Electronic Colloquium on Computational Complexity, Report No. 2 (2018)

Which Distribution Distances are Sublinearly Testable?

Constantinos Daskalakis*
EECS & CSAIL, MIT
costis@csail.mit.edu

Gautam Kamath†
EECS & CSAIL, MIT
g@csail.mit.edu

John Wright‡
Physics, MIT
jswright@mit.edu

January 1, 2018

Abstract

Given samples from an unknown distribution p and a description of a distribution q, are p and q close or far? This question of "identity testing" has received significant attention in the case of testing whether p and q are equal or far in total variation distance. However, in recent work [VV11a, ADK15, DP17], the following questions have been been critical to solving problems at the frontiers of distribution testing:

- Alternative Distances: Can we test whether p and q are far in other distances, say Hellinger?
- Tolerance: Can we test when p and q are *close*, rather than equal? And if so, close in which distances?

Motivated by these questions, we characterize the complexity of distribution testing under a variety of distances, including total variation, ℓ_2, Hellinger, Kullback-Leibler, and χ^2. For each pair of distances d_1 and d_2, we study the complexity of testing if p and q are close in d_1 versus far in d_2, with a focus on identifying which problems allow *strongly* sublinear testers (i.e., those with complexity $O(n^{1-\gamma})$ for some $\gamma > 0$ where n is the size of the support of the distributions p and q). We provide matching upper and lower bounds for each case. We also study these questions in the case where we only have samples from q (equivalence testing), showing qualitative differences from identity testing in terms of when tolerance can be achieved. Our algorithms fall into the classical paradigm of χ^2-statistics, but require crucial changes to handle the challenges introduced by each distance we consider. Finally, we survey other recent results in an attempt to serve as a reference for the complexity of various distribution testing problems.

1 Introduction

The arch problem in science is determining whether observations of some phenomenon conform to a conjectured model. Often, phenomena of interest are probabilistic in nature, and so are our models of these phenomena; hence, testing their validity becomes a statistical hypothesis testing problem. In mathematical notation, suppose that we have access to samples from some unknown distribution p over some set Σ of size n. We also have a hypothesis distribution q, and our goal is to distinguish whether $p = q$ or $p \neq q$. For instance, we may want to test whether the sizes of some population of insects are normally distributed around their mean by sampling insects and measuring their sizes.

Of course, our models are usually imperfect. In our insect example, perhaps our estimation of the mean and variance of the insect sizes is a bit off. Furthermore, the sizes will clearly always be positive numbers. Yet a Normal distribution could still be a good fit. To get a meaningful testing problem some slack may be introduced, turning the problem into that of distinguishing whether $d_1(p, q) \leq \varepsilon_1$ versus $d_2(p, q) \geq \varepsilon_2$, for some distance measures $d_1(\cdot, \cdot)$ and $d_2(\cdot, \cdot)$ between distributions over Σ and some choice of ε_1 and ε_2 which may potentially depend on Σ or even q. Regardless, for the problem to be well-defined, the sets of distributions $\mathcal{C} = \{p \mid d_1(p, q) \leq \varepsilon_1\}$ and $\mathcal{F} = \{p \mid d_2(p, q) \geq \varepsilon_2\}$ should be disjoint. In fact, as our goal is to distinguish between $p \in \mathcal{C}$ and $p \in \mathcal{F}$ from samples, we cannot possibly draw the right conclusion with

*Supported by NSF CCF-1617730, CCF-1650733, and ONR N00014-12-1-0999.
†Supported by NSF CCF-1617730, CCF-1650733, and ONR N00014-12-1-0999. Part of this work was done while the author was an intern at Microsoft Research New England.
‡Supported by NSF grant CCF-6931885.

ISSN 1433-8092

probability 1 or detect the most minute deviations of p from $\mathcal{C}$ or $\mathcal{F}$. So our guarantee should be probabilistic, and there should be some "gap" between the sets $\mathcal{C}$ and $\mathcal{F}$. In sum, the problem is the following:

> (d_1, d_2)-*Identity Testing*: Given an explicit description of a distribution q over Σ, sample access to a distribution p over Σ, and bounds $\varepsilon_1 \geq 0$, and $\varepsilon_2, \delta > 0$, distinguish with probability at least $1 - \delta$ between $d_1(p, q) \leq \varepsilon_1$ and $d_2(p, q) \geq \varepsilon_2$, whenever p satisfies one of these two inequalities.

A related problem is when we have sample access to both p and q. For example, we might be interested in whether two populations of insects have distributions that are close or far. The resulting problem is the following:

> (d_1, d_2)-*Equivalence (or Closeness) Testing*: Given sample access to distributions p and q over Σ, and bounds $\varepsilon_1 \geq 0$, and $\varepsilon_2, \delta > 0$, distinguish with probability at least $1 - \delta$ between $d_1(p, q) \leq \varepsilon_1$ and $d_2(p, q) \geq \varepsilon_2$, whenever p, q satisfy one of these two inequalities.

The above questions are of course fundamental, and widely studied since the beginning of statistics. However, most tests only detect certain types of deviations of p from q, or are designed for distributions in parametric families. Moreover, most of the emphasis has been on the asymptotic sample regime. To address these challenges, there has been a surge of recent interest in information theory, property testing, and sublinear-time algorithms aiming at finite sample and d_1-close vs. d_2-far distinguishers, as in the formulations above; see e.g. [BFF$^+$01, BKR04, Pan08, VV17, ADK15, CDGR16, DK16]. This line of work has culminated in computationally efficient and sample optimal testers for several choices of distances d_1 and d_2, as well as error parameters ε_1 and ε_2, for example:

- for identity testing, when:
  - d_2 is taken to be the total variation distance, and $\varepsilon_1 = 0$ [BFF$^+$01, Pan08, VV17];
  - d_1 is taken to be the χ^2-divergence, d_2 is taken to be the total variation distance, and $\varepsilon_1 = (\varepsilon_2)^2/4$ [ADK15, DK16];
- for equivalence testing, when d_2 is taken to be the total variation distance, and $\varepsilon_1 = 0$ [BFR$^+$13, Val11, CDVV14].

There are also several other sub-optimal results known for other combinations of d_1, d_2, ε_1 and ε_2, and for many combinations there are no known testers. A more extensive discussion of the literature is provided in Section 1.2.

The goal of this paper is to *provide a complete mapping of the optimal sample complexity required to obtain computationally efficient testers for identity testing and equivalence testing under the most commonly used notions of distances d_1 and d_2*. Our results are summarized in Tables 1, 2, and 3 and discussed in detail in Section 1.1. In particular, we obtain computationally efficient and sample optimal testers for distances d_1 and d_2 ranging in the set $\{\ell_2$-distance, total variation distance, Hellinger distance, Kullback-Leibler divergence, χ^2-divergence$\}$,[1] and for combinations of these distances and choice of errors ε_1 and ε_2 which give rise to meaningful testing problems as discussed above. The sample complexities stated in the tables are for probability of error $1/3$. Throwing in extra factors of $O(\log 1/\delta)$ boosts the probability of error to $1 - \delta$, as usual.[2]

Our motivation for this work is primarily the fundamental nature of identity and equivalence testing, as well as of the distances under which we study these problems. It is also the fact that, even though distribution testing is by now a mature subfield of information theory, property testing, and sublinear-time algorithms, several of the testing questions that we consider have had unknown statuses prior to our work. This gap is accentuated by the fact that, as we establish, closely related distances may have radically different behavior. To give a quick example, it is easy to see that χ^2-divergence is the second-order Taylor expansion of KL-divergence. Yet, as we show, the sample complexity for identity testing changes radically when d_2 is taken to be total variation or Hellinger distance, and d_1 transitions from χ^2 to KL or weaker distances;

[1] These distances are nicely nested, as discussed in Section 2, from the weaker ℓ_2 to the stronger χ^2-divergence.

