Proceedings of Machine Learning Research vol 132:1–30, 2021 32nd International Conference on Algorithmic Learning Theory

Testing Product Distributions: A Closer Look

Arnab Bhattacharyya ARNABB@NUS.EDU.SG
National University of Singapore, Singapore

Sutanu Gayen SUTANUGAYEN@GMAIL.COM
National University of Singapore, Singapore

Saravanan Kandasamy SK3277@CORNELL.EDU
Cornell University, Ithaca NY, USA

N. V. Vinodchandran VINOD@CSE.UNL.EDU
University of Nebraska-Lincoln, Lincoln NE, USA

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

We study the problems of *identity* and *closeness testing* of n-dimensional product distributions. Prior works of Canonne et al. (2017) and Daskalakis and Pan (2017) have established tight sample complexity bounds for *non-tolerant testing over a binary alphabet*: given two product distributions P and Q over a binary alphabet, distinguish between the cases $P = Q$ and $d_{\mathrm{TV}}(P, Q) > \epsilon$. We build on this prior work to give a more comprehensive map of the complexity of testing of product distributions by investigating *tolerant testing with respect to several natural distance measures and over an arbitrary alphabet*. Our study gives a fine-grained understanding of how the sample complexity of tolerant testing varies with the distance measures for product distributions. In addition, we also extend one of our upper bounds on product distributions to bounded-degree Bayes nets.

Keywords: property testing, distribution testing, product distributions

1. Introduction

The main goal of this work is to give a comprehensive investigation to the sample complexity of several distribution testing problems over *high-dimensional product distributions*. Testing properties of distributions from samples has been actively investigated for several decades from the perspectives of classical statistics and, more recently, from a property testing viewpoint in theoretical computer science. Hypothesis testing is a classical problem investigated in statistics with significant practical applications. From the property testing viewpoint, the two most well studied distribution testing problems are *identity testing* and *closeness testing.*[1]

In the *identity testing problem*, we are given a known reference distribution Q and sample access to an unknown distribution P over the same sample space as that of Q, and the goal is to distinguish between the cases $P = Q$ or P is ϵ-far from Q with respect to a given distance measure. It is known that $\Theta(\sqrt{m}/\epsilon^2)$ samples are necessary and sufficient to solve the identity testing problem with respect to the total variation distance, where m is the size of the sample space (Valiant and Valiant, 2014; Paninski, 2008). In the *closeness testing problem*, we have sample access to a pair of unknown distributions P and Q on a common sample space, and the goal is to distinguish between the cases $P = Q$ or P is ϵ-far from Q with respect to a certain distance measure. It is known that

1. Identity testing is also known as *goodness-of-fit testing* or *one-sample testing* in the literature. Similarly, closeness testing is also known as *two-sample testing*.

$\Theta(\max(m^{2/3}\epsilon^{-4/3}, \sqrt{m}\epsilon^{-2}))$ samples are necessary and sufficient to solve the closeness testing problem with respect to the total variation distance, where m is the size of the sample space (Chan, Diakonikolas, Valiant, and Valiant, 2014). See the surveys Rubinfeld (2012); Canonne (2020) and the references therein for pointers to the extensive research on identity and closeness testing as well as related problems.

One of the main bottlenecks resulting from the above-mentioned complexity bounds is that these testing problems are (provably) hard for arbitrary distributions over large sample spaces. For example, for distributions over an n-dimensional Boolean hypercube $m = 2^n$ and hence $\Theta(2^{\frac{n}{2}})$ samples are necessary and sufficient for identity testing (for a constant ϵ). To overcome this bottleneck, very recently researchers have started investigating testing problems over high-dimensional sample spaces by imposing natural structural assumptions over distributions. Such assumptions restrict the class of distributions and open up the possibility of designing testers with substantially smaller sample complexity than required for the the general case. Among them *product distributions* over a finite alphabet are a natural class that is both practically relevant and simple enough to serve as a test ground for algorithm design. Indeed, prior works of Canonne, Diakonikolas, Kane, and Stewart (2017) and Daskalakis and Pan (2017) have established tight sample complexity bounds for identity and closeness testing of product distributions over a binary alphabet.

A drawback of the testing problems as stated is their one-sided or *non-tolerant* aspect: on the one side of the decision, we only need to distinguish from the case where two distributions are *exactly equal*. This is a significant restriction specially for high-dimensional distributions which require a large number of parameters to be specified. For example, in the case of identity testing, it is unlikely that we can ever hypothesize a reference distribution Q such that it exactly equals the data distribution P. Similarly, for closeness testing, two data distributions P and Q are most likely not exactly equal. The *tolerant* version of testing problems addresses this issue as it seeks to design testers for identity and closeness that *tolerate* errors on both decision cases. That is, in the tolerant version we would like to distinguish between the cases $d_{\mathrm{TV}}(P, Q) \leq \epsilon_1$ and $d_{\mathrm{TV}}(P, Q) > \epsilon_2$, where $\epsilon_1 < \epsilon_2$ are user-supplied error parameters. The tolerance requirement makes the testing problems more expensive. For arbitrary distributions on a set of size m it is known (Valiant and Valiant, 2010) that tolerant identity and closeness testing of arbitrary distributions supported on a set of size m require $\Omega(m/\log m)$ samples for constants $\epsilon_1 < \epsilon_2$, if the distance measure used is the total variation distance.

The main focus of this paper is to take a closer look at the complexity of testing of product distributions by investigating *tolerant testing with respect to several natural distance measures and over an arbitrary alphabet*. Such an investigation is important because a complete picture on the complexity of testing product distribution will shed light on possibilities and challenges in algorithm design for testing high dimensional structured distributions.

1.1. Our Contributions

We investigate tolerant testing of product distributions with respect to the following distance measures: total variation distance (d_{TV}), Hellinger distance (d_{H}), Kullback-Leibler divergence (d_{KL}), and Chi-squared distance (d_{χ^2})[2]. The following relationship is well-known among them:

$$d_{\mathrm{H}}^2(P, Q) \leq d_{\mathrm{TV}}(P, Q) \leq \sqrt{2} d_{\mathrm{H}}(P, Q) \leq \sqrt{d_{\mathrm{KL}}(P, Q)} \leq \sqrt{d_{\chi^2}(P, Q)} \tag{1}$$

2. Refer to Section 2 for the notations and definitions.

We fix a pair of distance functions $d_1 \leq d_2$ from the above equation and investigate the problem of deciding $d_1(P,Q) \leq \epsilon/3$ versus $d_2(P,Q) > \epsilon$ with 2/3 probability, which we call d_1-versus-d_2 testing. When both P and Q are only accessed by samples, this problem is called d_1-versus-d_2 closeness testing. When Q is a reference distribution given to us and P is accessed by samples, the problem is called d_1-versus-d_2 identity testing. The problem of distinguishing $P = Q$ versus $d_2(P,Q) > \epsilon$ is called non-tolerant testing w.r.t. d_2. Clearly, tolerant testing is at least as hard as non-tolerant testing.

Our contributions regarding d_1-versus-d_2 identity and closing testing problems over product distributions are summarized in Table 1 and Table 2. Each cell of the tables represents the sample complexity of testing whether the two product distributions are close or far in terms of the distance corresponding to that row and column respectively. The problems become harder as we traverse the table down or to the right due to Equation (1). Daskalakis, Kamath, and Wright (2018) have shown that non-tolerant testing w.r.t. d_{KL} is not testable in a finite set of samples. Hence, only d_{TV} and d_{H} are meaningful for d_2.

Table 1: Sample complexity upper and lower bounds for d_1-*vs*-d_2 *identity testing* of product distributions for various distance measures. First column (row) lists $d_1(P,Q) \leq \epsilon_1$ (respectively, $d_2(P,Q) > \epsilon_2$). The problem becomes computationally more difficult, and hence the sample complexity is non-decreasing, as we traverse the table down or to the right. [†], [∗] and [‡] are from Daskalakis and Pan (2017), Canonne, Diakonikolas, Kane, and Stewart (2017) and Bhattacharyya, Gayen, Meel, and Vinodchandran (2020), respectively. Note that for some of the cells, to get to the bound we need to follow a chain of directions.

	$d_{\mathrm{TV}}(P,Q) > \epsilon$	$\sqrt{2}d_{\mathrm{H}}(P,Q) > \epsilon$
$P = Q$	*UB* : $O(\sqrt{n}/\epsilon^2)$ (for $\lvert\Sigma\rvert = 2$) [†,∗] *LB* : $\Omega(\sqrt{n}/\epsilon^2)$ (for $\lvert\Sigma\rvert = 2$) [†,∗] *LB* : $\Omega(\sqrt{n\lvert\Sigma\rvert}/\epsilon^2)$ (for $\lvert\Sigma\rvert > 2$) Theorem 2.3	*UB* : *Below* *LB* : *Left*
$d_{\chi^2}(P,Q) \leq \epsilon^2/9$	*UB* : *Right* *LB* : *Above*	*UB* : $O(\sqrt{n\lvert\Sigma\rvert}/\epsilon^2)$ Theorem 2.1 *LB* : *Left*
$d_{\mathrm{KL}}(P,Q) \leq \epsilon^2/9$	*UB* : *Below* *LB* : $\Omega(n/\log n)$ Theorem 2.4	*UB* : *Below* *LB* : *Left*
$\sqrt{2}d_{\mathrm{H}}(P,Q) \leq \epsilon/3$	*UB* : *Right* *LB* : *Above*	*UB* : $O(n\lvert\Sigma\rvert/\epsilon^2)$ Theorem 2.2 *LB* : *Left*
$d_{\mathrm{TV}}(P,Q) \leq \epsilon/3$	*UB* : $O(n\lvert\Sigma\rvert/\epsilon^2)$ [‡] *LB* : $\Omega(n/\log n)$[∗]	*Not Well Defined*

We informally present our main results below. We would like to note that the only algorithmic results known regarding the complexity of testing product distributions prior to our work are:

(1) $\Theta(\sqrt{n}/\epsilon^2)$ sample complexity bound for the non-tolerant identity testing over the binary alphabet (Daskalakis and Pan, 2017; Canonne, Diakonikolas, Kane, and Stewart, 2017),

(2) $\Theta(\max(\sqrt{n}/\epsilon^2, n^{3/4}/\epsilon))$ sample complexity bound for non-tolerant closeness testing problem over the binary alphabet (Canonne, Diakonikolas, Kane, and Stewart, 2017), and

(3) $O(n|\Sigma|/\epsilon^2)$ upper bound for d_{TV}-vs-d_{TV} tolerant identity and closeness testing (Bhattacharyya, Gayen, Meel, and Vinodchandran, 2020).

Table 2: Sample complexity bounds for of d_1-*vs*-d_2 *closeness testing* of product distributions. As in the case of identity testing, sample complexity is non-decreasing as we traverse the table down or to the right. [$*$] and [‡] are from Canonne, Diakonikolas, Kane, and Stewart (2017) and Bhattacharyya, Gayen, Meel, and Vinodchandran (2020), respectively.

	$d_{\mathrm{TV}}(P,Q) > \epsilon$	$\sqrt{2}d_{\mathrm{H}}(P,Q) > \epsilon$
$P = Q$ $(\Sigma = 2)$	*UB* : $O(\max(\sqrt{n}/\epsilon^2, n^{3/4}/\epsilon))$ [*] *LB* : $\Omega(\max(\sqrt{n}/\epsilon^2, n^{3/4}/\epsilon))$ [*]	*UB* : $O(n^{3/4}/\epsilon^2)$, Theorem 2.6 *LB* : *Left*
(Any Σ*)*	*UB* : $O\Big(\max\Big\{\sqrt{n\lvert\Sigma\rvert}/\epsilon^2, (n\lvert\Sigma\rvert)^{3/4}/\epsilon\Big\}\Big)$ Theorem 2.7 *LB* : *Above*	*UB* : $O((n\lvert\Sigma\rvert)^{3/4}/\epsilon^2)$, Theorem 2.6 *LB* : *Above*
$d_{\chi^2}(P,Q) \le \epsilon^2/9$	*UB* : *Below* *LB* : $\Omega(n/\log n)$ Theorem 2.5	*UB* : *Below* *LB* : *Left*
$d_{\mathrm{KL}}(P,Q) \le \epsilon^2/9$	*UB* : *Below* *LB* : *Above*	*UB* : *Below* *LB* : *Left*
$\sqrt{2}d_{\mathrm{H}}(P,Q) \le \epsilon/3$	*UB* : Right *LB* : *Above*	*UB* : $O(n\lvert\Sigma\rvert/\epsilon^2)$ Theorem 2.2 *LB* : *Left*
$d_{\mathrm{TV}}(P,Q) \le \epsilon/3$	*UB* : $O(n\lvert\Sigma\rvert/\epsilon^2)$ [‡] *LB* : $\Omega(n/\log n)$[*]	*Not Well-defined*

IDENTITY TESTING: P UNKNOWN AND Q GIVEN

- We present a tolerant identity testing algorithm that distinguishes $d_{\chi^2}(P,Q) \le \epsilon^2/9$ versus $d_{\mathrm{H}}(P,Q) > \epsilon$ with $O(\sqrt{n|\Sigma|}/\epsilon^2)$ sample complexity. Since the condition for the tester rejecting $d_{\mathrm{H}}(P,Q) > \epsilon$ is stronger than $d_{\mathrm{TV}}(P,Q) > \epsilon$ due to Equation (1), we get the same bound when the second distance is d_{TV}. Our algorithm applies for an arbitrary Σ and has optimal dependence on n, $|\Sigma|$ and ϵ.
- We present an algorithm for d_{H}-vs-d_{H} identity testing with sample complexity $O(n|\Sigma|/\epsilon^2)$.
- Our third result is a lower bound: we establish the optimality of the sample complexity of our non-tolerant identity tester w.r.t. d_{TV}in terms of n, $|\Sigma|$ and ϵ. Such a lower bound was previously known only for $|\Sigma| = 2$.
- Our next result is another lower bound: we show that the identity testing problem of distinguishing $d_{\mathrm{KL}}(P,Q) \le \epsilon^2/9$ versus $d_{\mathrm{TV}}(P,Q) > \epsilon$ requires at least $\Omega(n/\log n)$ samples. This shows a jump in sample complexity when we move to d_{KL} from χ^2 distance. Previously, Canonne, Diakonikolas, Kane, and Stewart (2017) had shown the $\Omega(n/\log n)$ lower bound for d_{TV}-vs-d_{TV} testing; we strengthen it to d_{KL}-vs-d_{TV} testing.

