity testing in Hellinger instead of a closeness one as the first building block). An immediate consequence of this is to yield an improved identity tester for Bayes net with sample complexity $\tilde{O}(\frac{2^{d/2} n}{\varepsilon^2} + \frac{n^2}{\varepsilon^4})$, thus improving on the prior bound $\tilde{O}(\frac{2^{d/2} n^{3/2}}{\varepsilon^2} + \frac{n^2}{\varepsilon^4})$ [15, Theorem 1.2].

■ **Algorithm 2** Closeness testing for Bayesian networks.

Require: Sample access to p and q , both in-degree d Bayesian networks over $\{0, 1\}^n$; accuracy parameter $\varepsilon \in (0, 1]$.

- 1: Set $\varepsilon_1 \leftarrow \Theta\left(\frac{\varepsilon^2}{n}\right)$, $\varepsilon_2 \leftarrow \Theta\left(\frac{\varepsilon^2}{d \cdot n \cdot \log(dn/\varepsilon)}\right)$, $\delta \leftarrow \frac{1}{20\binom{n}{d+1}}$
- 2: Set

$$m \leftarrow \tilde{O}\left(\min\left\{\frac{2^{3d/4} \cdot n}{\varepsilon^2}, \frac{2^{2d/3} \cdot n^{4/3}}{\varepsilon^{8/3}}\right\}\right) + O\left(\frac{d^3 n^2 \cdot \log^2(n/\varepsilon)}{\varepsilon^4}\right)$$

- 3: Take a multiset set S_p (resp. S_q) of m samples from $\tilde{p}$ (resp. $\tilde{q}$), defined by

$$\tilde{p} := \left(1 - \frac{C_3 \cdot \varepsilon^2}{dn \log(n/\varepsilon)}\right) p + \frac{C_3 \cdot \varepsilon^2}{dn \log(n/\varepsilon)} U \quad \text{and} \quad \tilde{q} := \left(1 - \frac{C_3 \cdot \varepsilon^2}{dn \log(n/\varepsilon)}\right) q + \frac{C_3 \cdot \varepsilon^2}{dn \log(n/\varepsilon)} U$$

- 4: **for all** subset $T \subseteq [n]$ with $|T| = d + 1$ **do** $\triangleright \binom{n}{d+1}$ possible subsets
- 5: Let $\mathcal{T} \leftarrow \{0, 1\}^T$
- 6: Test $\tilde{p}_{\mathcal{T}} = \tilde{q}_{\mathcal{T}}$ vs. $|H(\tilde{p}_{\mathcal{T}}) - H(\tilde{q}_{\mathcal{T}})| \geq \varepsilon_1$ with failure probability δ , using S_p, S_q
- 7: $\triangleright$ By Corollary 2
- 8: Test $\tilde{p}_{\mathcal{T}} = \tilde{q}_{\mathcal{T}}$ vs. $d_H^2(\tilde{p}_{\mathcal{T}}, \tilde{q}_{\mathcal{T}}) \geq \varepsilon_2$ with failure probability δ , using S_p, S_q $\triangleright$ By [28]
- 9: **if** either test rejects **then return reject**
- 10: **return accept**

► **Corollary 18.** *Given sample access to a Bayesian networks p and a full description of Bayesian network q over $\{0, 1\}^n$ with in-degree at most d , as well as a parameter $\varepsilon \in (0, 1]$, there is an algorithm that takes*

$$s = \tilde{O}\left(\frac{2^{d/2} n}{\varepsilon^2} + \frac{n^2}{\varepsilon^4}\right),$$

samples from p , and distinguishes with probability at least $2/3$ between (1) $p = q$ and (2) $d_{TV}(p, q) \geq \varepsilon$. Moreover, the algorithm runs in time $\tilde{O}(n^{d+1}s)$.

4 Lower Bound

In this section, we establish our sample complexity lower bound for EET (Theorem 3). The $\tilde{\Omega}(\min\{n^{3/4}/\varepsilon, n^{2/3}/\varepsilon^{4/3}\})$ term follows from a reduction to mutual information (MI) testing combined with [37, Corollary 6.2], and the $\Omega(\log^2 n/\varepsilon^2)$ term follows directly from [15, Lemma 2.7], whose two-point construction is also a hard instance for EET.