[2] Namely, one can repeat the test $O(\log 1/\delta)$ times and output the majority result. One can analyze the resulting probability of success by the Chernoff bound.

see Table 1. Prior to this work we knew about a transition somewhere between χ^2-divergence and total variation distance, but our work identifies a more refined understanding of the point of transition. Similar fragility phenomena are identified by our work for equivalence testing, when we switch from total variation to Hellinger distance, as seen in Tables 2 and 3.

Adding to the fundamental nature of the problems we consider here, we should also emphasize that a clear understanding of the different tradeoffs mapped out by our work is critical at this point for the further development of the distribution testing field, as recent experience has established. Let us provide a couple of recent examples, drawing from our prior work. Acharya, Daskalakis, and Kamath [ADK15] study whether properties of distributions, such as unimodality or log-concavity, can be tested in total variation distance. Namely, given sample access to a distribution p, how many samples are needed to test whether it has some property (modeled by a set $\mathcal{P}$ of distributions) or whether it is far from having the property, i.e. $d_{\mathrm{TV}}(p, \mathcal{P}) > \varepsilon$, for some error ε? Their approach is to first learn a proxy distribution $\hat{p} \in \mathcal{P}$ that satisfies $d'(p, \hat{p}) \leq \varepsilon'$ for some distance d', whenever $p \in \mathcal{P}$, then reduce the property testing problem to (d', d_{TV})-identity testing of p to $\hat{p}$. Interestingly, rather than picking d' to be total variation distance, they take it to be χ^2-divergence, which leads to optimal testers of sample complexity $O(\sqrt{n}/\varepsilon^2)$ for several $\mathcal{P}$'s such as monotone, unimodal, and log-concave distributions over $[n]$. Had they picked d' to be total variation distance, they would be stuck with a $\Omega(n/\log n)$ sample complexity in the resulting identity testing problem, as Table 1 illustrates, which would lead to a suboptimal overall tester. The choice of χ^2-divergence in the work of Acharya et al. was somewhat ad hoc. By providing a full mapping of the sample complexity tradeoffs in the use of different distances, we expect to help future work in identifying better where the bottlenecks and opportunities lie.

Another example supporting our expectation can be found in recent work of Daskalakis and Pan [DP17]. They study equivalence testing of Bayesian networks under total variation distance. Bayesian networks are flexible models expressing combinatorial structure in high-dimensional distributions in terms of a directed acyclic graph (DAG) specifying their conditional dependence structure. The challenge in testing Bayes nets is that their support scales exponentially in the number of nodes, and hence naive applications of known equivalence tests lead to sample complexities that are exponential in the number of nodes, even when the in-degree δ of the underlying DAGs is bounded. To address this challenge, Daskalakis and Pan establish "localization-of-distance" results of the following form, for various choices of distance d: "If two Bayes nets P and Q are ε-far in total variation distance, then there exists a small set of nodes S (whose size is $\Delta + 1$, where Δ is again the maximum in-degree of the underlying DAG where P and Q are defined) such that the marginal distributions of P and Q over the nodes of set S are ε'-far under distance d." When they take d to be total variation distance, they can show $\varepsilon' = \Omega(\varepsilon/m)$, where m is the number of nodes in the underlying DAG (i.e. the dimension). Given this localization of distance, to test whether two Bayes nets P and Q satisfy $P = Q$ vs. $d_{\mathrm{TV}}(P, Q) \geq \varepsilon$, it suffices to test, for all relevant marginals P_S and Q_S whether $P_S = Q_S$ vs. $d_{\mathrm{TV}}(P_S, Q_S) = \Omega(\varepsilon/m)$. From Table 2 it follows that this requires sample size superlinear in m, which is suboptimal. Interestingly, when they take d to be the square Hellinger distance, they can establish a localization-of-distance result with $\varepsilon' = \varepsilon^2/2m$. By Table 2, to test each S they need sample complexity that is linear in m, leading to an overall dependence of the sample complexity on m that is $\tilde{O}(m)$,[3] which is optimal up to log factors. Again, switching to a different distance results in near-optimal overall sample complexity, and our table is guidance as to where the bottlenecks and opportunities lie.

Finally, we comment that tolerant testing (i.e., when $\varepsilon_1 > 0$) is perhaps one of the most interesting questions in the design of practically useful testers. Indeed, as mentioned before, in many statistical settings there may be model misspecification. For example, why should one expect to be receiving samples from *precisely* the uniform distribution? As such, one may desire that a tester is *robust* to small errors, and accepts all distributions which are *close* to uniform. Unfortunately, Valiant and Valiant [VV11a] ruled out the possibility of a strongly sublinear tester which has total variation tolerance, showing that such a problem requires $\Theta\left(\frac{n}{\log n}\right)$ samples. However, as shown by Acharya, Daskalakis, and Kamath [ADK15], χ^2-tolerance is possible with only $O\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ samples. This raises the following question: Which distances can a tester be tolerant to, while maintaining a strongly sublinear sample complexity? We outline what is possible.

[3]The extra log factors are to guarantee that the tests performed on all sets S of size $\delta + 1$ succeed.

	$d_{\mathrm{TV}}(p,q) \geq \varepsilon$	$d_{\mathrm{H}}(p,q) \geq \varepsilon/\sqrt{2}$	$d_{\mathrm{KL}}(p,q) \geq \varepsilon^2$	$d_{\chi^2}(p,q) \geq \varepsilon^2$
$p = q$	$\Omega\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$[Pan08]		Untestable [Theorem 7]	
$d_{\chi^2}(p,q) \leq \varepsilon^2/4$		$O\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$[Theorem 1]		
$d_{\mathrm{KL}}(p,q) \leq \varepsilon^2/4$	$\Omega\left(\frac{n}{\log n}\right)$[Theorem 8]			
$d_{\mathrm{H}}(p,q) \leq \varepsilon/2\sqrt{2}$				
$d_{\mathrm{TV}}(p,q) \leq \varepsilon/2$ or $\varepsilon^2/4$[5]		$O\left(\frac{n}{\log n}\right)$[Corollary 3]		

Table 1: Identity Testing. Rows correspond to completeness of the tester, and columns correspond to soundness.