CLOSENESS TESTING: P AND Q UNKNOWN

- We design an efficient algorithm that distinguishes $\sqrt{2}d_{\mathrm{H}}(P,Q) \le \epsilon/3$ versus $\sqrt{2}d_{\mathrm{H}}(P,Q) > \epsilon$ with $O(n|\Sigma|/\epsilon^2)$ sample complexity. (Note that this result appears in both Table 1 and Table 2). Our upper bound works for distributions over arbitrary alphabet.
- We complement the above upper bound with a new lower bound. We show that given sample access to two unknown distributions P and Q, distinguishing $d_{\chi^2}(P,Q) \le \epsilon^2/9$ from

$d_{\mathrm{TV}}(P, Q) > \epsilon$ requires $\Omega(n/\log n)$ samples, even for $|\Sigma| = 2$ and constant ϵ. Note that this is in contrast to identity testing, where Table 1 shows that the same problem can be solved using $O(\sqrt{n|\Sigma|}/\epsilon^2)$ samples. This also strengthens the d_{TV}-vs-d_{TV} lower bound of Canonne, Diakonikolas, Kane, and Stewart (2017).

We also establish new upper bounds for non-tolerant closeness testing over arbitrary alphabet. Prior work considered only the binary alphabet and the extension to arbitrary alphabet is not completely straightforward.

TOLERANT TESTING FOR BAYES NETS

A more general class of probability distributions, containing product distributions as a special case, is bounded-degree *Bayesian networks* (or Bayes nets in short). Formally, a probability distribution P over n variables $X_1, \ldots, X_n \in \Sigma$ is said to be a *Bayesian network on a directed acyclic graph* G with n nodes if[3] for every $i \in [n]$, X_i is conditionally independent of $X_{\text{non-descendants}(i)}$ given $X_{\text{parents}(i)}$. Equivalently, P admits the factorization:

$$\Pr_{X\sim P}[X = x] = \prod_{i=1}^{n} \Pr_{X\sim P}[X_i = x_i \mid \forall j \in \text{parents}(i), X_j = x_j] \qquad \text{for all } x \in \Sigma^n \tag{2}$$

For example, product distributions are Bayes nets on the empty graph. A *degree-d Bayes net* is a Bayes net on a graph with in-degree bounded by d.

We consider tolerant closeness testing of degree-d Bayes nets on known directed acyclic graphs. Bhattacharyya, Gayen, Meel, and Vinodchandran (2020) designed an algorithm for tolerant d_{TV}-vs-d_{TV} closeness testing with $\tilde{O}(|\Sigma|^{d+1} n\epsilon^{-2})$ sample complexity. Our main result for Bayes nets extends this same bound to d_{H}-vs-d_{H} testing, which is the hardest variant of the tolerant testing problems considered above. Moreover, our test is computationally efficient (in terms of time complexity). Note that a computationally inefficient test readily follows from available *learning* algorithms for fixed-structure Bayes nets with respect to KL divergence (Dasgupta, 1997; Bhattacharyya, Gayen, Meel, and Vinodchandran, 2020). Indeed, the main technical component in our result is a novel efficient estimator for Hellinger distance between two distributions when given access to samples generated from them as well as their probability mass functions. This estimator may be of independent interest.

1.2. Related Work

The history of identity tests goes back to Pearson's chi-squared test in 1900. The traditional spirit of analyzing such tests is to consider a fixed distribution P and to let the number of samples go to infinity. Work on understanding the performance of hypothesis tests with a finite number of samples mostly started only quite recently. Goldreich and Ron (2011) studied the problem of distinguishing whether an input distribution P is uniform over its support or ϵ-far from uniform in total variation distance (in fact, they showed a *tolerant* tester with respect to the ℓ_2-norm). Paninski showed that $\Theta(\sqrt{m}/\epsilon^2)$ samples are necessary for uniformity testing, and gave an optimal tester when $\epsilon > m^{-1/4}$ (where m is the size of the support). For the more general problem of testing identity to an arbitrary given distribution, Batu, Fortnow, Rubinfeld, Smith, and White (2013) showed an upper bound of

3. We use the notation X_S to denote $\{X_i : i \in S\}$ for a set $S \subseteq [n]$.

$\tilde{O}(\sqrt{m}/\epsilon^6)$. This was then refined by Valiant and Valiant (2014) to the tight bound of $\Theta(\sqrt{m}/\epsilon^2)$. Batu, Fortnow, Rubinfeld, Smith, and White (2013) also studied the problem of testing closeness between two input distributions and showed an upper bound of $\tilde{O}(m^{2/3}\text{poly}(1/\epsilon))$ on the sample complexity. The tight bound of $\Theta(\max(m^{2/3}\epsilon^{-4/3}, \sqrt{m}\epsilon^{-2}))$ was achieved by Chan, Diakonikolas, Valiant, and Valiant (2014). Tolerant versions of uniformity, identity, and closeness testing with respect to the total variation distance require $\Omega(m/\log m)$ samples Valiant and Valiant (2011), which is also tight Valiant and Valiant (2010). To circumvent this lower bound, tolerant identity testing with respect to chi-squared distance was initiated by Acharya, Daskalakis, and Kamath (2015) and was thoroughly studied in Daskalakis, Kamath, and Wright (2018) for a number of pairs of distances.

The study of testing distributions over high-dimensional domains was initiated recently independently and concurrently in Daskalakis, Dikkala, and Kamath (2019); Canonne, Diakonikolas, Kane, and Stewart (2017); Daskalakis and Pan (2017), who recognized that since testing arbitrary distributions over Σ^n would require an exponential number of samples, it is important to make structural assumptions on the distribution. In particular, in Daskalakis, Dikkala, and Kamath (2019), they make the assumption that the input distributions are drawn from an Ising model. In Canonne, Diakonikolas, Kane, and Stewart (2017) and Daskalakis and Pan (2017), the authors considered identity testing and closeness testing for distributions given by Bayes networks of bounded in-degree. These works also considered the special case of product distributions (equivalently, distributions over a Bayes network consisting of isolated nodes). It's shown that $\Theta(\sqrt{n}/\epsilon^2)$ and $\Theta(\max(\sqrt{n}/\epsilon^2, n^{3/4}/\epsilon))$ samples are necessary and sufficient for identity testing and closeness testing respectively of pairs of product distributions when $|\Sigma| = 2$. The identity tester of Canonne, Diakonikolas, Kane, and Stewart (2017) is claimed to have certain weaker ($O(\epsilon^2)$ in d_{TV}, see Remark 8) tolerance. A reduction from testing problems for product distributions over alphabet Σ, to that for the Bayes nets of degree $\lfloor \log_2 |\Sigma| \rfloor - 1$, was given in Canonne, Diakonikolas, Kane, and Stewart (2017) (Remark 55 of their paper). Canonne, Diakonikolas, Kane, and Stewart (2017) also show that for product distributions, $\Omega(n/\log n)$ samples are necessary for tolerant identity and closeness testing with respect to the total variation distance. Very recently, Bhattacharyya, Gayen, Meel, and Vinodchandran (2020) designed tolerant testers for certain classes of high-dimensional distributions (including product distributions) with respect to d_{TV}.

2. Preliminaries and Formal Statements of Results

We use $\mathrm{Bern}(\delta)$ to denote the Bernoulli distribution with $\Pr[1] = \delta$. We define various distance measures between distributions that we use in this paper.

Definition 1 *Let $P = (p_1, p_2, \ldots, p_m)$ and $Q = (q_1, q_2, \ldots, q_m)$ be two distributions over sample space $[m]$. Then the distance measures, total variational distance, chi-squared distance, Hellinger distance, and KL distance, respectively are defined as follows.*

$$d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\sum_i |p_i - q_i|; \qquad d_{\chi^2}(P,Q) = \sum_i (p_i - q_i)^2/q_i = \sum_i p_i^2/q_i - 1;$$

$$d_{\mathrm{H}}^2(P,Q) = \frac{1}{2}\sum_i (\sqrt{p_i} - \sqrt{q_i})^2 = 1 - \sum_i \sqrt{p_i q_i}; \qquad d_{\mathrm{KL}}(P,Q) = \sum_i p_i \ln \frac{p_i}{q_i}$$

Lemma 2 *(folklore, see Daskalakis, Kamath, and Wright (2018) for a proof) For two distributions P and Q, the following relation holds.*

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

2.1. Formal Statements of Main Results

Here we list the formal statements of the main theorems we prove in the paper. First we state the two main upper bounds.

Theorem 2.1 (d_{χ^2}-versus-d_{H} identity tester) *There is an algorithm with sample access to an unknown product distribution $P = \prod_{i=1}^n P_i$ and input a known product distribution $Q = \prod_{i=1}^n Q_i$, both over the common sample space Σ^n, that decides between cases $d_{\chi^2}(P,Q) \le \epsilon^2/9$ versus $\sqrt{2}d_{\mathrm{H}}(P,Q) > \epsilon$. The algorithm takes $O(\sqrt{n|\Sigma|}/\epsilon^2)$ samples from P and runs in time $O(n\ell + n^{3/2}\sqrt{\ell}/\epsilon^2)$. The algorithm has a success probability at least $2/3$.*

Theorem 2.2 (d_{H}-versus-d_{H} closeness tester) *There is an algorithm with sample access to two unknown product distribution $P = \prod_{i=1}^n P_i$ and $Q = \prod_{i=1}^n Q_i$, both over the common sample space Σ^n, that decides between cases $\sqrt{2}d_{\mathrm{H}}(P,Q) \le \epsilon/3$ versus $\sqrt{2}d_{\mathrm{H}}(P,Q) > \epsilon$. The algorithm takes $O(n(|\Sigma| + \log n)/\epsilon^2)$ samples from P and Q and runs in time $O(n^2(|\Sigma| + \log n)/\epsilon^2)$. The algorithm has a success probability at least $2/3$.*

We complement the above upper bounds on sample complexity with the following lower bounds.

Theorem 2.3 *Uniformity testing with w.r.t. d_{TV} distance for product distributions over $[\ell]^n$ needs $\Omega(\sqrt{n\ell}/\epsilon^2)$ samples.*

Theorem 2.4 (d_{KL}-versus-d_{TV} identity testing lower bound) *There exists a constant $0 < \epsilon < 1$ and three product distributions F^{yes}, F^{no} and F, each over the sample space $\{0,1\}^n$ such that $d_{\mathrm{KL}}(F^{yes}, F) \le \epsilon^2/9$, whereas $d_{\mathrm{TV}}(F^{no}, F) > \epsilon$, and given only sample accesses to F^{yes}, F^{no}, and complete knowledge about F, distinguishing F^{yes} versus F^{no} with probability $> 2/3$, requires $\Omega(n/\log n)$ samples.*

Theorem 2.5 (d_{χ^2}-versus-d_{TV} closeness testing lower bound) *There exists a constant $0 < \epsilon < 1$ and three product distributions F^{yes}, F^{no} and F, each over the sample space $\{0,1\}^n$ such that $d_{\chi^2}(F^{yes}, F) \le \epsilon^2/9$, whereas $d_{\mathrm{TV}}(F^{no}, F) > \epsilon$, and given only sample accesses to F^{yes}, F^{no} and F, distinguishing F^{yes} versus F^{no} with probability $> 2/3$, requires $\Omega(n/\log n)$ samples.*

Earlier work has designed non-tolerant closeness tester for product distribution over a binary alphabet. Here we extend it to arbitrary alphabets.

Theorem 2.6 (Exact-versus-d_{H} closeness tester) *There is an algorithm with sample access to two unknown product distribution $P = \prod_{i=1}^n P_i$ and $Q = \prod_{i=1}^n Q_i$, both over the common sample space Σ^n, that decides between cases $P = Q$ versus $\sqrt{2}d_{\mathrm{H}}(P,Q) > \epsilon$. The algorithm takes $m = O((n|\Sigma|)^{3/4}/\epsilon^2)$ samples from P and Q and runs in time $O(mn)$. The algorithm has a success probability at least $2/3$.*

Theorem 2.7 (Exact-versus-d_{TV} closeness tester) *There is an algorithm with sample access to two unknown product distribution $P = \prod_{i=1}^n P_i$ and $Q = \prod_{i=1}^n Q_i$, both over the common sample space Σ^n, that decides between cases $P = Q$ versus $d_{\mathrm{TV}}(P, Q) > \epsilon$. The algorithm takes $m = O(\max\{\sqrt{n|\Sigma|}/\epsilon^2, (n|\Sigma|)^{3/4}/\epsilon\})$ samples and runs in time $O(mn)$. The algorithm has a success probability at least $2/3$.*

Finally, we state our result for d_{H}-vs-d_{H} closeness testing of fixed-structure Bayes nets.

Theorem 2.8 (d_{H}-vs-d_{H} closeness tester for Bayes nets) *Given samples from two unknown Bayesian networks P and Q over Σ^n on potentially different but known pair of graphs of indegree at most d, we can distinguish the cases $d_{\mathrm{H}}(P, Q) \le \epsilon/2$ versus $d_{\mathrm{H}}(P, Q) > \epsilon$ with 2/3 probability using $m = O(|\Sigma|^{d+1} n \log(|\Sigma|^{d+1} n)\epsilon^{-2})$ samples and $O(|\Sigma|^{d+1} mn + n\epsilon^{-4})$ time.*

3. Efficient Tolerant Testers

3.1. d_{χ^2}-vs-d_{H} Tolerant Identity Tester

In this section, we generalize the testers of Daskalakis and Pan (2017); Canonne, Diakonikolas, Kane, and Stewart (2017) that distinguishes $P = Q$ ('yes class') versus $d_{\mathrm{TV}}(P, Q) \ge \epsilon$ ('no class') using $O(\sqrt{n}/\epsilon^2)$ samples, where P and Q are product distributions over $\{0, 1\}^n$. Our first contribution is to generalize their tester in the following three ways. Firstly, our 'no class' is defined as $\sqrt{2} d_{\mathrm{H}}(P, Q) \ge \epsilon$, which is more general than $d_{\mathrm{TV}}(P, Q) \ge \epsilon$. Secondly, our tester works for any general alphabet size $|\Sigma| \ge 2$. Finally, we give a d_{χ^2} tolerant tester i.e. our 'yes class' is defined as $d_{\chi^2}(P, Q) < \epsilon^2/9$.

Our tester relies on certain factorizations of d_{χ^2} and d_{H}^2 for product distributions. We shall proceed to discuss those relations.