For a joint distribution P_{AC} over $[k_A] \times [k_C]$ with marginals P_A, P_C , the *mutual information* between coordinates is

$$I(A : C) := H(A) + H(C) - H(A, C),$$

and equals zero if and only if $P_{AC} = P_A \otimes P_C$. The MI testing problem asks, given sample access to P_{AC} , to distinguish $I(A : C) = 0$ from $I(A : C) \geq \varepsilon$ with probability at least $2/3$. [37, Corollary 6.2] established the lower bound

$$\tilde{\Omega}\left(\min\left\{\frac{k_A^{3/4} k_C^{1/4}}{\varepsilon}, \frac{k_A^{2/3} k_C^{1/3}}{\varepsilon^{4/3}}\right\}\right)$$

for this task.

Reduction from MI testing to EET

► **Lemma 19.** *Let $k_A, k_C \geq 2$ and $n = k_A k_C$. Any EET algorithm $\mathcal{A}$ over $[n]$ with sample complexity T yields an MI testing algorithm over $[k_A] \times [k_C]$ with sample complexity $3T$ and the same success probability.*

Proof. Map $[k_A] \times [k_C]$ with $[n]$ via some fixed bijection. Using $\mathcal{A}$, we construct a MI tester for P_{AC} over $[k_A] \times [k_C]$ as follows:

1. Draw $3T$ i.i.d. samples $(a_1, c_1), \dots, (a_{3T}, c_{3T})$ from P_{AC} .
2. Form T samples of $P := P_{AC}$, denoted by (a_i, c_i) for $i = 1, \dots, T$.
3. Form T samples of $Q := P_A \otimes P_C$, denoted by (a_{T+2i-1}, c_{T+2i}) for $i = 1, \dots, T$. Since the samples are i.i.d., the a - and c -coordinates of each pair are independent draws from P_A and P_C respectively.
4. Run $\mathcal{A}$ on the T samples of P and the T samples of Q .

Observe that $|H(P) - H(Q)| = |H(A, C) - H(A) - H(C)| = I(A : C)$. Then:

- If $I(A : C) = 0$, then $P_{AC} = P_A \otimes P_C$, so $P = Q$ and $\mathcal{A}$ returns **accept**.
- If $I(A : C) \geq \varepsilon$, then $|H(P) - H(Q)| \geq \varepsilon$ and $\mathcal{A}$ returns **reject**. ◀

► **Theorem 20** (EET lower bound: Theorem 3, restated). *Any algorithm that, given sample access to two unknown distributions p, q over $[n]$ and parameter $\varepsilon \in (0, 1]$, distinguishes $p = q$ from $|H(p) - H(q)| \geq \varepsilon$ with probability at least $2/3$ requires*

$$\tilde{\Omega}\left(\min\left\{\frac{n^{3/4}}{\varepsilon}, \frac{n^{2/3}}{\varepsilon^{4/3}}\right\}\right) + \Omega\left(\frac{\log^2 n}{\varepsilon^2}\right)$$

samples from each of p and q .

Proof. Apply Lemma 19 with $k_A = n/2$ and $k_C = 2$. Combining with [37, Corollary 6.2] gives the first term:

$$\tilde{\Omega}\left(\min\left\{\frac{(n/2)^{3/4} \cdot 2^{1/4}}{\varepsilon}, \frac{(n/2)^{2/3} \cdot 2^{1/3}}{\varepsilon^{4/3}}\right\}\right) = \tilde{\Omega}\left(\min\left\{\frac{n^{3/4}}{\varepsilon}, \frac{n^{2/3}}{\varepsilon^{4/3}}\right\}\right).$$

For the second term, [15, Lemma 2.7] exhibits distributions p, q over $[n]$ with $|H(p) - H(q)| \geq \varepsilon$ that no algorithm can distinguish from $p = q$ using fewer than $\Omega(\log^2 n / \varepsilon^2)$ samples from p when q is given. Any EET algorithm using T samples from each of p, q yields such a tester by simulating T samples from the known q , so EET also requires $\Omega(\log^2 n / \varepsilon^2)$ samples. ◀