	$d_{\mathrm{TV}}(p,q) \geq \varepsilon$	$d_{\mathrm{H}}(p,q) \geq \varepsilon/\sqrt{2}$	$d_{\mathrm{KL}}(p,q) \geq \varepsilon^2$	$d_{\chi^2}(p,q) \geq \varepsilon^2$
$p = q$	$O\left(\max\left\{\frac{n^{1/2}}{\varepsilon^2}, \frac{n^{2/3}}{\varepsilon^{4/3}}\right\}\right)$[CDVV14] $\Omega\left(\max\left\{\frac{n^{1/2}}{\varepsilon^2}, \frac{n^{2/3}}{\varepsilon^{4/3}}\right\}\right)$[CDVV14]	$O\left(\min\left\{\frac{n^{3/4}}{\varepsilon^2}, \frac{n^{2/3}}{\varepsilon^{8/3}}\right\}\right)$[Theorem 5] $\Omega\left(\min\left\{\frac{n^{3/4}}{\varepsilon^2}, \frac{n^{2/3}}{\varepsilon^{8/3}}\right\}\right)$[DK16]	Untestable [Theorem 7]	
$d_{\chi^2}(p,q) \leq \varepsilon^2/4$	$\Omega\left(\frac{n}{\log n}\right)$ [Theorem 9]			
$d_{\mathrm{KL}}(p,q) \leq \varepsilon^2/4$				
$d_{\mathrm{H}}(p,q) \leq \varepsilon/2\sqrt{2}$				
$d_{\mathrm{TV}}(p,q) \leq \varepsilon/2$ or $\varepsilon^2/4$[5]		$O\left(\frac{n}{\log n}\right)$[Corollary 3]		

Table 2: Equivalence Testing. Rows correspond to completeness of the tester, and columns correspond to soundness.

1.1 Results

Our results are pictorially presented in Tables 1, 2, and 3. We note that these tables are intended to provide only references to the *sample complexity* of each testing problem, rather than exhaustively cover all prior work. As such, several references are deferred to Section 1.2. In Tables 1 and 2, each cell contains the complexity of testing whether two distributions are close in the distance for that row, versus far in the distance for that column.[4] These distances and their relationships are covered in detail in Section 2, but we note that the distances are scaled and transformed such that problems become harder as we traverse the table down or to the right. In other words, lower bounds hold for cells which are down or to the right in the table, and upper bounds hold for cells which are up or to the left; problems with the same complexity are shaded with the same color. The dark grey boxes indicate problems which are not well-defined, i.e. two distributions could simultaneously be close in KL and far in χ^2-divergence.

We highlight some of our results:

1. We give an $O(\sqrt{n}/\varepsilon^2)$ sample algorithm for identity testing whether $d_{\chi^2}(p,q) \leq \varepsilon^2/4$ or $d_{\mathrm{H}}(p,q) \geq \varepsilon/\sqrt{2}$ (Theorem 1). This is the first algorithm which achieves the optimal dependence on both n and ε for identity testing with respect to Hellinger distance (even non-tolerantly). We note that a $O(\sqrt{n}/\varepsilon^4)$ algorithm was known, due to optimal identity testers for total variation distance and the quadratic relationship between total variation and Hellinger distance.

2. In the case of identity testing, a stronger form of tolerance (i.e., KL divergence instead of χ^2) causes the sample complexity to jump to $\Omega(n/\log n)$ (Theorem 8). We find this a bit surprising, as χ^2-divergence is the second-order Taylor expansion of KL divergence, so one might expect that the testing problems have comparable complexities.

3. In the case of equivalence testing, *even* χ^2*-tolerance* comes at the cost of an $\Omega(n/\log n)$ sample complexity (Theorem 9). This is a qualitative difference from identity testing, where χ^2-tolerance came at no cost.

[4]Note that we chose constants in our theorem statements for simplicity of presentation, and they may not match the constants presented in the table. This can be remedied by appropriate changing of constants in the algorithms and constant factor increases in the sample complexity.

[5]We note that we must use $\varepsilon/2$ or $\varepsilon^2/4$ depending on whether we are testing with respect to TV or Hellinger. For more details and other discussion of the $n/\log n$ region of this chart, see Section 1.1.2.

	Identity Testing	Equivalence Testing
$d(p,q) \leq f_d(n,\varepsilon)$ vs. $d_{\ell_2}(p,q) \geq \varepsilon$	$\Theta\left(\frac{1}{\varepsilon^2}\right)$ [Corollary 2]	$\Theta\left(\frac{1}{\varepsilon^2}\right)$ [Corollary 2]
$d_{\ell_2}(p,q) \leq \frac{\varepsilon}{\sqrt{n}}$ vs. $d_{\mathrm{TV}}(p,q) \geq \varepsilon$	$\Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ [Theorem 2]	$\Theta\left(\max\left\{\frac{n^{1/2}}{\varepsilon^2}, \frac{n^{2/3}}{\varepsilon^{4/3}}\right\}\right)$ [Theorem 4]
$d_{\ell_2}(p,q) \leq \frac{\varepsilon^2}{\sqrt{n}}$ vs. $d_{\mathrm{H}}(p,q) \geq \varepsilon$	$\Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ [Theorem 3]	$\Theta\left(\min\left\{\frac{n^{3/4}}{\varepsilon^2}, \frac{n^{2/3}}{\varepsilon^{8/3}}\right\}\right)$ [Theorem 5]

Table 3: ℓ_2 Testing. $f_d(n,\varepsilon)$ is a quantity such that $d(p,q) \leq f_d(n,\varepsilon)$ and $d_{\ell_2}(p,q) \geq \varepsilon$ are disjoint.

4. However, in both identity and equivalence testing, ℓ_2 tolerance comes at no additional cost (Theorems 2, 3, 4, and 5). Thus, in many cases, ℓ_2 tolerance is the best one can do if one wishes to maintain a strongly sublinear sample complexity.

From a technical standpoint, our algorithms are χ^2-statistical tests, and most closely resemble those of [ADK15] and [CDVV14] (similar χ^2-tests were employed in [VV17, DKN15, CDGR16]). However, crucial changes are required to satisfy the more stringent requirements of testing with respect to Hellinger distance. In our identity tester for Hellinger, we deal with this different distance measure by pruning light domain elements of q less aggressively than [ADK15], in combination with a preliminary test to reject early if the difference between p and q is contained exclusively within the set of light elements – this is a new issue that cannot arise when testing with respect to total variation distance. In our equivalence tester for Hellinger, we follow an approach, similar to [CDVV14] and [DK16], of analyzing the light and heavy domain elements separately, with the challenge that the algorithm does not know which elements are which. Finally, to achieve ℓ_2 tolerance in these cases, we use a "mixing" strategy in which instead of testing based solely on samples from p and q, we mix in some number (depending on our application) of samples from the uniform distribution. At a high level, the purpose of mixing is to make our distributions *well-conditioned*, i.e. to ensure that all probability values are sufficiently large. Such a strategy was recently employed by Goldreich in [Gol16] for uniformity testing.

1.1.1 Comments on ℓ_2-tolerance

ℓ_2 tolerance has been indirectly considered in [GR00, BFF$^+$01, BFR$^+$13] through their weak tolerance for total variation distance and the relationship with ℓ_2 distance, though these results have suboptimal sample complexity. Our equivalence testing results improve upon [CDVV14] by adding ℓ_2-tolerance. We note that [DK16] also provides ℓ_2-tolerant testers (as well as [DKN15] for the case of uniformity), comparable to those obtained in Theorems 2, 3, and 5, though this tolerance is not explicitly analyzed in their paper. This can be seen by noting that the underlying tester from [CDVV14] is tolerant, and the "flattening" operation they apply reduces the ℓ_2-distance between the distributions. The testers in [DK16] are those of Propositions 2.7, 2.10, and 2.15, combined with the observation of Remark 2.8. We rederive these results for completeness, and to show a direct way of proving ℓ_2-tolerance. Note that Theorem 5 also improves upon Proposition 2.15 of [DK16] by removing log factors in the sample complexity.