Lemma 3 *(folklore, see Acharya, Daskalakis, and Kamath (2015) for a proof) Let $P = \prod_{i=1}^n P_i$ and $Q = \prod_{i=1}^n Q_i$ be two distributions, both over the common sample space Σ^n. Then $d_{\chi^2}(P, Q) = \prod_{i=1}^n (1 + d_{\chi^2}(P_i, Q_i)) - 1$. In particular, $d_{\chi^2}(P, Q) \ge \sum_i d_{\chi^2}(P_i, Q_i)$*

Fact 4 *(folklore) Let $P = \prod_{i=1}^n P_i$ and $Q = \prod_{i=1}^n Q_i$ be two distributions over Σ^n. It holds that $1 - d_{\mathrm{H}}^2(P, Q) = \Pi_{i=1}^n (1 - d_{\mathrm{H}}^2(P_i, Q_i))$. In particular, ${d_{\mathrm{H}}}^2(P, Q) \le \sum_i {d_{\mathrm{H}}}^2(P_i, Q_i)$.*

We get the following useful corollary from Lemma 2 and Fact 4.

Corollary 5 *Let $P = \prod_{i=1}^n P_i$ and $Q = \prod_{i=1}^n Q_i$ be two distributions, both over the common sample space Σ^n. Then $d_{\mathrm{H}}^2(P, Q) \le \sum_i d_{\chi^2}(P_i, Q_i)/2$.*

To avoid low probabilities in the denominator of the test statistic, we need to ensure that for each distribution Q_i, each element in the sample space Σ gets at least a sufficiently large probability $\Omega(\epsilon^2/|\Sigma|n)$. We do this by *slightly randomizing* Q to get a new distribution S. The randomization process to get S from Q is given below. This is similar to the reduction given in Daskalakis and Pan (2017) and Canonne, Diakonikolas, Kane, and Stewart (2017) for the case $\Sigma = \{0, 1\}$ and for the case when the 'no' class is defined with respect to d_{TV}. Let $\mathrm{Bern}(\delta)^n$ be the product distribution of n copies of $\mathrm{Bern}(\delta)$.

Lemma 6 *For a product distribution $P = \prod_{i=1}^{n} P_i$, where the P_i's are over a sample space Σ, and $0 < \delta < 1$, let P^δ be the distribution over Σ^n defined by the following sampling process. In order to produce a sample $(X_1, X_2, \ldots, X_n)$ of P^δ,*

- *Sample $(r_1, r_2, \ldots, r_n) \sim \mathrm{Bern}(\delta)^n$ and sample $(Y_1, Y_2, \ldots, Y_n) \sim P$*
- *For every i, if $r_i = 1$, $X_i \leftarrow$ uniform sample from Σ, if $r_i = 0$, $X_i \leftarrow Y_i$.*

Then, the following is true.

- *P^δ is a product distribution $\prod_i P_i^\delta$ and each sample from P^δ can be simulated by 1 sample from P.*
- *For every $i : 1 \le i \le n$ and $j \in \Sigma$, $P_i^\delta(j) \ge \delta/|\Sigma|$.*
- *$d_{\mathrm{H}}^2(P, P^\delta) \le 2n\delta$*

Proof The first part is obvious from the sampling process. For the second part, $P_i^\delta(j) = (1 - \delta)P_i(j) + \delta/|\Sigma| \ge \delta/|\Sigma|$ for every i, j.

The proof of the third part can be obtained by generalizing the proof of Daskalakis, Kamath, and Wright (2018). Consider the i-th component of P and P^δ, denoted P_i and Q_i respectively for convenience. Let E_i be the event that $r_i = 0$. Also note that conditioned on the event E_i, for any item $j \in \Sigma$, the probability values satisfy $Q_i(j \mid E_i) = P_i(j)$.

$$\begin{aligned}
d_{\mathrm{H}}^2(P_i, Q_i) &= \sum_{j\in\Sigma}\left(\sqrt{Q_i(j)} - \sqrt{P_i(j)}\right)^2 \\
&= \sum_{j\in\Sigma}\left(\sqrt{Q_i(j \mid E_i)\Pr(E_i) + Q_i(j \mid \bar{E}_i)\Pr(\bar{E}_i)} - \sqrt{P_i(j)}\right)^2 \\
&= \sum_{j}\left(\sqrt{P_i(j)\Pr(E_i)} - \sqrt{P_i(j)}\right)^2 + \sum_{j} Q_i(j \mid \bar{E}_i)\Pr(\bar{E}_i) \\
&\text{(Using } (\sqrt{a+b} - \sqrt{c+d})^2 \le (\sqrt{a} - \sqrt{c})^2 + (\sqrt{b} - \sqrt{d})^2 \text{ for non-negative } a, b, c, d) \\
&= (1 - \sqrt{\Pr(E_i)})^2 + \Pr(\bar{E}_i) \\
&= (1 - \sqrt{1 - \Pr(\bar{E}_i)})^2 + \Pr(\bar{E}_i) \\
&\le 2\Pr(\bar{E}_i) = 2\delta \qquad \text{(Using } (1 - \sqrt{1-x})^2 \le x \text{ for } 0 \le x \le 1)
\end{aligned}$$

Then Fact 4 gives us $d_{\mathrm{H}}^2(P, P^\delta) \le 2n\delta$ due to sub-additivity. ■

Lemma 7 *Let $P = \prod_{i=1}^{n} P_i$ and $Q = \prod_{i=1}^{n} Q_i$ be two distributions, both over the common sample space Σ^n. Let $\ell = |\Sigma|$, $R = P^\delta$, $S = Q^\delta$ with $\delta = \epsilon^2/50n$. $R = \prod_{i=1}^{n} R_i$ and $S = \prod_{i=1}^{n} S_i$, where $R_i = \langle r_{i1}, r_{i2}, \ldots, r_{i\ell}\rangle$ and $S_i = \langle s_{i1}, s_{i2}, \ldots, s_{i\ell}\rangle$ for every i. Then*

(1) If $d_{\chi^2}(P, Q) \le \epsilon^2/9$ then $\sum_{i,j} \frac{(r_{ij} - s_{ij})^2}{s_{ij}} < 0.12\epsilon^2$.

(2) If $\sqrt{2}d_{\mathrm{H}}(P, Q) \ge \epsilon$ then $\sum_{i,j} \frac{(r_{ij} - s_{ij})^2}{s_{ij}} > 0.18\epsilon^2$.

Proof (*Proof of (1)*) We have that $r_{ij} = (1-\delta)p_{ij} + \delta/\ell$ and $s_{ij} = (1-\delta)q_{ij} + \delta/\ell$. Then, $\sum_{i,j} \frac{(r_{ij}-s_{ij})^2}{s_{ij}} = \sum_{i,j} \frac{(1-\delta)^2(p_{ij}-q_{ij})^2}{(1-\delta)q_{ij}+\delta/\ell} \leq \sum_{i,j} \frac{(1-\delta)^2(p_{ij}-q_{ij})^2}{(1-\delta)q_{ij}} = (1-\delta)\sum_{i,j} \frac{(p_{ij}-q_{ij})^2}{q_{ij}} < \sum_i \sum_j \frac{(p_{ij}-q_{ij})^2}{q_{ij}} = \sum_i d_{\chi^2}(P_i, Q_i) \leq d_{\chi^2}(P, Q) < 0.12\epsilon^2$. The second last step is due to Lemma 3.
(*Proof of (2)*). From Lemma 6, for $\delta = \epsilon^2/50n$, it follows that $d_{\mathrm{H}}^2(P, R) \leq \epsilon^2/25$ and $d_{\mathrm{H}}^2(Q, S) \leq \epsilon^2/25$. By triangle inequality we get $d_{\mathrm{H}}(P, Q) \leq d_{\mathrm{H}}(R, S) + d_{\mathrm{H}}(P, R) + d_{\mathrm{H}}(Q, S)$. It follows that if $\sqrt{2}d_{\mathrm{H}}(P, Q) \geq \epsilon$ then $d_{\mathrm{H}}(R, S) \geq \epsilon(1/\sqrt{2} - 2/5)$. Then Corollary 5 gives $\sum_{i,j} \frac{(r_{ij}-s_{ij})^2}{s_{ij}} = \sum_i d_{\chi^2}(R_i, S_i)) \geq 2d_{\mathrm{H}}^2(R, S) > 0.18\epsilon^2$. ■

At this point it remains to test $\sum_{i,j} \frac{(r_{ij}-s_{ij})^2}{s_{ij}} > 0.18\epsilon^2/10$ versus $< 0.12\epsilon^2/10$, which we perform using the tester of Acharya, Daskalakis, and Kamath (2015). We state their tester with necessary modifications and prove it in the Appendix.

Theorem 3.1 (Modified from Acharya, Daskalakis, and Kamath (2015)) *Let m be an integer and $0 < \epsilon < 1$ be an error parameter. Let $r_1, r_2, \ldots, r_K$ be K non-negative real numbers. Let $s_1, s_2, \ldots, s_K$ be non-negative real numbers such that $s_i \geq \epsilon^2/50K$. For $1 \leq i \leq K$, let $N_i \sim \mathrm{Poi}(mr_i)$ be independent samples from $\mathrm{Poi}(mr_i)$. Then there exists a test statistic T, computable in time $O(K)$ from inputs N_is and s_is, with the following guarantees.*

- $\mathrm{E}[T] = m\sum_i \frac{(r_i-s_i)^2}{s_i}$
- $\mathrm{Var}[T] \leq 2K + 7\sqrt{K}\mathrm{E}[T] + 4K^{1/4}(\mathrm{E}[T])^{3/2}$, *for a constant c and $m \geq c\sqrt{K}/\epsilon^2$.*

Remark The test T of Acharya, Daskalakis, and Kamath (2015) is given by $T = \Sigma_{i=1}^{n} \frac{(N_i - ms_i)^2 - N_i}{ms_i}$. Their paper gives the upper bound $\mathrm{Var}[T] \leq 4n + 9\sqrt{n}\mathrm{E}[T] + \frac{2}{5}n^{\frac{1}{4}}\mathrm{E}[T]^{3/2}$ under the assumption $s_i \geq \epsilon/50n$ for every i. In our application, ℓ is the alphabet size and we will need the bound to depend on ℓ. In addition, we also need the bounds to work when $s_i \geq \epsilon^2/50n\ell$. Both these can be achieved by modifying their proof.

It remains to sample numbers $N_{ij} \sim \mathrm{Poi}(mr_{ij})$ independently for every i, j. We do this via poissonization followed by sampling from each coordinate of the product distribution R independently. We present Algorithm 1 with its correctness.

Proof (*of Theorem 2.1*) Let $\ell = |\Sigma|$. First, we transform the distributions P and Q into the distributions R and S respectively according to the modification process mentioned in Lemma 7. This gives:

- Each sample from R can be simulated by 1 sample from P.
- R and S are product distributions, $R = \prod_{i=1}^{n} R_i$ and $S = \prod_{i=1}^{n} S_i$, where $R_i = \langle r_{i1}, r_{i2}, \ldots, r_{i\ell} \rangle$ and $S_i = \langle s_{i1}, s_{i2}, \ldots, s_{i\ell} \rangle$ for every i.
- For every i, j, $s_{ij} \geq \epsilon^2/50n\ell$.
- If $d_{\chi^2}(P, Q) \leq \epsilon^2/9$ then $\sum_{i,j} \frac{(r_{ij}-s_{ij})^2}{s_{ij}} < 0.12\epsilon^2$.
- If $\sqrt{2}d_{\mathrm{H}}(P, Q) > \epsilon$ then $\sum_{i,j} \frac{(r_{ij}-s_{ij})^2}{s_{ij}} > 0.18\epsilon^2$.

Algorithm 1: Given samples from an unknown distribution $R = \prod_{i=1}^{n} R_i$ and a known distribution $S = \prod_{i=1}^{n} S_i$ over Σ^n, decide $d_{\chi^2}(R, S) \leq \epsilon^2/9$ ('yes') versus $\sqrt{2} d_{\mathrm{H}}(R, S) > \epsilon$ ('no'). Let $\ell = |\Sigma|$, $R_i = \langle r_{i1}, r_{i2}, \ldots, r_{i\ell} \rangle$, $S_i = \langle s_{i1}, s_{i2}, \ldots, s_{i\ell} \rangle$ with $s_{ij} \geq \epsilon^2/50n\ell$ for every j, for every i

```
for i = 1 to n do
    Sample N_i ~ Poi(m) independently;
end
N = max_i N_i;
X ← Take N samples from R;
for i = 1 to n do
    X_i ← Sequence of symbols in the i-th coordinate of first N_i samples of X;
    ⟨N_i1, N_i1, ..., N_iℓ⟩ ← histogram of symbols in X_i;
end
Compute statistic T of Theorem 3.1 using N_ij and s_ij values for every i, j;
if T ≤ 0.15mε² then
    output 'yes.';
else
    output 'no.'
end
```

Henceforth, we focus on distinguishing $\sum_{i,j} \frac{(r_{ij}-s_{ij})^2}{s_{ij}} < 0.12\epsilon^2$ versus $> 0.18\epsilon^2$, under the assumption $s_{ij} \geq \epsilon^2/50n\ell$ for every i, j, by sampling from R. We use the tester T of Acharya, Daskalakis, and Kamath (2015) stated in Theorem 3.1 with $K = n\ell$, for this. Firstly, note that in Algorithm 1, the samples S_i is a set of $N_i \sim \mathrm{Poi}(m)$ samples from R_i, independently for every i's. This is because the set of samples are taken from the product distribution $R = R_1 \times R_2 \times \cdots \times R_n$ and the N_i values are independent for different i's. Due to Poissonization it follows $N_{ij} \sim \mathrm{Poi}(r_{ij})$ independently for every i, j. The tester T requires $m \geq c\sqrt{n\ell}/\epsilon^2$, for some constant c and satisfies $\mathrm{E}[T] = m \sum_{i,j} \frac{(r_{ij}-s_{ij})^2}{s_{ij}}$, $\mathrm{Var}[T] \leq 2n\ell + 7\sqrt{n\ell}\mathrm{E}[T] + 4(n\ell)^{1/4}(\mathrm{E}[T])^{3/2}$.

If $\sum_{i,j} \frac{(r_{ij}-s_{ij})^2}{s_{ij}} < 0.12\epsilon^2$ then we get $\mathrm{E}[T] \leq 0.12m\epsilon^2$ and $\mathrm{Var}[T] \leq \left(\frac{2}{c^2} + \frac{0.84}{c} + \frac{4(0.12)^3}{\sqrt{c}}\right) m^2\epsilon^4$, using $m \geq c\sqrt{nl}/\epsilon^2$ and the upper bound for $\mathrm{E}[T]$. By Chebyshev's inequality $T < 0.15m\epsilon^2$ with probability at least 4/5, where $c = \Omega(1)$ is an appropriate constant.