References

- 1 Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *NeurIPS*, pages 3591–3599, 2015. URL: <https://proceedings.neurips.cc/paper/2015/hash/1f36c15d6a3d18d52e8d493bc8187cb9-Abstract.html>.
- 2 Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. The complexity of estimating rényi entropy. In *SODA*, pages 1855–1869. SIAM, 2015. doi:10.1137/1.9781611973730.124.
- 3 Tugkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating the entropy. *SIAM J. Comput.*, 35(1):132–150, 2005. doi:10.1137/S0097539702403645.
- 4 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. doi:10.1109/SFCS.2001.959920.

- 5 Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *FOCS*, pages 259–269. IEEE Computer Society, 2000. doi: 10.1109/SFCS.2000.892113.
- 6 Arnab Bhattacharyya, Clément L. Canonne, and Joy Qiping Yang. Independence testing for bounded degree bayesian networks. In *NeurIPS*, 2022.
- 7 Arnab Bhattacharyya, Davin Choo, Sutanu Gayen, and Dimitrios Myrisiotis. Learnability of parameter-bounded bayes nets. In *AAAI*, pages 15559–15566. AAAI Press, 2025. doi: 10.1609/AAAI.V39I15.33708.
- 8 Arnab Bhattacharyya, Sutanu Gayen, Philips George John, Sayantan Sen, and N. V. Vinodchandran. Distribution learning meets graph structure sampling. In *Advances in Neural Information Processing Systems*, volume 38, pages 21411–21432, 2025.
- 9 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. doi:10.1137/22M1489678.
- 10 Clément L. Canonne. *A Survey on Distribution Testing: Your Data is Big. But is it Blue?* Number 9 in Graduate Surveys. Theory of Computing Library, 2020. doi:10.4086/toc.gs.2020.009.
- 11 Clément L. Canonne. Topics and techniques in distribution testing: A biased but representative sample. *Found. Trends Commun. Inf. Theory*, 19(6):1032–1198, 2022. doi: 10.1561/0100000114.
- 12 Clément L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Testing Bayesian networks. In *Conference on Learning Theory*, pages 370–448. PMLR, 2017. URL: <http://proceedings.mlr.press/v65/canonnel17a.html>.
- 13 Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing conditional independence of discrete distributions. In *STOC*, pages 735–748. ACM, 2018. doi:10.1145/3188745.3188756.
- 14 Clément L. Canonne, Yash Pote, Jonathan Scarlett, and Joy Qiping Yang. Entropy equivalence testing. *CoRR*, abs/2605.23225, 2026. doi:10.48550/arXiv.2605.23225.
- 15 Clément L Canonne and Joy Q Yang. Entropy testing and its application to testing Bayesian networks. *Advances in Neural Information Processing Systems*, 2024.
- 16 Clément L. Canonne. A short note on learning discrete distributions, 2020. arXiv:2002.11457.
- 17 Siu On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203. SIAM, 2014. doi:10.1137/1.9781611973402.88.
- 18 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. URL: <https://jmlr.org/papers/volume5/chickering04a/chickering04a.pdf>.
- 19 Davin Choo, Joy Qiping Yang, Arnab Bhattacharyya, and Clément L. Canonne. Learning bounded-degree polytrees with known skeleton. In *ALT*, Proceedings of Machine Learning Research, pages 402–443. PMLR, 2024. URL: <https://proceedings.mlr.press/v237/choo24a.html>.
- 20 C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theory*, 14(3):462–467, 1968. doi:10.1109/TIT.1968.1054142.
- 21 Imre Csiszár and János Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.
- 22 Sanjoy Dasgupta. Learning polytrees. In *UAI*, pages 134–141. Morgan Kaufmann, 1999. URL: https://dslpitt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=162&proceeding_id=15.
- 23 Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing ising models. In *SODA*, pages 1989–2007. SIAM, 2018. doi:10.1137/1.9781611975031.130.