1.1.2 Comments on the $\Theta(n/\log n)$ Results

Our upper bounds in the bottom-left portion of the table are based off the total variation distance estimation algorithm of Jiao, Han, and Weissman [JHW16], where an $\Theta(n/\log n)$ complexity is only derived for $\varepsilon \geq 1/\operatorname{poly}(n)$. Similarly, in [VV10a], the lower bounds are only valid for constant ε. We believe that the precise characterization is a very interesting open problem. In the present work, we focus on the case of constant ε for these testing problems.

We wish to draw attention to the bottom row of the table, and note that the two testing problems are $d_{\mathrm{TV}}(p,q) \leq \varepsilon/2$ versus $d_{\mathrm{TV}}(p,q) \geq \varepsilon$, and $d_{\mathrm{TV}}(p,q) \leq \varepsilon^2/4$ versus $d_{\mathrm{H}}(p,q) \geq \varepsilon/\sqrt{2}$. This difference in parameterization is required to make the two cases in the testing problem disjoint. With this parameterization, we conjecture that the latter problem has a greater dependence on ε as it goes to 0 (namely, ε^{-4} versus ε^{-2}), so we colour the box a slightly darker shade of orange.

1.2 Related Work

The most classic distribution testing question is uniformity testing, which is identity testing when $\varepsilon_1 = 0$, d_2 is total variation distance, and q is the uniform distribution. This was first studied in theoretical computer science in [GR00]. Paninski gave an optimal algorithm (for when ε_2 is not too small) with a complexity of $O(\sqrt{n}/\varepsilon^2)$ and a matching lower bound [Pan08]. More generally, letting q be an arbitrary distribution, exact total variation identity testing was studied [BFF$^+$01], and an (instance) optimal algorithm was given by Valiant and Valiant [VV17], with the same complexity of $O(\sqrt{n}/\varepsilon^2)$. Optimal algorithms for this problem were rediscovered several times, see i.e. [DKN15, ADK15, DK16, DGPP16].

Equivalence (or closeness) testing was studied in [BFR$^+$13], in the same setting ($\varepsilon_1 = 0$, d_2 is total variation distance). A lower bound of $\Omega(n^{2/3})$ was given by [Val11]. Tight upper and lower bounds were given in [CDVV14], which shows interesting behavior of the sample complexity as the parameter ε goes from large to small. This problem was also studied in the setting where one has unequal sample sizes from the two distributions [BV15, DK16]. When the distance d_1 is Hellinger, the complexity is qualitatively different, as shown by [DK16]. They prove a nearly-optimal upper bound and a tight lower bound for this problem.

[Wag15, DBNNR11] also consider testing problems with other distances, namely ℓ_p distances and earth mover's distance (also known as Wasserstein distance), respectively.

Tolerant identity testing (where $\varepsilon_1 = O(\varepsilon)$ and d_1 is total variation distance) was studied in [VV10a, VV10b, VV11a, VV11b], through the (equivalent) lens of estimating total variation distance between distributions. In these works, $\Theta\left(n/\log n\right)$ bounds were proven for the sample complexity. Several other related problems (i.e., support size and entropy estimation) share the same sample complexity, and have enjoyed significant study [AOST17, WY16, ADOS17]. The closest related results to our work are those on estimating distances between distributions [JHW16, JVHW17, HJW16].

χ^2-tolerance (when d_1 is χ^2-divergence and $\varepsilon_1 = O(\varepsilon^2)$) was introduced and applied by [ADK15] for testing families of distributions, i.e., testing if a distribution is monotone or far from being monotone. It was shown that this tolerance comes at no additional cost over vanilla identity testing; that is, the sample complexity is still $O(\sqrt{n}/\varepsilon^2)$. Testing such families of distributions was also studied by [CDGR16].

Testing with respect to Hellinger distance was applied in [DP17] for testing Bayes networks. Since lower bounds of [ADK15] show that distribution testing suffers from the curse of dimensionality, further structural assumptions must be made if one wishes to test multivariate distributions. This "high-dimensional frontier" has also been studied on graphical models by [DDK18] and [CDKS17] (for Ising models and Bayesian networks, respectively).

Our work focuses on characterizing the complexity of identity and equivalence testing in the worst case over pairs p and q. Related works attempt to nail down the sample complexity of identity testing on an *instance-by-instance* basis [VV17, JHW16, DK16, BCG17] – that is, reducing the sample complexity depending on which distribution q is given as input (and sometimes depending on p as well). We consider this to be an interesting open question for different distances d_1 and d_2. For example, Theorem 7 states that identity testing is impossible when d_2 is the KL divergence. However, if q is the uniform distribution, then the complexity becomes $\Theta(\sqrt{n})$. An instance-by-instance analysis would allow one to bypass some of these strong lower bounds.

This is only a fraction of recent results; we direct the reader to [Can15] for an excellent recent survey of distribution testing.

1.3 Organization

The organization of this paper is as follows. In Section 2, we state preliminaries and notation used in this paper. In Sections 3 and 4, we prove upper bounds for identity testing and equivalence testing (respectively) based on χ^2-style statistics. In Section 5, we prove upper bounds for distribution testing based on distance estimation. Finally, in Section 6, we prove testing lower bounds.

2 Preliminaries

In this paper, we will focus on discrete probability distributions over $[n]$. For a distribution p, we will use the notation p_i to denote the mass p places on symbol i. For a set $S \subseteq [n]$ and a distribution p over $[n]$, p_S

is the vector p restricted to the coordinates in S. We will call this a *restriction* of distribution p.

The following probability distances and divergences are of interest to us:

Definition 1. *The* total variation distance *between p and q is defined as*

$$d_{\mathrm{TV}}(p,q) = \max_{S\subseteq[n]} p(S) - q(S) = \frac{1}{2}\sum_{i\in[n]} |p_i - q_i| = \|p-q\|_1 \in [0,1].$$

Definition 2. *The* KL divergence *between p and q is defined as*

$$d_{\mathrm{KL}}(p,q) = \sum_{i\in[n]} p_i \log\left(\frac{p_i}{q_i}\right) \in [0,\infty).$$

This definition uses the convention that $0\log 0 = 0$.

Definition 3. *The* Hellinger distance *between p and q is defined as*

$$d_{\mathrm{H}}(p,q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i\in[n]} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2} \in [0,1].$$

Definition 4. *The χ^2-*divergence *(or* chi-squared divergence*) between p and q is defined as*

$$d_{\chi^2}(p,q) = \sum_{i\in[n]} \frac{(p_i - q_i)^2}{q_i} \in [0,\infty).$$

Definition 5. *The ℓ_2* distance *between p and q is defined as*

$$d_{\ell_2}(p,q) = \sqrt{\sum_{i\in[n]} (p_i - q_i)^2} = \|p-q\|_2 \in [0,1].$$

We also define these distances for restrictions of distributions p_S and q_S by replacing the summations over $i \in [n]$ with summations over $i \in S$.

We have the following relationships between these distances. These are well-known for distributions, i.e., see [GS02], but we prove them more generally for restrictions of distributions in Section A.