If $\sum_{i,j} \frac{(r_{ij}-s_{ij})^2}{s_{ij}} > 0.18\epsilon^2$ then we get $\mathrm{E}[T] > 0.18m\epsilon^2 \geq 0.18c\sqrt{nl}$ and $\mathrm{Var}[T] \leq (\frac{2}{(0.18c)^2} + \frac{7}{0.18c} + \frac{4}{\sqrt{0.18c}})\mathrm{E}^2[T]$. By Chebyshev's inequality $T > 0.15m\epsilon^2$ with probability at least 4/5, for an appropriate constant $c = \Omega(1)$.

Hence, for some constant c', $m \geq c'\sqrt{n\ell}/\epsilon^2$ suffices for the tester T to distinguish the above two cases. It also follows from the concentration of the Poisson distribution that the number of samples required is $\max_i N_i \leq 2m$, except for probability at most $n \cdot \exp(-m) < 1/10$ (using union bound).

The histograms can be computed by a single pass over the n-dimensional sample set S. The statistic T can be computed in time $O(n\ell)$. So the time complexity is $O(n\ell + n^{3/2}\sqrt{\ell}/\epsilon^2)$. ∎

3.2. d_{H}-vs-d_{H} Tolerant Closeness Tester

In this section, we give a tester for distinguishing $d_{\mathrm{H}}(P, Q) \leq \epsilon$ versus $d_{\mathrm{H}}(P, Q) > 3\epsilon$ for two unknown product distributions P and Q over support Σ^n. To get Theorem 2.2, we rescale ϵ down to $\epsilon/\sqrt{2}$. We take a testing-by-learning approach: we first learn P and Q in Hellinger distance $\epsilon/2$ using the following known result. Then the Hellinger distance between the learnt distributions can be computed exactly.

Theorem 3.2 *(Acharya, Daskalakis, and Kamath, 2015) Given samples from an unknown product distribution D over Σ^n, $\hat{D}$, the product of component-wise empirical distributions on m samples satisfy $d_{\mathrm{H}}(D, \hat{D}) \leq \epsilon$ with $9/10$ probability if $m \geq \Theta(n|\Sigma|/\epsilon^2)$.*

Proof (*of Theorem 2.2*) We first learn P and Q as $\hat{P}$ and $\hat{Q}$ using Theorem 3.2 such that $d_{\mathrm{H}}(P, \hat{P}) \leq \epsilon/2$ and $d_{\mathrm{H}}(Q, \hat{Q}) \leq \epsilon/2$, together with 4/5 probability. Conditioned on this. we compute $d_{\mathrm{H}}(\hat{P}, \hat{Q})$ exactly using Fact 4.

Due to triangle inequality, $d_{\mathrm{H}}(\hat{P}, \hat{Q}) \leq 2\epsilon$ or not would decide $d_{\mathrm{H}}(P, Q) \leq \epsilon$ or $> 3\epsilon$. ■

4. Lower Bounds

In this section, we give lower bounds for tolerant testing of product distributions. Our lower bounds use a reduction from testing the class of unstructured distributions over n items to testing the class of product distributions over $\{0, 1\}^n$, given by Canonne, Diakonikolas, Kane, and Stewart (2017) (Section 4.5 of their paper). However, in order to apply this reduction, we need to establish certain new bounds relating the distances in the unstructured setting to the setting of product distribution. We first define how to construct a product distribution from the corresponding unstructured distribution. In particular, for a $\delta < 1$, this construction produces a product distribution $F_\delta(P)$ over $\{0, 1\}^n$ from a given distribution P over n symbols.

Definition 8 (Construction of $F_\delta(P)$) *Let P be a distribution over a sample space of n items and $0 < \delta \leq 1$ be a constant. Let S be a random set of $\mathrm{Poi}(\delta)$ samples from P. For every item $i \in [n]$, let x_i be the indicator variable such that $x_i = 1$ iff i appears in S. Let $F_\delta(P)$ be the joint distribution of $\langle x_1, x_2, \ldots, x_n \rangle$ over the sample space $\{0, 1\}^n$.*

The following property can be observed using the property of Poissonization.

Fact 9 *Let P be a distribution over a sample space of n items with probability vector $\langle p_1, p_2, \ldots, p_n \rangle$ and $0 < \delta \leq 1$ be a constant. Then $F_\delta(P)$ is a product distribution such that $F_\delta(P) = \prod_{i=1}^n F_\delta(P_i)$ where $F_\delta(P_i) \sim \mathrm{Bern}(1 - e^{-\delta p_i})$.*

We use the following crucial lemma.

Lemma 10 *For any $0 < \delta \leq 1$ and distributions P, Q, $d_{\mathrm{TV}}(F_\delta(P), F_\delta(Q)) \geq \delta e^{-\delta} d_{\mathrm{TV}}(P, Q)$, with equality holding iff $P = Q$.*

Proof Let $P = \langle p_1, \dots, p_i, \dots, p_n \rangle$ and $Q = \langle q_1, \dots, q_i, \dots, q_n \rangle$ be the probability values of P and Q.

$$\begin{aligned} d_{\mathrm{TV}}(F_\delta(P), F_\delta(Q)) &= \sum_{x \in \{0,1\}^n} |F_\delta(P)(x) - F_\delta(Q)(x)| \\ &\geq \sum_{i=1}^{n} |F_\delta(P)(e_i) - F_\delta(Q)(e_i)| && \text{(unit vector } e_i \text{ has } i\text{-th value 1)} \\ &= \sum_{i=1}^{n} |(1 - e^{-\delta p_i})\Pi_{j \neq i} e^{-\delta p_j} - (1 - e^{-\delta q_i})\Pi_{j \neq i} e^{-\delta q_j}| \\ &= \sum_{i=1}^{n} e^{-\delta} |e^{\delta p_i} - e^{\delta q_i}| && \text{(Since } \Pi_j e^{-\delta p_j} = \Pi_j e^{-\delta q_j} = e^{-\delta}) \\ &= e^{-\delta} \sum_{i=1}^{n} |\delta(p_i - q_i) + \delta^2(p_i^2 - q_i^2)/2! + \dots + \delta^j(p_i^j - q_i^j)/j! + \dots|. \end{aligned}$$

We analyze the expression under modulus under two cases: 1) if $p_i > q_i$, it is more than $\delta(p_i - q_i)$, 2) if $p_i < q_i$, it is more than $\delta(q_i - p_i)$.

$$\begin{aligned} d_{\mathrm{TV}}(F_\delta(P), F_\delta(Q)) &\geq e^{-\delta} \sum_{i=1}^{n} |\delta(p_i - q_i)| \\ &= \delta e^{-\delta} d_{\mathrm{TV}}(P, Q). \end{aligned}$$

■

4.1. Hardness of d_{χ^2}-vs-d_{TV} Tolerant Closeness Testing

Here we show that for two unknown product distribution P, Q over $\{0,1\}^n$, distinguishing $d_{\chi^2}(P, Q) \leq \epsilon^2/9$ versus $d_{\mathrm{TV}}(P, Q) > \epsilon$, for a constant ϵ, can not be decided in general with a truly sublinear sample complexity. We use a reduction to the following difficult problem, for hardness of χ^2-tolerance for closeness testing of unstructured distributions over n items, given in Daskalakis, Kamath, and Wright (2018). We restate the theorem with changes in the constants.

Theorem 4.1 *There exists a constant $0 < \epsilon < 1$ and three distributions P^{yes}, P^{no} and Q, each over the sample space $[n]$ such that: (1) $d_{\chi^2}(P^{yes}, Q) \leq \epsilon^2/216$, whereas $d_{\mathrm{TV}}(P^{no}, Q) \geq \epsilon$ and (2) given only sample accesses to one of P^{yes} or P^{no}, and Q, distinguishing P^{yes} versus P^{no} with probability $> 4/5$, requires $\Omega(n/\log n)$ samples.*

We use the following important property about the χ^2-distance between the reduced distributions.

Lemma 11 $d_{\chi^2}(F_\delta(P), F_\delta(Q)) \leq \exp(4\delta \cdot \chi^2(P, Q)) - 1$, *for any* $0 < \delta \leq 1$.

Proof From Fact 9, both $F_\delta(P)$ and $F_\delta(Q)$ are product distributions, the distribution of the i-th component being $F_\delta(P_i)$ and $F_\delta(Q_i)$ respectively. Let $P = \langle p_1, p_2, \dots, p_n \rangle$ and $Q = \langle q_1, q_2, \dots, q_n \rangle$.

Then $F_\delta(P_i) \sim \mathrm{Bern}(1-e^{-\delta p_i})$ and $F_\delta(Q_i) \sim \mathrm{Bern}(1-e^{-\delta q_i})$.

$$\begin{aligned}
d_{\chi^2}(F_\delta(P), F_\delta(Q)) &= \prod_i (1 + d_{\chi^2}(F_\delta(P_i), F_\delta(Q_i))) - 1 && \text{(From Lemma 3)}\\
&\leq \prod_i \exp(d_{\chi^2}(F_\delta(P_i), F_\delta(Q_i))) - 1 && \text{(Since } e^x \geq (1+x) \text{ for } x \geq 0)\\
&= \exp(\sum_i d_{\chi^2}(F_\delta(P_i), F_\delta(Q_i))) - 1\\
&= \exp\left(\sum_i (e^{-\delta p_i} - e^{-\delta q_i})^2 \left(\frac{1}{e^{-\delta q_i}} + \frac{1}{1-e^{-\delta q_i}}\right)\right) - 1\\
&\qquad \text{(Since } F_\delta(P_i^{yes}) \sim \mathrm{Bern}(1-e^{-\delta p_i}) \text{ and } F_\delta(Q_i) \sim \mathrm{Bern}(1-e^{-\delta q_i}))\\
&= \exp(\sum_i (e^{-\delta p_i} - e^{-\delta q_i})^2 / e^{-\delta q_i}(1-e^{-\delta q_i})) - 1\\
&= \exp(\sum_i (e^{\delta(q_i - p_i)} - 1)^2 / (e^{\delta q_i} - 1)) - 1\\
&= \exp(\sum_i (e^{\delta(q_i - p_i)} - 1)^2 / (e^{\delta q_i} - 1)) - 1\\
&\leq \exp(\sum_i (2\delta(p_i - q_i))^2 / \delta q_i)) - 1\\
&\qquad \text{(Since } (e^x - 1) \geq x \text{ and } (|e^x - 1| \leq 2|x| \text{ for } 0 < |x| < 1)\\
&= \exp(4\delta\chi^2(P, Q)) - 1.
\end{aligned}$$

■

We are set to present the main lower bound result of this section.

Theorem 2.5 (d_{χ^2}-versus-d_{TV} closeness testing lower bound) *There exists a constant $0 < \epsilon < 1$ and three product distributions F^{yes}, F^{no} and F, each over the sample space $\{0,1\}^n$ such that $d_{\chi^2}(F^{yes}, F) \leq \epsilon^2/9$, whereas $d_{\mathrm{TV}}(F^{no}, F) > \epsilon$, and given only sample accesses to F^{yes}, F^{no} and F, distinguishing F^{yes} versus F^{no} with probability $> 2/3$, requires $\Omega(n/\log n)$ samples.*

Proof We start with the hard distributions P^{yes}, P^{no} and Q from Theorem 4.1. Then $d_{\chi^2}(P^{yes}, Q) \leq \epsilon^2/216$ and $d_{\mathrm{TV}}(P^{no}, Q) \geq \epsilon$ for some constant $0 < \epsilon < 1$. We apply the reduction of Definition 8 with $\delta = 1/3$ to these three distributions. Then from Lemma 10 and Lemma 11 we get the following two inequalities:

- $d_{\chi^2}(F_\delta(P^{yes}), F_\delta(Q)) \leq \exp(4\chi^2(P^{yes}, Q)/3) - 1 < \epsilon^2/160$.
- $d_{\mathrm{TV}}(F_\delta(P^{no}), F_\delta(Q)) > (1/3e^{1/3})\epsilon$.

It follows if we can distinguish $d_{\chi^2}(F_\delta(P^{yes}), F_\delta(Q)) \leq \epsilon^2/160$ versus $d_{\mathrm{TV}}(F_\delta(P^{no}), F_\delta(Q)) > (1/3e^{1/3})\epsilon$, then we are able to decide the hard instance of Theorem 4.1. Moreover, in order to simulate each sample from the distribution $F_{1/2}(P)$, we need $\mathrm{Poi}(1/2)$ samples from P. So, if we need m samples in total, from the additive property of the Poisson distribution, we need

$\mathrm{Poi}(m/2) = O(m)$ samples from P in total, except for $\exp(-m)$ probability. It follows, if we can decide the problem given in the theorem statement in $o(n/\log n)$ samples, we can decide the hard problem of Theorem 4.1 in $o(n/\log n)$ samples as well. This leads to a contradiction. Replacing the constant ϵ by $3e^{1/3}\epsilon_1$, we get Theorem 2.5. ■

4.2. Hardness of d_{KL}-vs-d_{TV} Tolerant Identity Testing

In this section we show that for an unknown product distribution P and a known product distribution Q over $\{0,1\}^n$, distinguishing $d_{\mathrm{KL}}(P,Q) \le \epsilon^2/9$ versus $d_{\mathrm{TV}}(P,Q) > \epsilon$, for a constant ϵ, cannot be decided in general with a truly sublinear sample complexity. We use a reduction to the following hardness result, for identity testing of unstructured distributions over n items under KL-tolerance, given in Daskalakis, Kamath, and Wright (2018). We restate the theorem with changes in the constants. For a probability distribution $P = \langle p_1, p_2, \ldots, p_n \rangle$ over n items, $||P||_2^2 = \sum_i p_i^2$.

Theorem 4.2 *There exists a constant $0 < \epsilon < 1$ and three distributions P^{yes}, P^{no} and Q, each over the sample space $[n]$ such that: (1) $d_{\mathrm{KL}}(P^{yes}, Q) \le \epsilon^2/216$, whereas $d_{\mathrm{TV}}(P^{no}, Q) \ge \epsilon$, (2) $||P^{yes}||_2^2 = O(\log^2 n/n)$, and (3) given only sample accesses to one of P^{yes} or P^{no}, and complete knowledge of Q, distinguishing P^{yes} versus P^{no} with probability $> 4/5$, requires $\Omega(n/\log n)$ samples.*

Proof The proof of this Theorem appears in Daskalakis, Kamath, and Wright (2018) (in Theorem 6.2 of this version), except the fact $||P^{yes}||_2^2 = O(\log^2 n/n)$ is not explicitly claimed. We prove this claim in the following, by observing from the original construction given in the paper by Valiant and Valiant (2010).