- 24 Constantinos Daskalakis, Gautam Kamath, and John Wright. Which distribution distances are sublinearly testable? In *SODA*, pages 2747–2764. SIAM, 2018. doi:10.1137/1.9781611975031.175.
- 25 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. URL: <http://proceedings.mlr.press/v65/daskalakis17a.html>.
- 26 Ilias Diakonikolas, Themis Gouleakis, Daniel M. Kane, John Peebles, and Eric Price. Optimal testing of discrete distributions with high probability. In *STOC*, pages 542–555. ACM, 2021. doi:10.1145/3406325.3450997.
- 27 Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *ICALP*, LIPIcs, pages 41:1–41:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018. doi:10.4230/LIPIcs.ICALP.2018.41.
- 28 Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *FOCS*, pages 685–694. IEEE Computer Society, 2016. doi:10.1109/FOCS.2016.78.
- 29 Oded Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017. doi:10.1017/9781108135252.
- 30 Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *J. ACM*, 45(4):653–750, 1998. doi:10.1145/285055.285060.
- 31 Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory*, 61(5):2835–2885, 2015. doi:10.1109/TIT.2015.2412945.
- 32 Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Trans. Inf. Theory*, 54(10):4750–4755, 2008. doi:10.1109/TIT.2008.928987.
- 33 Yury Polyanskiy and Yihong Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.
- 34 Ronitt Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012. doi:10.1145/2331042.2331052.
- 35 Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM J. Comput.*, 25(2):252–271, 1996. doi:10.1137/S0097539793255151.
- 36 Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Trans. Inf. Theory*, 62(11):5973–6006, 2016. doi:10.1109/TIT.2016.2603151.
- 37 Jan Seyfried, Sayantan Sen, and Marco Tomamichel. Testing (conditional) mutual information. *Electron. Colloquium Comput. Complex.*, TR25, 2025. URL: <https://eccc.weizmann.ac.il/report/2025/076>.
- 38 Flemming Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Trans. Inf. Theory*, 46(4):1602–1609, 2000. doi:10.1109/18.850703.
- 39 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. doi:10.1145/1993636.1993727.
- 40 Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *FOCS*, pages 51–60. IEEE Computer Society, 2014. doi:10.1109/FOCS.2014.14.
- 41 Yihong Wu and Pengkun Yang. Optimal entropy estimation on large alphabets via best polynomial approximation. In *ISIT*, pages 824–828. IEEE, 2015. doi:10.1109/ISIT.2015.7282570.
- 42 Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inf. Theory*, 2016. doi:10.1109/TIT.2016.2548468.

A

 Deferred Proofs

► **Fact 5.** Let X be a Poisson random variable with parameter λ , and $f: \mathbb{R} \rightarrow \mathbb{R}$ be any function, then, we have that $\mathbb{E}[(X)_m \cdot f(X)] = \lambda^m \cdot \mathbb{E}[f(X + m)]$. In particular, choosing $f \equiv 1$, we have that $\mathbb{E}[(X)_m] = \lambda^m$ for all $m \geq 0$.

Proof. We first show that $\mathbb{E}[X \cdot f(X)] = \lambda \cdot \mathbb{E}[f(X + 1)]$ by standard Poisson calculations:

$$\begin{aligned} \mathbb{E}[X \cdot f(X)] &= \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} \cdot k \cdot f(k) = \lambda \cdot \sum_{k=1}^{\infty} \frac{\lambda^{k-1} e^{-\lambda}}{(k-1)!} \cdot f(k) = \lambda \cdot \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} f(k+1) \\ &= \lambda \cdot \mathbb{E}[f(X + 1)] \end{aligned}$$

The proof is then completed by induction, by noting that

$$\begin{aligned} \mathbb{E}[(X)_m \cdot f(X)] &= \mathbb{E}[X \cdot (X-1)_{m-1} \cdot f(X)] \\ &= \lambda \cdot \mathbb{E}[(X-1+1)_{m-1} \cdot f(X+1)] \\ &= \lambda^2 \cdot \mathbb{E}[(X)_{m-2} \cdot f(X+2)] \\ &= \dots \\ &= \lambda^m \cdot \mathbb{E}[f(X+m)]. \end{aligned}$$

◀