Proposition 1. *Letting p_S and q_S be restrictions of distributions p and q to $S \subseteq [n]$,*

$$d_{\mathrm{H}}^2(p_S,q_S) \le d_{\mathrm{TV}}(p_S,q_S) \le \sqrt{2}d_{\mathrm{H}}(p_S,q_S) \le \sqrt{\sum_{i\in S}(q_i - p_i) + d_{\mathrm{KL}}(p_S,q_S)} \le \sqrt{d_{\chi^2}(p_S,q_S)}.$$

We recall that d_{ℓ_2} fits into the picture by its relationship with total variation distance:

Proposition 2. *Letting p and q be distributions over $[n]$,*

$$d_{\ell_2}(p,q) \le 2d_{\mathrm{TV}}(p,q) \le \sqrt{n}d_{\ell_2}(p,q).$$

The second inequality follows from Cauchy-Schwarz.

We will also need to following bound for Hellinger distance:

Proposition 3. $2d_{\mathrm{H}}^2(p,q) \le \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{p_i + q_i} \le 4d_{\mathrm{H}}^2(p,q).$

Proof. Expanding the Hellinger-squared distance,

$$d_{\mathrm{H}}^2(p,q) = \frac{1}{2}\sum_{i=1}^{n}(\sqrt{p_i} - \sqrt{q_i})^2 = \frac{1}{2}\sum_{i=1}^{n}\frac{(p_i - q_i)^2}{(\sqrt{p_i} + \sqrt{q_i})^2}.$$

The fact now follows because $(p_i + q_i) \le (\sqrt{p_i} + \sqrt{q_i})^2 \le 2(p_i + q_i)$. □

The quantity $\sum_{i=1}^{n}(p_i - q_i)^2/(p_i + q_i)$ is sometimes called the *triangle distance*. However, we see here that it is essentially the Hellinger distance (up to constant factors).

Proposition 4. *Given a number $\delta \in [0,1]$ and a discrete distribution $r = (r_1, \ldots, r_n)$, define*

$$r^{+\delta} := (1-\delta) \cdot r + \delta \cdot (\tfrac{1}{n}, \ldots, \tfrac{1}{n}).$$

Then given two discrete distributions $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$,

$$d_{\mathrm{TV}}(p^{+\delta}, q^{+\delta}) = (1-\delta) d_{\mathrm{TV}}(p,q), \quad d_{\ell_2}(p^{+\delta}, q^{+\delta}) = (1-\delta) d_{\ell_2}(p,q).$$

In addition, $d_{\mathrm{H}}(p^{+\delta}, q^{+\delta}) \geq d_{\mathrm{H}}(p,q) - 2\sqrt{\delta}$.

Proof. The statements for total variation and ℓ_2 distance are immediate. As for the Hellinger distance, we have by the triangle inequality that

$$d_{\mathrm{H}}(p,q) \leq d_{\mathrm{H}}(p, p^{+\delta}) + d_{\mathrm{H}}(p^{+\delta}, q^{+\delta}) + d_{\mathrm{H}}(q^{+\delta}, q).$$

We can bound the first term by

$$d_{\mathrm{H}}^2(p, p^{+\delta}) \leq d_{\mathrm{TV}}(p, p^{+\delta}) = \tfrac{1}{2} \cdot \|\delta \cdot p - \delta \cdot (\tfrac{1}{n}, \ldots, \tfrac{1}{n})\|_1 \leq \delta,$$

where the last step is by the triangle inequality, and a similar argument bounds the third term by $\sqrt{\delta}$ as well. Thus, $d_{\mathrm{H}}(p^{+\delta}, q^{+\delta}) \geq d_{\mathrm{H}}(p,q) - 2\sqrt{\delta}$. □

A similar technique was employed in [Gol16].

At times, our algorithms will employ *Poisson sampling*. Instead of taking m samples from a distribution p, we instead take Poisson(m) samples. As a result, letting N_i be the number of occurences of symbol i, all N_i will be independent and distributed as Poisson($m \cdot p_i$). We note that this method of sampling is for purposes of analysis – concentration bounds imply that $\mathrm{Poi}(m) = O(m)$ with high probability, so such an algorithm can be converted to one with a fixed budget of samples at a constant-factor increase in the sample complexity.

3 Upper Bounds for Identity Testing

In this section, we prove the following theorems for identity testing.

Theorem 1. *There exists an algorithm for identity testing between p and q distinguishing the cases:*

- $d_{\chi^2}(p,q) \leq \varepsilon^2$;
- $d_{\mathrm{H}}(p,q) \geq \varepsilon$.

The algorithm uses $O\left(\frac{n^{1/2}}{\varepsilon^2}\right)$ samples.

Theorem 2. *There exists an algorithm for identity testing between p and q distinguishing the cases:*

- $d_{\ell_2}(p,q) \leq \frac{\varepsilon}{\sqrt{n}}$;
- $d_{\mathrm{TV}}(p,q) \geq \varepsilon$.

The algorithm uses $O\left(\frac{n^{1/2}}{\varepsilon^2}\right)$ samples.

Theorem 3. *There exists an algorithm for identity testing between p and q distinguishing the cases:*

- $d_{\ell_2}(p,q) \leq \frac{\varepsilon^2}{\sqrt{n}}$;
- $d_{\mathrm{H}}(p,q) \geq \varepsilon$.

The algorithm uses $O\left(\frac{n^{1/2}}{\varepsilon^2}\right)$ samples.

We prove Theorem 1 in Section 3.1, and Theorems 2 and 3 in Section 3.2.

3.1 Identity Testing with Hellinger Distance and χ^2-Tolerance

We prove Theorem 1 by analyzing Algorithm 1. We will set $c_1 = \frac{1}{100}, c_2 = \frac{6}{25}$, and let C be a sufficiently large constant.

Algorithm 1 χ^2-close versus Hellinger-far testing algorithm

1: **Input:** ε; an explicit distribution q; sample access to a distribution p
2: Implicitly define $\mathcal{A} \leftarrow \{i : q_i \geq c_1\varepsilon^2/n\}$, $\bar{\mathcal{A}} \leftarrow [n] \setminus \mathcal{A}$
3: Let $\hat{p}$ be the empirical distribution[6] from drawing $m_1 = \Theta(1/\varepsilon^2)$ samples from p
4: **if** $\hat{p}(\bar{\mathcal{A}}) \geq \frac{3}{4}c_2\varepsilon^2$ **then**
5: **return** REJECT
6: **end if**
7: Draw a multiset S of Poisson(m_2) samples from p, where $m_2 = C\sqrt{n}/\varepsilon^2$
8: Let N_i be the number of occurrences of the ith domain element in S
9: Let S' be the set of domain elements observed in S
10: $Z \leftarrow \sum_{i \in S' \cap \mathcal{A}} \frac{(N_i - m_2 q_i)^2 - N_i}{m_2 q_i} + m_2(1 - q(S' \cap \mathcal{A}))$
11: **if** $Z \leq \frac{3}{2}m_2\varepsilon^2$ **then**
12: **return** ACCEPT
13: **else**
14: **return** REJECT
15: **end if**

We note that the sample and time complexity are both $O(\sqrt{n}/\varepsilon^2)$. We draw $m_1 + m_2 = \Theta(\sqrt{n}/\varepsilon^2)$ samples total. All steps of the algorithm only involve inspecting domain elements where a sample falls, and it runs linearly in the number of such elements. Indeed, Step 10 of the algorithm is written in an unusual way in order to ensure the running time of the algorithm is linear.