The hard distribution P^{yes} is the distribution $p^-_{\log k,\phi}$ as defined in Definition 12 of Valiant and Valiant (2010). We use the following facts about this distribution $p^-_{\log k,\phi}$, given in Fact 11, Definition 12 and (in the end of the second paragraph in the proof of) Lemma 13 in Valiant and Valiant (2010):

- ϕ is a small enough constant
- The support size n and the parameter k are related as $n = 32k \log k/\phi$
- The 'un-normalized' mass at each point is $x/32k$, where $j = \log k$ and $x \le 4j$
- The 'normalizing constant' c_2 (which makes the probability values sum up to 1) is at most ϕ/j where $j = \log k$

From these facts we conclude each probability mass is $c_2 \cdot x/32k \le \phi/8k$, where $n = 32k \log k/\phi$ for some constant ϕ. Hence, $||P^{yes}||_2^2 \le \phi^2/64k^2 \cdot 32k \log k/\phi = \phi \log k/2k = O(\log^2 n/n)$. ■

We use the reduction given in Definition 8. We establish the following lemma, relating KL distances between the original and the reduced distributions.

Lemma 12 $d_{\mathrm{KL}}(F_\delta(P), F_\delta(Q)) \le \left(\delta + \frac{\delta^2}{2}\right) d_{\mathrm{KL}}(P,Q) + \frac{3\delta^2}{2}||P||_2^2$, *for any* $0 < \delta \le 1$.

Proof We use the following fact about the KL-distance between two product distributions.

Fact 13 *For two distributions $P = \prod_{i=1}^n P_i$ and $Q = \prod_{i=1}^n Q_i$ over the same discrete sample space, it holds that $d_{\mathrm{KL}}(P, Q) = \sum_{i=1}^n d_{\mathrm{KL}}(P_i, Q_i)$.*

From Fact 9, both $F_\delta(P)$ and $F_\delta(Q)$ are product distributions, the distribution of the i-th component being $F_\delta(P_i)$ and $F_\delta(Q_i)$ respectively. Let $P = \langle p_1, p_2, \ldots, p_n\rangle$ and $Q = \langle q_1, q_2, \ldots, q_n\rangle$. Then $F_\delta(P_i) \sim \mathrm{Bern}(1 - e^{-\delta p_i})$ and $F_\delta(Q_i) \sim \mathrm{Bern}(1 - e^{-\delta q_i})$.

$$
\begin{aligned}
&d_{\mathrm{KL}}(F_\delta(P), F_\delta(Q)) \\
&\quad = \sum_i d_{\mathrm{KL}}(F_\delta(P_i), F_\delta(Q_i)) \\
&\quad = \sum_i \left[(1 - e^{-\delta p_i}) \ln\left(\frac{1 - e^{-\delta p_i}}{1 - e^{-\delta q_i}}\right) + e^{-\delta p_i} \ln \frac{e^{-\delta p_i}}{e^{-\delta q_i}}\right] \\
&\quad = \sum_i \ln\left(\frac{1 - e^{-\delta p_i}}{1 - e^{-\delta q_i}}\right) + \sum_i e^{-\delta p_i} \ln \frac{e^{-\delta p_i}(1 - e^{-\delta q_i})}{e^{-\delta q_i}(1 - e^{-\delta p_i})} \\
&\quad = \sum_i \ln \frac{e^{\delta q_i}(e^{\delta p_i} - 1)}{e^{\delta p_i}(e^{\delta q_i} - 1)} + \sum_i e^{-\delta p_i} \ln\left(\frac{e^{\delta q_i} - 1}{e^{\delta p_i} - 1}\right) \\
&\quad = \sum_i \ln \frac{e^{\delta q_i}}{e^{\delta p_i}} + \sum_i \ln\left(\frac{e^{\delta p_i} - 1}{e^{\delta q_i} - 1}\right) + \sum_i e^{-\delta p_i} \ln\left(\frac{e^{\delta q_i} - 1}{e^{\delta p_i} - 1}\right) \\
&\quad = \sum_i (q_i - p_i) + \sum_i (1 - e^{-\delta p_i}) \ln\left(\frac{e^{\delta p_i} - 1}{e^{\delta q_i} - 1}\right) \\
&\quad = \sum_i (1 - e^{-\delta p_i}) \ln\left(\frac{e^{\delta p_i} - 1}{e^{\delta q_i} - 1}\right) \qquad\qquad \text{(Since } \textstyle\sum_i p_i = \sum_i q_i = 1) \\
&\quad = \sum_{p_i > q_i} (1 - e^{-\delta p_i}) \ln\left(\frac{e^{\delta p_i} - 1}{e^{\delta q_i} - 1}\right) + \sum_{q_i > p_i} (1 - e^{-\delta p_i}) \ln\left(\frac{e^{\delta p_i} - 1}{e^{\delta q_i} - 1}\right) \\
&\quad \le \sum_{p_i > q_i} \delta p_i \ln\left(\frac{e^{\delta p_i} - 1}{e^{\delta q_i} - 1}\right) + \sum_{q_i > p_i} \left(\delta p_i - \frac{1}{2}\delta^2 p_i^2\right) \ln\left(\frac{e^{\delta p_i} - 1}{e^{\delta q_i} - 1}\right) \\
&\quad = \sum_i \delta p_i \ln\left(\frac{e^{\delta p_i} - 1}{e^{\delta q_i} - 1}\right) + \sum_{q_i > p_i} \frac{\delta^2 p_i^2}{2} \ln\left(\frac{e^{\delta q_i} - 1}{e^{\delta p_i} - 1}\right) \\
&\quad \le \sum_i \delta p_i \ln\left(\frac{\delta p_i(1 + \delta p_i)}{\delta q_i}\right) + \sum_{q_i > p_i} \frac{\delta^2 p_i^2}{2} \ln\left(\frac{\delta q_i(1 + \delta q_i)}{\delta p_i}\right) \\
&\quad = \delta\left(\sum_i p_i \ln \frac{p_i}{q_i} + \sum_i p_i \ln(1 + \delta p_i)\right) + \frac{\delta^2}{2}\left(\sum_{q_i > p_i} p_i^2 \ln \frac{q_i}{p_i} + \sum_{q_i > p_i} p_i^2 \ln(1 + \delta q_i)\right) \\
&\quad \le \left(\delta + \frac{\delta^2}{2}\right) \sum_i p_i \ln \frac{q_i}{p_i} + \frac{3\delta^2}{2} \sum_i p_i^2 \\
&\quad = \left(\delta + \frac{\delta^2}{2}\right) d_{\mathrm{KL}}(P, Q) + \frac{3\delta^2}{2} ||P||_2^2.
\end{aligned}
$$

■

Now we present the lower bound for closeness testing of product distributions.

Theorem 2.4 (d_{KL}-versus-d_{TV} identity testing lower bound) *There exists a constant* $0 < \epsilon < 1$ *and three product distributions* F^{yes}, F^{no} *and* F*, each over the sample space* $\{0,1\}^n$ *such that* $d_{\mathrm{KL}}(F^{yes}, F) \leq \epsilon^2/9$*, whereas* $d_{\mathrm{TV}}(F^{no}, F) > \epsilon$*, and given only sample accesses to* F^{yes}, F^{no}*, and complete knowledge about* F*, distinguishing* F^{yes} *versus* F^{no} *with probability* $> 2/3$*, requires* $\Omega(n/\log n)$ *samples.*

Proof We start with the distributions P^{yes}, P^{no} and Q from the hardness result Theorem 4.2. Then $d_{\mathrm{KL}}(P^{yes}, Q) \leq \epsilon^2/216$, $||P^{yes}||_2^2 = O(\log^2 n/n)$ and $d_{\mathrm{TV}}(P^{no}, Q) \geq \epsilon$ for some constant $0 < \epsilon < 1$. We apply the reduction of Definition 8, with $\delta = 1/3$ to these three distributions. Then from Lemma 10 and Lemma 12 we get the following two:

- $d_{\mathrm{KL}}(F_\delta(P^{yes}), F_\delta(Q)) \leq \epsilon^2/160$, for any large enough n.
- $d_{\mathrm{TV}}(F_\delta(P^{no}), F_\delta(Q)) > (1/3e^{1/3})\epsilon$.

It follows if we can distinguish $d_{\mathrm{KL}}(F_\delta(P^{yes}), F_\delta(Q)) \leq \epsilon^2/160$ versus $d_{\mathrm{TV}}(F_\delta(P^{no}), F_\delta(Q)) > (1/3e^{1/3})\epsilon$, then we will be able to decide the hard instance of Theorem 4.1. Moreover, in order to simulate each sample from the distribution $F_{1/2}(P)$, we need $\mathrm{Poi}(1/2)$ samples from P. So, if we need m samples in total, from the additive property of the Poisson distribution, we need $\mathrm{Poi}(m/2) = O(m)$ samples from P in total, except for $\exp(-m)$ probability. It follows, if we can decide the problem given in the Theorem statement in $o(n/\log n)$ samples, we can decide the hard problem of Theorem 4.2 in $o(n/\log n)$ samples as well. This leads to a contradiction. Replacing the constant ϵ by $3e^{1/3}\epsilon_1$, we get Theorem 2.4. ■

Before moving on to the next section, we note that recently the question of d_{TV}-versus-d_{TV} tolerant testing problem for uniformity testing of distributions over $[n]$ was settled to be $\Theta(\frac{n}{\log n}\frac{1}{\epsilon^2})$ by Jiao, Han, and Weissman (2018). In particular, this gives a stronger guarantee for Theorem 4.1 and Theorem 4.2 when ϵ is not a constant. This directly strengthens our Theorem 2.5; also Theorem 2.4 whenever $||P_{yes}||_2^2 = O(\epsilon^2)$, giving us a $\Omega(\frac{n}{\log n}\frac{1}{\epsilon^2})$ lower bound for any ϵ.

4.3. Hardness of non-tolerant d_{TV} Identity Testing for General Alphabets

Daskalakis, Dikkala, and Kamath (2019) and Canonne, Diakonikolas, Kane, and Stewart (2017) have given optimal lower bounds for non-tolerant testing w.r.t. d_{TV} distance when $|\Sigma| = 2$. In this section, we generalize their result for $\Sigma > 2$ case and get an optimal lower bound in regard to Theorem 2.1. We show the following theorem, generalizing the proof of Daskalakis, Dikkala, and Kamath (2019) specifically.

Theorem 2.3 *Uniformity testing with w.r.t.* d_{TV} *distance for product distributions over* $[\ell]^n$ *needs* $\Omega(\sqrt{n\ell}/\epsilon^2)$ *samples.*

Our hard distributions are as follows:
$P =$ the uniform distribution over $[\ell]^n$.

$Q =$ a radom distribution from the mixture $\left\{\left\{\frac{1}{\ell}\left(1 \pm \frac{\epsilon}{\sqrt{n}}\right)\right\}^{\frac{\ell}{2}}\right\}^n$. Each distribution of the mixture is a product distribution, whose i-th component is a distribution over $[\ell]$, which randomly assigns probability values either $\frac{1}{\ell}\left(1+\frac{\epsilon}{\sqrt{n}}\right), \frac{1}{\ell}\left(1-\frac{\epsilon}{\sqrt{n}}\right)$ or $\frac{1}{\ell}\left(1-\frac{\epsilon}{\sqrt{n}}\right), \frac{1}{\ell}\left(1+\frac{\epsilon}{\sqrt{n}}\right)$, based on a random vector from $\{0,1\}^{\frac{\ell}{2}}$, for every consecutive sample space items from $[\ell]$.

First claim we show is that each member of the mixture is $\Theta(\epsilon)$ far from P in d_{TV} distance.

Claim 14 *Let Q^* be any member of Q. Then $d_{\mathrm{TV}}(P, Q^*) \geq \Theta(\epsilon)$.*

Proof Note that all members of the mixture Q are permutations of each other. Since P is fixed to the uniform distribution, all of them have the same d_{TV} to P. We fix Q^* to be the distribution from Q, corresponding to $\{0,1\}^{\ell/2}$ at every component.

It is a known fact that applying a common function to the sample space items can only reduce d_{TV}. We apply the function which is parity of $x \in [\ell]$ component wise. Resulting sample space becomes $\{0,1\}^n$, Q^* becomes Bern $\left(1+\frac{\epsilon}{\sqrt{n}}\right)^n$ and P becomes Bern $\left(\frac{1}{2}\right)^n$. It is a standard fact that the d_{TV} of the later pair is at least $\Theta(\epsilon)$ (see eg. Canonne, Diakonikolas, Kane, and Stewart (2017)). ■

Next we show that distingishing k samples from P and Q is hard. Let $P^{\otimes k}, Q^{\otimes k}$ be their distributions. Noting that the components of both the mixture and the uniform distribution are independent and symmetric. We can upper bound them by n copies of the first component's distribution, using Pinsker's inequality and linearity of KL.

$$d_{\mathrm{TV}}^2(P^{\otimes k}, Q^{\otimes k}) \lesssim d_{\mathrm{KL}}(Q^{\otimes k}, P^{\otimes k}) \leq n \cdot d_{\mathrm{KL}}(Q_1^{\otimes k}, P_1^{\otimes k}) \tag{3}$$

Computing d_{H}^2 or d_{KL} are hard for the multinomial unlike Daskalakis, Dikkala, and Kamath (2019) (cf. Lemma 17). So, we use a reduction to simplify the calculations.

Recall $P_1^{\otimes k}$ is the distribution of the first k samples when P is the uniform distribution over $[\ell]$ and $Q_1^{\otimes k}$ is the same when we take a random distribution from the $2^{\frac{\ell}{2}}$ size mixture.

We reduce P_1 to the distribution $f(P)_1 = \mathrm{Bern}(\frac{1}{\ell})^l$ and any distribution from the mixture $Q_1^* = \langle q_1, \ldots, q_\ell \rangle$ to $f(Q)_1^* = \mathrm{Bern}(q_1) \times \cdots \times \mathrm{Bern}(q_\ell)$. We claim that this reduction changes KL by a constant factor, for any particular (randomly) chosen pair P_1, Q_1^* to start with.

Claim 15 $d_{\mathrm{H}}(Q_1^*, P_1) \leq d_{\mathrm{KL}}(f(Q)_1^*, f(P)_1)$ *and* $d_{\mathrm{KL}}({Q_1}^{\otimes k}, P_1^{\otimes k}) \leq 6 d_{\mathrm{KL}}({f(Q)_1}^{\otimes k}, {f(P)_1}^{\otimes k})$.