We first analyze the test in Step 4 of the algorithm. Folklore results state that with probability at least 99/100, this preliminary test will reject any p with $p(\bar{\mathcal{A}}) \geq c_2\varepsilon^2$, it will not reject any p with $p(\bar{\mathcal{A}}) \leq \frac{c_2}{2}\varepsilon^2$, and behavior for any other p is arbitrary. Condition on the event the test does not reject for the remainder of the proof. Note that since both thresholds here are $\Theta(\varepsilon^2)$, it only requires $m_1 = \Theta(1/\varepsilon^2)$ samples, rather than the "non-extreme" regime, where we would require $\Theta(1/\varepsilon^4)$ samples.

Remark 1. *We informally refer to this "extreme" versus "non-extreme" regime in distribution testing. To give an example of what we mean in these two cases, consider distinguishing $Ber(1/2)$ from $Ber(1/2 + \varepsilon)$. The complexity of this problem is $\Theta(1/\varepsilon^2)$, and we consider this to be in the non-extreme regime. On the other hand, distinguishing $Ber(\varepsilon)$ from $Ber(2\varepsilon)$ has a sample complexity of $\Theta(1/\varepsilon)$, and we consider this to be in the extreme regime.*

We justify that any p which may be rejected in Step 5 (i.e., any p such that $p(\bar{\mathcal{A}}) > \frac{c_2}{2}\varepsilon^2$) has the property that $d_{\chi^2}(p, q) > \varepsilon^2$ (in other words, we do not wrongfully reject any p).

Consider a p such that $p(\bar{\mathcal{A}}) \geq \frac{c_2}{2}\varepsilon^2$. Note that $d_{\chi^2}(p, q) \geq d_{\chi^2}(p_{\bar{\mathcal{A}}}, q_{\bar{\mathcal{A}}})$, which we lower bound as follows:

$$\begin{aligned}
d_{\chi^2}(p_{\bar{\mathcal{A}}}, q_{\bar{\mathcal{A}}}) &= \sum_{i \in \bar{\mathcal{A}}} \frac{(p_i - q_i)^2}{q_i} \\
&\geq \frac{n}{c_1\varepsilon^2} \sum_{i \in \bar{\mathcal{A}}} (p_i - q_i)^2 \\
&\geq \frac{n}{c_1\varepsilon^2} \cdot \frac{1}{n} \left(\sum_{i \in \bar{\mathcal{A}}} (p_i - q_i) \right)^2 \\
&\geq \frac{n}{c_1\varepsilon^2} \frac{\varepsilon^4 \left(\frac{c_2}{2} - c_1\right)^2}{n} \\
&= \frac{\left(\frac{c_2}{2} - c_1\right)^2}{c_1} \varepsilon^2
\end{aligned}$$

The first inequality is by the definition of $\bar{\mathcal{A}}$, the second is by Cauchy-Schwarz, and the third is since $p(\bar{\mathcal{A}}) \geq \frac{c_2}{2}\varepsilon^2$ and $q(\bar{\mathcal{A}}) \leq c_1\varepsilon^2$. By our setting of c_1 and c_2, this implies that $d_{\chi^2}(p,q) > \varepsilon^2$, and we are not rejecting any p which should be accepted.

For the remainder of the proof, we will implicitly assume that $p(\bar{\mathcal{A}}) \leq c_2\varepsilon^2$.

Let

$$Z' = \sum_{i \in \mathcal{A}} \frac{(N_i - m_2 q_i)^2 - N_i}{m_2 q_i}.$$

Note that the statistic Z can be rewritten as follows:

$$\begin{aligned}
Z &= \sum_{i \in S' \cap \mathcal{A}} \frac{(N_i - m_2 q_i)^2 - N_i}{m_2 q_i} + m_2(1 - q(S' \cap \mathcal{A})) \\
&= \sum_{i \in S' \cap \mathcal{A}} \frac{(N_i - m_2 q_i)^2 - N_i}{m_2 q_i} + \sum_{i \in \mathcal{A} \setminus S'} m_2 q_i + m_2 q(\bar{\mathcal{A}}) \\
&= \sum_{i \in S' \cap \mathcal{A}} \frac{(N_i - m_2 q_i)^2 - N_i}{m_2 q_i} + \sum_{i \in \mathcal{A} \setminus S'} \frac{(N_i - m_2 q_i)^2 - N_i}{m_2 q_i} + m_2 q(\bar{\mathcal{A}}) \\
&= Z' + m_2 q(\bar{\mathcal{A}})
\end{aligned}$$

We proceed by analyzing Z'. First, note that it has the following expectation and variance:

$$\mathbf{E}[Z'] = m_2 \cdot \sum_{i \in \mathcal{A}} \frac{(p_i - q_i)^2}{q_i} = m_2 \cdot d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) \tag{1}$$

$$\mathbf{Var}[Z'] = \sum_{i \in \mathcal{A}} \left[2\frac{p_i^2}{q_i^2} + 4m_2 \cdot \frac{p_i \cdot (p_i - q_i)^2}{q_i^2} \right] \tag{2}$$

These properties are proven in Section A of [ADK15].

We require the following two lemmas, which state that the mean of the statistic is separated in the two cases, and that the variance is bounded. The proofs largely follow the proofs of two similar lemmas in [ADK15].

Lemma 1. *If $d_{\chi^2}(p,q) \leq \varepsilon^2$, then $\mathbf{E}[Z'] \leq m_2\varepsilon^2$. If $d_{\mathrm{H}}(p,q) \geq \varepsilon$, then $\mathbf{E}[Z'] \geq (2 - c_1 - c_2)m_2\varepsilon^2$.*

Proof. The former case is immediate from (1).

For the latter case, note that

$$d_{\mathrm{H}}^2(p,q) = d_{\mathrm{H}}^2(p_{\mathcal{A}}, q_{\mathcal{A}}) + d_{\mathrm{H}}^2(p_{\bar{\mathcal{A}}}, q_{\bar{\mathcal{A}}}).$$

We upper bound the latter term as follows:

$$\begin{aligned}
d_{\mathrm{H}}^2(p_{\bar{\mathcal{A}}}, q_{\bar{\mathcal{A}}}) &\leq d_{\mathrm{TV}}(p_{\bar{\mathcal{A}}}, q_{\bar{\mathcal{A}}}) \\
&= \frac{1}{2} \sum_{i \in \bar{\mathcal{A}}} |p_i - q_i| \\
&\leq \frac{1}{2} \left(p(\bar{\mathcal{A}}) + q(\bar{\mathcal{A}}) \right) \\
&\leq \left(\frac{c_1 + c_2}{2} \right) \varepsilon^2
\end{aligned}$$

The first inequality is from Proposition 1, and the third inequality is from our prior condition that $p(\bar{\mathcal{A}}) \leq c_2\varepsilon^2$.

Since $d_{\mathrm{H}}^2(p,q) \geq \varepsilon^2$, this implies $d_{\mathrm{H}}^2(p_{\mathcal{A}}, q_{\mathcal{A}}) \geq \left(1 - \frac{c_1+c_2}{2}\right)\varepsilon^2$. Proposition 1 further implies that $d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) \geq (2 - c_1 - c_2)\,\varepsilon^2$. The lemma follows from (1). □

Lemma 2. *If $d_{\chi^2}(p,q) \le \varepsilon^2$, then $\mathbf{Var}[Z'] = O(m_2^2\varepsilon^4)$. If $d_{\mathrm{H}}(p,q) \ge \varepsilon$, then $\mathbf{Var}[Z'] \le O(\mathbf{E}[Z']^2)$. The constant in both expressions can be made arbitrarily small with the choice of the constant C.*

Proof. We bound the terms of (2) separately, starting with the first.

$$\begin{aligned}
2\sum_{i\in\mathcal{A}}\frac{p_i^2}{q_i^2} &= 2\sum_{i\in\mathcal{A}}\left(\frac{(p_i-q_i)^2}{q_i^2}+\frac{2p_iq_i-q_i^2}{q_i^2}\right)\\
&= 2\sum_{i\in\mathcal{A}}\left(\frac{(p_i-q_i)^2}{q_i^2}+\frac{2q_i(p_i-q_i)+q_i^2}{q_i^2}\right)\\
&\le 2n + 2\sum_{i\in\mathcal{A}}\left(\frac{(p_i-q_i)^2}{q_i^2}+2\frac{(p_i-q_i)}{q_i}\right)\\
&\le 4n + 4\sum_{i\in\mathcal{A}}\frac{(p_i-q_i)^2}{q_i^2}\\
&\le 4n + \frac{4n}{c_1\varepsilon^2}\sum_{i\in\mathcal{A}}\frac{(p_i-q_i)^2}{q_i}\\
&= 4n + \frac{4n}{c_1\varepsilon^2}\frac{E[Z']}{m_2}\\
&\le 4n + \frac{4}{c_1C}\sqrt{n}E[Z'] \qquad (3)
\end{aligned}$$

The second inequality is the AM-GM inequality, the third inequality uses that $q_i \ge \frac{c_1\varepsilon^2}{n}$ for all $i \in \mathcal{A}$, the last equality uses (1), and the final inequality substitutes a value $m_2 \ge C\frac{\sqrt{n}}{\varepsilon^2}$.

The second term can be similarly bounded:

$$\begin{aligned}
4m_2\sum_{i\in\mathcal{A}}\frac{p_i(p_i-q_i)^2}{q_i^2} &\le 4m_2\left(\sum_{i\in\mathcal{A}}\frac{p_i^2}{q_i^2}\right)^{1/2}\left(\sum_{i\in\mathcal{A}}\frac{(p_i-q_i)^4}{q_i^2}\right)^{1/2}\\
&\le 4m_2\left(4n+\frac{4}{c_1C}\sqrt{n}E[Z']\right)^{1/2}\left(\sum_{i\in\mathcal{A}}\frac{(p_i-q_i)^4}{q_i^2}\right)^{1/2}\\
&\le 4m_2\left(2\sqrt{n}+\frac{2}{\sqrt{c_1C}}n^{1/4}E[Z']^{1/2}\right)\left(\sum_{i\in\mathcal{A}}\frac{(p_i-q_i)^2}{q_i}\right)\\
&= \left(8\sqrt{n}+\frac{8}{\sqrt{c_1C}}n^{1/4}E[Z']^{1/2}\right)E[Z'].
\end{aligned}$$

The first inequality is Cauchy-Schwarz, the second inequality uses (3), the third inequality uses the monotonicity of the ℓ_p norms, and the equality uses (1).

Combining the two terms, we get

$$\mathbf{Var}[Z'] \le 4n + \left(8+\frac{4}{c_1C}\right)\sqrt{n}\mathbf{E}[Z'] + \frac{8}{\sqrt{c_1C}}n^{1/4}\mathbf{E}[Z']^{3/2}.$$

We now consider the two cases in the statement of our lemma.

- When $d_{\chi^2}(p,q) \le \varepsilon^2$, we know from Lemma 1 that $\mathbf{E}[Z'] \le m_2\varepsilon^2$. Combined with a choice of $m_2 \ge C\frac{\sqrt{n}}{\varepsilon^2}$ and the above expression for the variance, this gives:

$$\begin{aligned}
\mathbf{Var}[Z'] &\le \frac{4}{C^2}m_2^2\varepsilon^4 + \left(\frac{8}{C}+\frac{4}{c_1C^2}\right)m_2^2\varepsilon^4 + \frac{8}{C\sqrt{c_1}}m_2^2\varepsilon^4\\
&= \left(\frac{8}{C}+\frac{8}{C\sqrt{c_1}}+\frac{4}{C^2}+\frac{4}{c_1C^2}\right)m_2^2\varepsilon^4 = O(m_2^2\varepsilon^4).
\end{aligned}$$

- When $d_{\mathrm{H}}(p,q) \geq \varepsilon$, Lemma 1 and $m_2 \geq C\frac{\sqrt{n}}{\varepsilon^2}$ give:

$$\mathbf{E}[Z'] \geq (2 - c_1 - c_2)m_2\varepsilon^2 \geq C(2 - c_1 - c_2)\sqrt{n}.$$

Similar to before, combining this with our expression for the variance we get:

$$\begin{aligned}\mathbf{Var}[Z'] &\leq \left(\frac{8}{C(2-c_1-c_2)} + \frac{8}{C\sqrt{c_1(2-c_1-c_2)}} + \frac{4}{C^2(2-c_1-c_2)^2} + \frac{4}{C^2c_1(2-c_1-c_2)}\right)\mathbf{E}[Z']^2 \\ &= O(\mathbf{E}[Z']^2).\end{aligned}$$

□

To conclude the proof, we consider the two cases.

- Suppose $d_{\chi^2}(p,q) \leq \varepsilon^2$. By Lemma 1 and the definition of $\mathcal{A}$, we have that $\mathbf{E}[Z] \leq (1 + c_1)m_2\varepsilon^2$. By Lemma 2, $\mathbf{Var}[Z] = O(m_2^2\varepsilon^4)$. Therefore, for constant C sufficiently large, Chebyshev's inequality implies $\Pr(Z > \frac{3}{2}m_2\varepsilon^2) \leq 1/10$.
- Suppose $d_{\mathrm{H}}(p,q) \geq \varepsilon$. By Lemma 1, we have that $\mathbf{E}[Z'] \geq (2 - c_1 - c_2)m_2\varepsilon^2$. By Lemma 2, $\mathbf{Var}[Z'] = O(\mathbf{E}[Z']^2)$. Therefore, for constant C sufficiently large, Chebyshev's inequality implies $\Pr(Z' < \frac{3}{2}m_2\varepsilon^2) \leq 1/10$. Since $Z \geq Z'$, $\Pr(Z < \frac{3}{2}m_2\varepsilon^2) \leq 1/10$ as well.

3.2 Identity Testing with ℓ_2 Tolerance

In this section, we sketch the algorithms required to achieve ℓ_2 tolerance for identity testing. Since the algorithms and analysis are very similar to those of Algorithm 1 of [ADK15] and Algorithm 1, the full details are omitted.