Proof

$$\begin{aligned}
d_{\mathrm{KL}}(Q_1, P_1) &= \sum_j Q_{ij} \log \frac{Q_{1j}}{P_{1j}} \\
&= \sum_{j:Q_{1j} > \frac{1}{\ell}} \frac{1}{\ell}\left(1+\frac{\epsilon}{\sqrt{n}}\right) \log\left(1+\frac{\epsilon}{\sqrt{n}}\right) + \sum_{j:Q_{1j} < \frac{1}{\ell}} \frac{1}{\ell}\left(1-\frac{\epsilon}{\sqrt{n}}\right) \log\left(1-\frac{\epsilon}{\sqrt{n}}\right) \\
&\leq \frac{1}{2}\left(1+\frac{\epsilon}{\sqrt{n}}\right)\frac{\epsilon}{\sqrt{n}} + \frac{1}{2}\left(1-\frac{\epsilon}{\sqrt{n}}\right)\left(-\frac{\epsilon}{\sqrt{n}}\right) \\
&= \frac{\epsilon^2}{n}
\end{aligned}$$

$$d_{\mathrm{KL}}(f(Q)_1, f(P)_1) = \frac{\ell}{2} d_{\mathrm{KL}}\left(\mathrm{Bern}\left(\frac{1}{\ell}\left(1+\frac{\epsilon}{\sqrt{n}}\right), \mathrm{Bern}\left(\frac{1}{\ell}\right)\right)\right) + \frac{\ell}{2} d_{\mathrm{KL}}\left(\mathrm{Bern}\left(\frac{1}{\ell}\left(1-\frac{\epsilon}{\sqrt{n}}\right)\right), \mathrm{Bern}\left(\frac{1}{\ell}\right)\right)$$

$$\begin{aligned}
& d_{\mathrm{KL}}\left(\mathrm{Bern}\left(\frac{1}{\ell}\left(1+\frac{\epsilon}{\sqrt{n}}\right), \mathrm{Bern}\left(\frac{1}{\ell}\right)\right)\right) \\
&= \frac{1}{\ell}\left(1+\frac{\epsilon}{\sqrt{n}}\right)\log(1+\frac{\epsilon}{\sqrt{n}}) + \left(1-\frac{1}{\ell}-\frac{\epsilon}{\ell\sqrt{n}}\right)\log\left(1-\frac{\epsilon}{(\ell-1)\sqrt{n}}\right) \\
&\geq \frac{1}{\ell}\left(1+\frac{\epsilon}{\sqrt{n}}\right)\left(\frac{\epsilon}{\sqrt{n}}-\frac{\epsilon^2}{2n}\right) + \left(1-\frac{1}{\ell}-\frac{\epsilon}{\ell\sqrt{n}}\right)\left(-\frac{\epsilon}{\sqrt{n}(\ell-1)}-\frac{\epsilon^2}{n(\ell-1)^2}\right) \\
&\geq \frac{\epsilon^2}{3n\ell}
\end{aligned}$$

Therefore, $d_{\mathrm{KL}}(f(Q)_1, f(P)_1) \geq d_{\mathrm{KL}}(Q_1, P_1)/6$. Using linearity of KL, we get $d_{\mathrm{H}}(Q_1^{*\otimes k}, P_1^{\otimes k}) \leq 6 d_{\mathrm{KL}}(f(Q)_1^{*\otimes k}, f(P)_1^{\otimes k})$. Since this holds for the reduction on any chosen starting pair, we get that in general for the mixture,

$$d_{\mathrm{KL}}({Q_1}^{\otimes k}, P_1^{\otimes k}) \leq 6 d_{\mathrm{KL}}({f(Q)_1}^{\otimes k}, {f(P)_1}^{\otimes k})$$

■

Henceforth we focus on upper bounding $d_{\mathrm{KL}}(f(Q)_1^{\otimes k}, f(P)_1^{\otimes k})$. The reduction makes it a Boolean product distribution. Daskalakis, Dikkala, and Kamath (2019) gave such upper bounds when every component is randomly mixed and when the probabilities are close to $1/2$. Instead we need to mix every pairs of components and need to make the probabilities close to $1/\ell$.

Let $p_+ = \frac{1}{\ell}\left(1+\frac{\epsilon}{\sqrt{n}}\right)$ and $p_- = \frac{1}{\ell}\left(1-\frac{\epsilon}{\sqrt{n}}\right)$.

Proof (of Theorem 2.3) We firstly note that it suffices to upper bound the joint distribution of the count of 1's in the samples $f(Q)_1^{\otimes k}$ and $f(P)_1^{\otimes k}$ (Daskalakis, Dikkala, and Kamath (2019), Lemma 17). By symmetry, we can focus on the mixture on the first two components $1, 2$. Recall these are actually the first two of the ℓ components of the reduced distribution (where the reduction was performed on the first component of the original distribution). Note that Daskalakis, Dikkala, and Kamath (2019) could instead focus on a single component.

$$d_{\mathrm{KL}}(Q_1^{\otimes k}, P_1^{\otimes k}) \lesssim d_{\mathrm{KL}}(f(Q)_1^{\otimes k}, f(P)_1^{\otimes k}) = \frac{\ell}{2} d_{\mathrm{KL}}({f(Q)_1^{\otimes k}}_{12}, {f(P)_1^{\otimes k}}_{12}) \leq \frac{\ell}{2} d_{\chi^2}(R_{12}^{\otimes k}, S_{12}^{\otimes k}) \tag{4}$$

We use $R = f(Q)_1$ and $S = f(P)_1$ for simplicity. Then, $R_{12}^{\otimes k}$ and $S_{12}^{\otimes k}$ denote the joint distribution of the count of 1s in k samples at the first 2 components w.r.t the distributions $\mathrm{Bern}\left(\frac{1}{\ell}\right)$ and the $2^{\ell/2}$-sized mixture $\mathrm{Bern}\left(\frac{1}{\ell}\left(1 \pm \frac{\epsilon}{\sqrt{n}}\right)\right)$ (the former and the later are due to our reduction).

$$
\begin{aligned}
&1 + d_{\chi^2}(R_{12}^{\otimes k}, S_{12}^{\otimes k}) \\
&= \sum_{i=0}^{k}\sum_{j=0}^{k} \frac{\left[\frac{1}{2}\binom{k}{i}(p_+)^i(1-p_+)^{k-i} + \frac{1}{2}\binom{k}{j}(p_-)^j(1-p_-)^{k-j}\right]^2}{\binom{k}{i}\left(\frac{1}{\ell}\right)^i\left(1-\frac{1}{\ell}\right)^{k-i}\binom{k}{j}\left(\frac{1}{\ell}\right)^j\left(1-\frac{1}{\ell}\right)^{k-j}} \\
&= \frac{\ell^{2k}}{4}\sum_{i=0}^{k}\sum_{j=0}^{k}\binom{k}{i}\binom{k}{j}[(p_+^2)^i\left(\frac{(1-p_+)^2}{(\ell-1)}\right)^{k-i}(p_-^2)^j\left(\frac{(1-p_-)^2}{(\ell-1)}\right)^{k-j} + \\
&\qquad (p_-^2)^i\left(\frac{(1-p_-)^2}{(\ell-1)}\right)^{k-i}(p_+^2)^j\left(\frac{(1-p_+)^2}{(\ell-1)}\right)^{k-j} + \\
&\qquad 2\,(p_+p_-)^i\left(\frac{(1-p_+)(1-p_)}{(\ell-1)}\right)^{k-1}(p_-p_+)^j\left(\frac{(1-p_)(1-p_+)}{(\ell-1)}\right)^{k-j}] \\
&= \frac{\ell^{2k}}{4}\left[2\left(p_+^2 + \frac{(1-p_+)^2}{\ell-1}\right)^k\left(p_-^2 + \frac{(1-p_-)^2}{\ell-1}\right)^k + 2\left(p_+p_- + \frac{(1-p_+)(1-p_-)}{(\ell-1)}\right)^{2k}\right]
\end{aligned}
$$

$$
p_+^2 + \frac{(1-p_+)^2}{(\ell-1)} = \frac{1}{\ell^2}\left(1+\frac{\epsilon}{\sqrt{n}}\right)^2 + \frac{1}{(\ell-1)}\left(1-\frac{1}{\ell-\frac{\epsilon}{l\sqrt{n}}}\right)^2 \quad = \frac{1}{\ell} + \frac{\epsilon^2}{n\ell(\ell-1)} \qquad \text{(upon simplification)}
$$

$$
p_+^2 + \frac{(1-p_+)^2}{(\ell-1)} = \frac{1}{\ell} + \frac{\epsilon^2}{n\ell(\ell-1)} \qquad \text{(upon simplification)}
$$

$$
\left(p_+p_- + \frac{(1-p_+)(1-p_)}{(\ell-1)}\right) = \frac{1}{\ell} - \frac{\epsilon^2}{n\ell(\ell-1)} \qquad \text{(upon simplification)}
$$

Therefore,

$$
\begin{aligned}
1 + d_{\chi^2}(R_{12}^{\otimes k}, S_{12}^{\otimes k}) &\le \frac{\ell^{2k}}{2}\left[\left(\frac{1}{\ell} + \frac{\epsilon^2}{n\ell(\ell-1)}\right)^{2k} + \left(\frac{1}{\ell} - \frac{\epsilon^2}{n\ell(\ell-1)}\right)^{2k}\right] \\
&\approx 1 + \binom{2k}{2}\left(\frac{\epsilon^2}{n(\ell-1)}\right)^2 + \dots
\end{aligned}
$$

We get that if $k = o(\sqrt{nl}\epsilon^{-2})$, then $d_{\chi^2}(R_{12}^{\otimes k}, S_{12}^{\otimes k}) = o(\frac{1}{nl})$. Then we get from Equation (3) and Equation (4), $d_{\mathrm{TV}}(P^{\otimes k}, Q^{\otimes k}) = o(1)$, establishing Theorem 2.3. ■

5. Tolerant Testing in d_{H} for High-Dimensional Distributions

In this section, we give an algorithm for tolerant testing of two high-dimensional distributions w.r.t. the Hellinger distance. More specifically, given samples from such a pair of unknown distributions

P and Q our goal would be to distinguish between the cases: $d_{\mathrm{H}}(P,Q) \le \epsilon/2$ versus $d_{\mathrm{H}}(P,Q) > \epsilon$ with 2/3 probability, which can be amplified to $1-\delta$ using the majority of $O(\log \frac{1}{\delta})$ repetitions, for any $0 < \delta, \epsilon < 1$. This generalizes the work of Bhattacharyya, Gayen, Meel, and Vinodchandran (2020), who gave such tolerant testers w.r.t. d_{TV} using distance approximation.

We start with a result that additively estimates $d_{\mathrm{H}}^2(P,Q)$, when we have access to both the p.m.f.s and also to independent samples from P.

Theorem 5.1 *Consider P and Q be two unknown distributions over Σ^n. Suppose we have access to two circuits $\xi_P(x)$ and $\xi_Q(x)$ which on input x, outputs $P(x)$ and $Q(x)$ respectively. Then we can output a number e such that $|e - d_{\mathrm{H}}^2(P,Q)| \le \epsilon$ with 2/3 probability for any $0 < \epsilon < 1$, using $3\epsilon^{-2}$ independent samples from P and $3\epsilon^{-2}$ calls to each of $\xi_P(x)$ and $\xi_Q(x)$.*

Proof

$$\begin{aligned}
1 - d_{\mathrm{H}}^2(P,Q) &= \sum_{x\in\Sigma^n} \sqrt{P(x)Q(x)} \\
&= \sum_{x\in\Sigma^n} \sqrt{\frac{Q(x)}{P(x)}} P(x) \\
&= \mathrm{E}_{x\sim P}\left[\sqrt{\frac{Q(x)}{P(x)}}\right] && \text{(since } P(x) \neq 0\text{)}
\end{aligned}$$

Let $f(x) = \sqrt{\frac{Q(x)}{P(x)}}$. Therefore, it suffices to estimate $\mathrm{E}_{x\sim P}[f(x)]$ additively. Note that, $\mathrm{Var}_{x\sim P}[f(x)] \le \mathrm{E}_{x\sim P}\left[f^2(x)\right] = \sum_x Q(x) = 1$. We define our estimator to be e, the average of $(1-f(x))$ over R samples from P. Then e satisfies $\mathrm{E}[e] = d_{\mathrm{H}}^2(P,Q)$ and $\mathrm{Var}[e] \le 1/R$. Chebyshev's inequality gives us that for $R \ge 3\epsilon^{-2}$, $|e - d_{\mathrm{H}}^2(P,Q))| \le \epsilon$ with at least 2/3 probability. ∎

5.1. Application: Bayesian Networks

Bhattacharyya, Gayen, Meel, and Vinodchandran (2020) have given the following Algorithm for learning an unknown Bayesian network on a known graph of indegree at most d.

Theorem 5.2 *There is an algorithm that given a parameter $\epsilon > 0$ and sample access to an unknown Bayesian network distribution P on a known directed acyclic graph G of in-degree at most d, returns a Bayesian network $\hat{P}$ on G such that $d_{\mathrm{H}}(P,\hat{P}) \le \epsilon$ with probability $\ge 9/10$. Letting Σ denote the range of each variable X_i, the algorithm takes $m = O(|\Sigma|^{d+1} n \log(|\Sigma|^{d+1} n)\epsilon^{-2})$ samples and runs in $O(mn)$ time.*

We get the following result for tolerant resting of Bayesian networks in Hellinger distance.

Theorem 2.8 (d_{H}-vs-d_{H} closeness tester for Bayes nets) *Given samples from two unknown Bayesian networks P and Q over Σ^n on potentially different but known pair of graphs of indegree at most d, we can distinguish the cases $d_{\mathrm{H}}(P,Q) \le \epsilon/2$ versus $d_{\mathrm{H}}(P,Q) > \epsilon$ with 2/3 probability using $m = O(|\Sigma|^{d+1} n \log(|\Sigma|^{d+1} n)\epsilon^{-2})$ samples and $O(|\Sigma|^{d+1} mn + n\epsilon^{-4})$ time.*

Proof First we learn P and Q using Theorem 5.2 such that $d_{\mathrm{H}}(P, \hat{P}) \le \epsilon/12$ and $d_{\mathrm{H}}(Q, \hat{Q}) \le \epsilon/12$. This step costs $m = O(|\Sigma|^{d+1} n \log(|\Sigma|^{d+1} n)\epsilon^{-2})$ samples, runs in $O(mn)$ time, and succeeds with 4/5 probability. Note that $\hat{P}$ and $\hat{Q}$, once learnt, can be sampled and evaluated correctly in $O(n)$ time.

Next we estimate $d_{\mathrm{H}}^2(\hat{P}, \hat{Q})$ up to an additive $\epsilon^2/9$ error using Theorem 5.1. This step costs $O(n\epsilon^{-4})$ time and no further samples and succeeds with 4/5 probability. Due to the triangle inequality of d_{H}, in the first case, $d_{\mathrm{H}}^2(\hat{P}, \hat{Q}) \le 20\epsilon^2/36$ and in the second case $d_{\mathrm{H}}^2(\hat{P}, \hat{Q}) > 21\epsilon^2/36$, thus separating the two cases. ■

6. Non-tolerant Closeness Testers

6.1. d_{H}-tester

A non-tolerant tester for 2-sample testing of product distributions was given in Canonne, Diakonikolas, Kane, and Stewart (2017). Their tester distinguishes $P = Q$ from $d_{\mathrm{TV}}(P, Q) \ge \epsilon$ with sample complexity $O(\max\{n^{3/4}/\epsilon, \sqrt{n}/\epsilon^2\})$. Using the relation, $d_{\mathrm{TV}} \ge d_{\mathrm{H}}^2$ from Lemma 2, we immediately get a tester for distinguishing $P = Q$ from $d_{\mathrm{H}}(P, Q) \ge \epsilon$, with sample complexity $O(\max\{n^{3/4}/\epsilon^2, \sqrt{n}/\epsilon^4\})$. Here, we show an improved tester with $O(n^{3/4}/\epsilon^2)$ sample complexity in Algorithm 2. We analyze its correctness and complexity below.

Let $P = \prod_{i=1}^n P_i$ and $Q = \prod_{i=1}^n Q_i$, with $P_i = \langle p_{i1}, \ldots, p_{i\ell}\rangle$ and $Q_i = \langle q_{i1}, \ldots, q_{i\ell}\rangle$ as probability vectors. We assume $\min_{i,j} p_{ij} \ge \epsilon^2/50n\ell$ and $\min_{i,j} q_{ij} \ge \epsilon^2/50n\ell$, without loss of generality using the reduction of Lemma 6.

Analysis of Algorithm 2 can be divided into two cases: 'heavy' and 'light'. Let $V \subseteq [n] \times [\ell]$ be the 'light' set of indices (i, j), such that $\max\{p_{ij}, q_{ij}\} < 1/m$. The remaining indices in $U = [n] \times [\ell] \setminus V$ are 'heavy'. The following important lemma shows that for each case, a certain sum must deviate from zero substantially, for the 'no' class.

Lemma 16 *Suppose* $d_{\mathrm{H}}(P, Q) \ge \epsilon$. *Suppose* $\min_{i,j} p_{ij} \ge \epsilon^2/50n\ell$ *and* $\min_{i,j} q_{ij} \ge \epsilon^2/50n\ell$. *Then at least one of the following two must hold:*

1. $\sum_{(i,j)\in V} (p_{ij} - q_{ij})^2 \ge \epsilon^4/25n\ell$
2. $\sum_{(i,j)\in U} \frac{(p_{ij}-q_{ij})^2}{p_{ij}+q_{ij}} \ge \epsilon^2$.

Proof If $d_{\mathrm{H}}(P, Q) \ge \epsilon$, Fact 4 gives us $\sum_i d_{\mathrm{H}}^2(P_i, Q_i) \ge \epsilon^2$. We use the following standard Fact to get $\sum_{i=1}^n \sum_{j=1}^\ell \frac{(p_{ij}-q_{ij})^2}{p_{ij}+q_{ij}} \ge 2\sum_i d_{\mathrm{H}}^2(P_i, Q_i) \ge 2\epsilon^2$.

Fact 17 *(see eg. Daskalakis, Kamath, and Wright (2018)) For two distributions* $P = \{p_1, \ldots, p_\ell\}$ *and* $Q = \{q_1, \ldots, q_\ell\}$, $\sum_j \frac{(p_j-q_j)^2}{p_j+q_j} \ge 2d_{\mathrm{H}}^2(P, Q)$.

It follows that at least one of 1) $\sum_{(i,j)\in V} \frac{(p_{ij}-q_{ij})^2}{p_{ij}+q_{ij}}$ or 2) $\sum_{(i,j)\in U} \frac{(p_{ij}-q_{ij})^2}{p_{ij}+q_{ij}}$ is at least ϵ^2.

Algorithm 2: Given samples from two unknown distributions $P = P_1 \times \cdots \times P_n$ and $Q = Q_1 \times \cdots \times Q_n$ over Σ^n, decides $P = Q$ ('yes') versus $d_{\mathrm{H}}(P, Q) \geq \epsilon$ ('no'). Let $\ell = |\Sigma|$.

`/* Approximately identify heavy and light partitions */`

Take m samples from P and Q. Let $U' \subseteq [n] \times [\ell]$ be the set of indices (i, j), such that at least one sample from either P or Q has hit symbol $j \in \Sigma$ in the coordinate i;

Let $V' = [n] \times [\ell] \setminus U'$;

`/* Poisson sampling */`

For each $i \in [n]$, sample $M_i \sim \mathrm{Poi}(m)$ independently

For each $i \in [n]$, sample $M_i' \sim \mathrm{Poi}(m)$ independently

Let $M = \max_i\{M_i\}$ and $M' = \max_i\{M_i'\}$

If $\max\{M, M'\} \geq 2m$ output 'no'

Take M samples $X^1, \ldots, X^M$ from P

Take M' samples $Y^1, \ldots, Y^{M'}$ from Q

For every (i, j), let W_{ij} be the number of occurrences of symbol $j \in \Sigma$ in the i-th coordinate of the sample subset $X^1, \ldots, X^{M_i}$

For every (i, j), let V_{ij} be the number of occurrences of symbol $j \in \Sigma$ in the i-th coordinate of the sample subset$Y^1, \ldots, Y^{M_i'}$

`/* Test the heavy partition */`

$W_{heavy} = \sum_{(i,j) \in U'} \frac{(W_{ij} - V_{ij})^2 - (W_{ij} + V_{ij})}{(W_{ij} + V_{ij})}$

If $W_{heavy} > m\epsilon^2/120$ output 'no'

`/* Test the light partition */`

$W_{light} = \sum_{(i,j) \in V'} (W_{ij} - V_{ij})^2 - (W_{ij} + V_{ij})$

If $W_{light} > m^2\epsilon^4/1000n\ell$ output 'no'

Output 'yes'

In the first case,

$$\begin{aligned}\sum_{(i,j)\in V}(p_{ij}-q_{ij})^2 &= \sum_{(i,j)\in V}(\sqrt{p_{ij}}-\sqrt{q_{ij}})^2(\sqrt{p_{ij}}+\sqrt{q_{ij}})^2 \\ &\geq (2\epsilon^2/25n\ell)\cdot\sum_{(i,j)\in V}(\sqrt{p_{ij}}-\sqrt{q_{ij}})^2 \qquad (\text{as } \min_{i,j}\min\{p_{ij},q_{ij}\}\geq\epsilon^2/50n\ell) \\ &\geq (2\epsilon^2/25n\ell)\cdot\sum_{(i,j)\in V}\frac{(p_{ij}-q_{ij})^2}{(\sqrt{p_{ij}}+\sqrt{q_{ij}})^2} \\ &\geq (\epsilon^2/25n\ell)\cdot\sum_{(i,j)\in V}\frac{(p_{ij}-q_{ij})^2}{p_{ij}+q_{ij}} \\ &\geq \epsilon^4/25n\ell.\end{aligned}$$

■

The following lemma shows U' (V'), as obtained in Lines 1-2 of Algorithm 2, could be an acceptable proxy for U (V).

Lemma 18 *Let U', V' be as in Algorithm 2. Let $m = \Omega(\sqrt{n\ell}/\epsilon^2)$ for some sufficiently large constant. Then with probability at least 0.63 in each case, the following holds:*

1. $\sum_{(i,j)\in U}\frac{(p_{ij}-q_{ij})^2}{p_{ij}+q_{ij}}\geq\epsilon^2$ *implies* $\sum_{(i,j)\in U\cap U'}\frac{(p_{ij}-q_{ij})^2}{p_{ij}+q_{ij}}\geq\epsilon^2/20$
2. $\sum_{(i,j)\in V}(p_{ij}-q_{ij})^2\geq\epsilon^4/25n\ell$ *implies* $\sum_{(i,j)\in V'}(p_{ij}-q_{ij})^2\geq\epsilon^4/500n\ell$

Proof Note that $(i,j)\in V'$ with probability = $(1-p_{ij})^m(1-q_{ij})^m$, and $(i,j)\in U'$ with the remaining probability.

(*Proof of 1:*) Let $U'' = U\cap U'$. Suppose there exists $(i,j)\in U$ such that $\frac{(p_{ij}-q_{ij})^2}{p_{ij}+q_{ij}}\geq\epsilon^2/20$. Then this $(i,j)\in U''$ with probability $1-(1-p_{ij})^m(1-q_{ij})^m\geq 1-(1-1/m)^m\geq 0.63$, in which case the result follows. Otherwise, we consider sum of the independent random variables, $S=\sum_{(i,j)}1_{(i,j)\in U''}\frac{20(p_{ij}-q_{ij})^2}{\epsilon^2(p_{ij}+q_{ij})}$, each of which is in $[0,1]$. $\mathrm{E}[S]=\sum_{(i,j)\in U}(1-(1-p_{ij})^m(1-q_{ij})^m)\frac{20(p_{ij}-q_{ij})^2}{\epsilon^2(p_{ij}+q_{ij})}\geq 12.6$. We apply Chernoff's bound to get $S\geq 6.3$ with probability 0.63.

(*Proof of 2:*) Let $V''=V\cap V'$. We consider sum of the independent random variables, $S=\sum_{(i,j)}1_{(i,j)\in V''}m^2(p_{ij}-q_{ij})^2$, each of which is in $[0,1]$. $\mathrm{E}[S]=\sum_{(i,j)\in V}(1-p_{ij})^m(1-q_{ij})^m m^2(p_{ij}-q_{ij})^2 > (1-1/m)^{2m}\sum_{(i,j)\in V}m^2(p_{ij}-q_{ij})^2\geq m^2\epsilon^4/250n\ell$, for $m\geq 4$. We apply Chernoff's bound to get $S\geq m^2\epsilon^4/500n\ell$ except for probability at most $\exp(-m^2\epsilon^4/3000n\ell)$. ■

Combining Lemma 16 and Lemma 18, we get for the 'no' case, one of the two conditions of Lemma 18 must hold. Algorithm 2 uses the two tests W_{heavy} and W_{light} to check these two conditions separately. To analyze them, we use certain important results from Canonne, Diakonikolas, Kane, and Stewart (2017), assuming $W_{ij}\sim\mathrm{Poi}(p_{ij})$ and $V_{ij}\sim\mathrm{Poi}(q_{ij})$ for every i,j, which holds due to Poisson sampling. We assume the check of Line 7 goes through except 1/50 probability, using the concentration of Poisson distribution.

Analysis of W_{heavy}

Lemma 19 (Obtained from Claims 37 and 38 of Canonne, Diakonikolas, Kane, and Stewart (2017)) *If $P = Q$ then $\mathrm{E}[W_{heavy}] = 0$. If $\sum_{(i,j)\in U\cap U'} \frac{(p_{ij}-q_{ij})^2}{p_{ij}+q_{ij}} \geq \epsilon^2/20$ then $\mathrm{E}[W_{heavy}] \geq m\epsilon^2/60$. In both cases $\mathrm{Var}[W_{heavy}] \leq 7n\ell + 15\mathrm{E}[W_{heavy}]$*

By the application of Chebyshev's inequality we get the following.

Lemma 20 *Let $m = \Omega(\sqrt{n\ell}/\epsilon^2)$ for a sufficiently large constant. Then the following holds except for probability $\leq 1/25$ in each case,*

1. *$P = Q$ implies $W_{heavy} \leq m\epsilon^2/120$*
2. *$\sum_{(i,j)\in U\cap U'} \frac{(p_{ij}-q_{ij})^2}{p_{ij}+q_{ij}} \geq \epsilon^2/20$ implies $W_{heavy} > m\epsilon^2/120$.*

Analysis of W_{light} We use the following result given in Proposition 6 of Chan, Diakonikolas, Valiant, and Valiant (2014) and Claim 35 of Canonne, Diakonikolas, Kane, and Stewart (2017).

Lemma 21 *$\mathrm{E}[W_{light}] = m^2 \sum_{(i,j)\in V'} (p_{ij} - q_{ij})^2$ and $\mathrm{Var}[W_{light}] \leq 80m^3\sqrt{b} \sum_{(i,j)\in V'} (p_{ij} - q_{ij})^2 + 8m^2 b$, where $b = \max\{\sum_{(i,j)\in V'} p_{ij}^2, \sum_{(i,j)\in V'} q_{ij}^2\}$. Furthermore, $b \leq 50n\ell/m^2$ for a sufficiently large m, except for probability at most $1/50$.*

By the application of Chebyshev's inequality we get the following.

Lemma 22 *Let $m = \Omega((n\ell)^{3/4}/\epsilon^2)$ for a sufficiently large constant. Then the following holds except for probability $\leq 1/50$ in each case,*

1. *$P = Q$ implies $W_{light} \leq m^2\epsilon^4/1000n\ell$*
2. *$\sum_{(i,j)\in V'}(p_{ij} - q_{ij})^2 \geq \epsilon^4/500n\ell$ implies $W_{light} > m^2\epsilon^4/1000n\ell$.*

Together we get $O((n\ell)^{3/4}/\epsilon^2)$ samples are enough to distinguish $P = Q$ versus $d_{\mathrm{H}}(P,Q) \geq \epsilon$.

Theorem 2.6 (Exact-versus-d_{H} closeness tester) *There is an algorithm with sample access to two unknown product distribution $P = \prod_{i=1}^n P_i$ and $Q = \prod_{i=1}^n Q_i$, both over the common sample space Σ^n, that decides between cases $P = Q$ versus $\sqrt{2}d_{\mathrm{H}}(P,Q) > \epsilon$. The algorithm takes $m = O((n|\Sigma|)^{3/4}/\epsilon^2)$ samples from P and Q and runs in time $O(mn)$. The algorithm has a success probability at least $2/3$.*

6.2. d_{TV}-tester

A 2-sample tester for distinguishing $P = Q$ from $d_{\mathrm{TV}}(P,Q) \geq \epsilon$, for product distributions over $\{0,1\}^n$, was given in Canonne, Diakonikolas, Kane, and Stewart (2017) with sample complexity $O(\sqrt{n}/\epsilon^2, \max\{n^{3/4}/\epsilon\})$. In the following, we generalize this result for product distributions over Σ^n with sample complexity $m = O(\max\{\sqrt{n|\Sigma|}/\epsilon^2, (n|\Sigma|)^{3/4}/\epsilon\})$ which is optimal for $|\Sigma| = 2$ (Canonne, Diakonikolas, Kane, and Stewart, 2017). Let $\ell = |\Sigma|$. We assume $\min_{i,j} p_{ij} \geq \epsilon/50n\ell$ and $\min_{i,j} q_{ij} \geq \epsilon/50n\ell$ without loss of generality, using a reduction similar to Lemma 6 (Canonne, Diakonikolas, Kane, and Stewart, 2017; Daskalakis and Pan, 2017). Let $V \subseteq [n] \times [\ell]$ be the set of indices (i,j), such that $\max\{p_{ij}, q_{ij}\} < 1/m$. Let $U = [n] \times [\ell] \setminus V$.

Lemma 23 *Suppose* $d_{\mathrm{TV}}(P,Q) \geq \epsilon$. *Suppose* $\min_{i,j} p_{ij} \geq \epsilon/50n\ell$ *and* $\min_{i,j} q_{ij} \geq \epsilon/50n\ell$. *Then at least one of the following two must hold:*

1. $\sum_{(i,j)\in V}(p_{ij}-q_{ij})^2 \geq \epsilon^2/n\ell$
2. $\sum_{(i,j)\in U} \frac{(p_{ij}-q_{ij})^2}{p_{ij}+q_{ij}} \geq \epsilon^2/4$.

Proof We get $\sum_{(i,j)\in V}|p_{ij}-q_{ij}| + \sum_{(i,j)\in U}|p_{ij}-q_{ij}| = 2\sum_i d_{\mathrm{TV}}(P_i,Q_i) \geq 2d_{\mathrm{TV}}(P,Q) \geq 2\epsilon$, the second last inequality from the following Fact.

Fact 24 *For two product distributions* $P = \Pi_{i=1}^n P_i$ *and* $Q = \Pi_{i=1}^n Q_i$, $d_{\mathrm{TV}}(P,Q) \leq \sum_i d_{\mathrm{TV}}(P_i,Q_i)$.

Hence at least one of $\sum_{(i,j)\in V}|p_{ij}-q_{ij}|$ or $\sum_{(i,j)\in U}|p_{ij}-q_{ij}|$ is at least ϵ.

In the first case, we get

$$\sum_{(i,j)\in V}(p_{ij}-q_{ij})^2 \sum_{(i,j)\in V} 1 \geq (\sum_{(i,j)\in V}|p_{ij}-q_{ij}|)^2 \qquad \text{(Cauchy-Schwarz inequality)}$$
$$\sum_{(i,j)\in V}(p_{ij}-q_{ij})^2 \geq \epsilon^2/n\ell.$$

In the second case, the proof is similar to that of the standard Facts $d_{\mathrm{TV}} \leq d_{\mathrm{H}}$ and Fact 17 (see eg. Daskalakis, Kamath, and Wright (2018) for both).

$$\begin{aligned}
\epsilon^2 &\leq (\sum_{(i,j)\in U}|p_{ij}-q_{ij}|)^2 \\
&= (\sum_{(i,j)\in U}|\sqrt{p_{ij}}-\sqrt{q_{ij}}||\sqrt{p_{ij}}+\sqrt{q_{ij}}|)^2 \\
&\leq (\sum_{(i,j)\in U}(\sqrt{p_{ij}}-\sqrt{q_{ij}})^2)(\sum_{(i,j)\in U}(\sqrt{p_{ij}}+\sqrt{q_{ij}})^2) \qquad \text{(Cauchy-Schwarz inequality)} \\
&\leq (\sum_{(i,j)\in U}(\sqrt{p_{ij}}-\sqrt{q_{ij}})^2)(\sum_{(i,j)\in U}2(p_{ij}+q_{ij})) \\
&\leq 4\sum_{(i,j)\in U}\frac{(p_{ij}-q_{ij})^2}{(\sqrt{p_{ij}}+\sqrt{q_{ij}})^2} \\
&\leq 4\sum_{(i,j)\in U}\frac{(p_{ij}-q_{ij})^2}{p_{ij}+q_{ij}}.
\end{aligned}$$

■

We skip the rest of the details of the algorithm and its analysis since it closely follows that of Section 6.1. We identify the partitions U and V approximately by checking which indices are hit in m samples. For $m = \Omega(\sqrt{n\ell}/\epsilon)$, this approximation is acceptable using a result similar to Lemma 18. Lemma 19, Lemma 20 and Lemma 21 are as before up to the constants. Only in Lemma 22, the threshold for the light part changes to $m^2\epsilon^2/40n\ell$, and the sample complexity for the light part changes to $m = \Theta((n\ell)^{3/4}/\epsilon)$.

Theorem 2.7 (Exact-versus-d_{TV} closeness tester) *There is an algorithm with sample access to two unknown product distribution $P = \prod_{i=1}^n P_i$ and $Q = \prod_{i=1}^n Q_i$, both over the common sample space Σ^n, that decides between cases $P = Q$ versus $d_{\mathrm{TV}}(P, Q) > \epsilon$. The algorithm takes $m = O(\max\{\sqrt{n|\Sigma|}/\epsilon^2, (n|\Sigma|)^{3/4}/\epsilon\})$ samples and runs in time $O(mn)$. The algorithm has a success probability at least $2/3$.*

Acknowledgement We thank Clément Canonne for his comments on an earlier version of this paper. We also thank the anonymous reviewers of ALT 21 for improving the paper. Arnab and Sutanu were supported by the NRF-AI Fellowship R-252-100-B13-281. Arnab was also supported by an Amazon Research Award. Saravanan was supported by the NSF 1846300 CAREER award. Vinod was supported by NSF CCF-184908 and NSF HDR:TRIPODS-1934884 awards.

References

Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3591–3599, 2015. URL http://papers.nips.cc/paper/5839-optimal-testing-for-properties-of-distributions.

Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White. Testing closeness of discrete distributions. *Journal of the ACM (JACM)*, 60(1):4, 2013.

Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, and N. V. Vinodchandran. Efficient distance approximation for structured high-dimensional distributions via learning. *CoRR*, abs/2002.05378, 2020. URL https://arxiv.org/abs/2002.05378.

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. URL http://www.theoryofcomputing.org/library.html.

Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing bayesian networks. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 370–448, 2017. URL http://proceedings.mlr.press/v65/canonne17a.html.

Siu-On Chan, Ilias Diakonikolas, Paul Valiant, and Gregory Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1193–1203. SIAM, 2014.

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

Constantinos Daskalakis and Qinxuan Pan. Square hellinger subadditivity for bayesian networks and its applications to identity testing. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 697–703, 2017. URL http://proceedings.mlr.press/v65/daskalakis17a.html.

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*, SODA '18, pages 2747–2764, Philadelphia, PA, USA, 2018. Society for Industrial and Applied Mathematics. ISBN 978-1-6119-7503-1. URL http://dl.acm.org/citation.cfm?id=3174304.3175479.

Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing ising models. *IEEE Trans. Inf. Theory*, 65(11):6829–6852, 2019. doi: 10.1109/TIT.2019.2932255. URL https://doi.org/10.1109/TIT.2019.2932255.

Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation*, pages 68–75. Springer, 2011.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the l_1 distance. *IEEE Trans. Inf. Theory*, 64(10):6672–6706, 2018. doi: 10.1109/TIT.2018.2846245. URL https://doi.org/10.1109/TIT.2018.2846245.

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

Ronitt Rubinfeld. Taming big probability distributions. *ACM Crossroads*, 19(1):24–28, 2012. doi: 10.1145/2331042.2331052. URL https://doi.org/10.1145/2331042.2331052.

Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:183, 2010. URL http://eccc.hpi-web.de/report/2010/179.

Gregory Valiant and Paul Valiant. The power of linear estimators. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pages 403–412. IEEE, 2011.

Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, FOCS '14, pages 51–60, Washington, DC, USA, 2014. IEEE Computer Society. ISBN 978-1-4799-6517-5. doi: 10.1109/FOCS.2014.14. URL http://dx.doi.org/10.1109/FOCS.2014.14.

Appendix A. Proof of Theorem 3.1

Theorem 3.1 (Modified from Acharya, Daskalakis, and Kamath (2015)) *Let m be an integer and $0 < \epsilon < 1$ be an error parameter. Let $r_1, r_2, \ldots, r_K$ be K non-negative real numbers. Let $s_1, s_2, \ldots, s_K$ be non-negative real numbers such that $s_i \geq \epsilon^2/50K$. For $1 \leq i \leq K$, let $N_i \sim \mathrm{Poi}(mr_i)$ be independent samples from $\mathrm{Poi}(mr_i)$. Then there exists a test statistic T, computable in time $O(K)$ from inputs N_is and s_is, with the following guarantees.*

- $\mathrm{E}[T] = m \sum_i \frac{(r_i - s_i)^2}{s_i}$
- $\mathrm{Var}[T] \leq 2K + 7\sqrt{K}\mathrm{E}[T] + 4K^{1/4}(\mathrm{E}[T])^{3/2}$, *for a constant c and $m \geq c\sqrt{K}/\epsilon^2$.*

Proof The test T of Acharya, Daskalakis, and Kamath (2015) is given by $T = \Sigma_{i=1}^{K} \frac{(N_i - ms_i)^2 - N_i}{ms_i}$.

$$
\begin{aligned}
\mathrm{E}[T] &= \mathrm{E}\left[\sum_i \frac{(N_i - ms_i)^2 - N_i}{ms_i}\right] \\
&= \sum_i \frac{\mathrm{E}[(N_i - ms_i)^2 - N_i]}{ms_i} \\
&= \sum_i \frac{\mathrm{E}[N_i^2] + m^2 s_i^2 - 2ms_i \mathrm{E}[N_i] - \mathrm{E}[N_i]}{ms_i} \\
&= \sum_i \frac{mr_i(1 + mr_i) + m^2 s_i^2 - 2ms_i \cdot mr_i - mr_i}{ms_i} && \text{(Since } N_i \sim \mathrm{Poi}(mr_i)\text{)} \\
&= m \sum_i \frac{(r_i - s_i)^2}{s_i}.
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}[T] &= \mathrm{Var}\left[\sum_i \frac{(N_i - ms_i)^2 - N_i}{ms_i}\right] \\
&= \sum_i \mathrm{Var}\left[\frac{(N_i - ms_i)^2 - N_i}{ms_i}\right] && \text{(Since } N_i\text{s are independent for different } i\text{s)} \\
&= \sum_i \frac{1}{m^2 s_i^2} \mathrm{Var}[N_i^2 - (2ms_i + 1)N_i] \\
&= \sum_i \frac{1}{m^2 s_i^2} [\mathrm{Var}[N_i^2] + (2ms_i + 1)^2 \mathrm{Var}[N_i] - 2(2ms_i + 1)\mathrm{Cov}(N_i^2, N_i)] \\
&= \sum_i \frac{1}{m^2 s_i^2} [(\mathrm{E}[N_i^4] - \mathrm{E}^2[N_i^2]) + (2ms_i + 1)^2 (\mathrm{E}[N_i^2] - \mathrm{E}^2[N_i]) \\
&\qquad - 2(2ms_i + 1)(\mathrm{E}[N_i^3] - \mathrm{E}[N_i^2]\mathrm{E}[N_i])] \\
&= \sum_i \frac{1}{m^2 s_i^2} [\lambda(1 + 5\lambda + 4\lambda^2) + \lambda(2ms_i + 1)^2 - 2(2ms_i + 1)(\lambda(2\lambda + 1))] \\
& && \text{(Since } N_i \sim \mathrm{Poi}(\lambda)\text{, where } \lambda = mr_i\text{)}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_i \frac{1}{m^2 s_i^2} \lambda[\lambda + (2ms_i - 2\lambda)^2] \\
&= \sum_i \frac{r_i^2}{s_i^2} + \sum_i 4mr_i \frac{(r_i - s_i)^2}{s_i^2}.
\end{aligned}
$$

We bound the above two summations separately.

$$
\begin{aligned}
\sum_i \frac{r_i^2}{s_i^2} &= \sum_i \frac{(r_i - s_i)^2 + 2s_i(r_i - s_i) + s_i^2}{s_i^2} \\
&= \sum_i \frac{(r_i - s_i)^2}{s_i^2} + 2\sum_i \frac{r_i - s_i}{s_i} + \sum_i 1 \\
&\leq 2\left(\sum_i \frac{(r_i - s_i)^2}{s_i^2} + \sum_i 1\right) && \text{(Using } a^2 + 1 \geq 2a\text{)} \\
&\leq 2\left(\frac{50K}{\epsilon^2} \sum_i \frac{(r_i - s_i)^2}{s_i} + K\right) && \text{(Using } s_i \geq \epsilon^2/50K\text{)} \\
&= 2(\frac{50K}{\epsilon^2} \frac{\mathrm{E}[T]}{m} + K) \\
&\leq \sqrt{K}\mathrm{E}[T] + 2K. && \text{(Using } m \geq c\sqrt{K}/\epsilon^2 \text{ for c sufficiently large)}
\end{aligned}
$$

$$
\begin{aligned}
\sum_i 4mr_i \frac{(r_i - s_i)^2}{s_i^2} &\leq 4m \sqrt{\sum_i \frac{r_i^2}{s_i^2}} \sqrt{\sum_i \frac{(r_i - s_i)^4}{s_i^2}} && \text{(Using Cauchy-Schwarz inequality)} \\
&\leq 4m \sqrt{\sqrt{K}\mathrm{E}[T] + 2K} \sum_i \frac{(r_i - s_i)^2}{s_i} \\
&\leq 4\mathrm{E}[T](K^{1/4}\sqrt{\mathrm{E}[T]} + \sqrt{2K}).
\end{aligned}
$$

Together we get

$$
\begin{aligned}
\mathrm{Var}[T] &\leq \sqrt{K}\mathrm{E}[T] + 2K + 4\mathrm{E}[T](K^{1/4}\sqrt{\mathrm{E}[T]} + \sqrt{2K}) \\
&\leq 2K + 7\sqrt{K}\mathrm{E}[T] + 4K^{1/4}(\mathrm{E}[T])^{3/2}.
\end{aligned}
$$

■