First, we prove Theorem 2. The algorithm is Algorithm 1 of [ADK15], but instead of testing on p and q, we instead test on $p^{+\frac{1}{2}}$ and $q^{+\frac{1}{2}}$, as defined in Proposition 4. By this proposition, this operation preserves total variation and ℓ_2 distance, up to a factor of 2, and also makes it so that the minimum probability element of $q^{+\frac{1}{2}}$ is at least $1/2n$. In the case where $d_{\ell_2}(p,q) \leq \frac{\varepsilon}{\sqrt{n}}$, we have the following upper bound on $\mathbf{E}[Z]$:

$$\mathbf{E}[Z'] = m\sum_{i\in\bar{\mathcal{A}}}\frac{(p_i - q_i)^2}{q_i} \leq O\left(m \cdot n \cdot d_{\ell_2}^2(p,q)\right) \leq O(m_2\varepsilon^2).$$

This is the same bound as in Lemma 2 of [ADK15]. The rest of the analysis follows identically to that of Algorithm 1 of [ADK15], giving us Theorem 2.

Next, we prove Theorem 3. We observe that Algorithm 1 as stated can be considered as ℓ_2-tolerant instead of χ^2-tolerant, if desired. First, we do not wrongfully reject any p (i.e., those with $d_{\ell_2}(p,q) \leq \frac{\varepsilon^2}{\sqrt{n}}$) in Step 5. This is because we reject in this step if there is $\geq \Omega(\varepsilon^2)$ total variation distance between p and q (witnessed by the set $\bar{\mathcal{A}}$), which implies that p and q are far in ℓ_2-distance by Proposition 2. It remains to prove an upper bound on $\mathbf{E}[Z']$ in the case where $d_{\ell_2}(p,q) \leq \frac{\varepsilon^2}{\sqrt{n}}$.

$$\mathbf{E}[Z'] = m_2 d_{\chi^2}(p,q) = m_2\sum_{i\in\bar{\mathcal{A}}}\frac{(p_i - q_i)^2}{q_i} \leq O\left(m_2 \cdot \left(\frac{n}{\varepsilon^2}\right) \cdot d_{\ell_2}^2(p,q)\right) \leq O(m_2\varepsilon^2).$$

We note that this is the same bound as in Lemma 1. With this bound on the mean, the rest of the analysis is identical to that of Theorem 1, giving us Theorem 3.

4 Upper Bounds for Equivalence Testing

In this section, we prove the following theorems for equivalence testing.

Theorem 4. *There exists an algorithm for equivalence testing between p and q distinguishing the cases:*

- $d_{\ell_2}(p,q) \leq \frac{\varepsilon}{2\sqrt{n}}$

- $d_{\mathrm{TV}}(p,q) \geq \varepsilon$

The algorithm uses $O\left(\max\left\{\frac{n^{2/3}}{\varepsilon^{4/3}}, \frac{n^{1/2}}{\varepsilon^2}\right\}\right)$ *samples.*

Theorem 5. *There exists an algorithm for equivalence testing between p and q distinguishing the cases:*

- $d_{\ell_2}(p,q) \leq \frac{\varepsilon^2}{32\sqrt{n}}$
- $d_{\mathrm{H}}(p,q) \geq \varepsilon$

The algorithm uses $O\left(\min\left\{\frac{n^{2/3}}{\varepsilon^{8/3}}, \frac{n^{3/4}}{\varepsilon^2}\right\}\right)$ *samples.*

Consider drawing Poisson(m) samples from two unknown distributions $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$. Given the resulting histograms $\boldsymbol{X}$ and $\boldsymbol{Y}$, [CDVV14] define the following statistic:

$$\boldsymbol{Z} = \sum_{i=1}^{n} \frac{(\boldsymbol{X}_i - \boldsymbol{Y}_i)^2 - \boldsymbol{X}_i - \boldsymbol{Y}_i}{\boldsymbol{X}_i + \boldsymbol{Y}_i}. \tag{4}$$

This can be viewed as a modification to the empirical triangle distance applied to $\boldsymbol{X}$ and $\boldsymbol{Y}$. Both of our equivalence testing upper bounds will be obtained by appropriate thresholding of the statistic $\boldsymbol{Z}$.

The organization of this section is as follows. In Section 4.1, we prove some basic properties of $\boldsymbol{Z}$. In Section 4.2, we prove Theorem 4. In Section 4.3, we prove Theorem 5.

4.1 Some facts about Z

Chan et al. [CDVV14] give the following expressions for the mean and variance of $\boldsymbol{Z}$.

Proposition 5 ([CDVV14])**.** *Consider the function*

$$f(x) = \left(1 - \frac{1 - e^{-x}}{x}\right).$$

Then for any subset $A \subseteq [n]$,

$$\mathbf{E}[\boldsymbol{Z}_A] = \sum_{i \in A} \frac{(p_i - q_i)^2}{p_i + q_i} m \cdot f(m(p_i + q_i)). \tag{5}$$

As a result, $\boldsymbol{Z}$ is mean-zero when $p = q$. Furthermore,

$$\mathbf{Var}[\boldsymbol{Z}] \leq 2\min\{m, n\} + \sum_{i=1}^{n} 5m \frac{(p_i - q_i)^2}{p_i + q_i}.$$

Applying Proposition 3, we immediately have the following corollary.

Corollary 1. $\mathbf{Var}[\boldsymbol{Z}] \leq 2\min\{m, n\} + 20m d_{\mathrm{H}}(p,q)^2$.

Without the corrective factor of $f(m(p_i + q_i))$, Equation (5) would just be m times the triangle distance between p and q. Our goal then is to understand the function $f(x)$ and how it affects this quantity. Aside from the removable discontinuity at $x = 0$, f is a monotonically increasing function, and for $x > 0$, it is strictly bounded between 0 and 1. Furthermore, for $x > 0$ there are roughly two "regimes" that $f(x)$ exhibits: when $x < 1$, where $f(x)$ is well-approximated by $x/2$, and when $x \geq 1$, where $f(x)$ is "morally the constant one," slowly increasing from e^{-1} to 1. In fact, we have the following explicit bound on $f(x)$.

Fact 1. *For all $x > 0$, $f(x) \leq \min\{1, x\}$.*

In terms of $f(m(p_i + q_i))$, these regimes correspond to whether $p_i + q_i$ is less than or greater than $\frac{1}{m}$. Hence, the expression for the mean of $\boldsymbol{Z}$ (i.e. Equation (5) for $A = [n]$) splits in two: those terms for "large" $p_i + q_i$ look roughly like the triangle distance (times m), and those terms for "small" $p_i + q_i$ look roughly like the ℓ_2^2 distance (times m^2). This is why we have given ourselves the flexibility to consider subsets A of the domain.

We will now prove several upper and lower bounds on $\mathbf{E}[\boldsymbol{Z}_A]$, based in part on whether we will apply them in the large or small $p_i + q_i$ regime. Let us begin with a pair of upper bounds.

Proposition 6. *Suppose for every $i \in A$, $p_i + q_i \geq \delta$. Then*

$$\mathbf{E}[\boldsymbol{Z}_A] \leq \frac{m}{\delta} d_{\ell_2}^2(p_A, q_A).$$

Proof. Because $f(x) \leq 1$ for all $x > 0$,

$$\mathbf{E}[\boldsymbol{Z}_A] = \sum_{i \in A} \frac{(p_i - q_i)^2}{p_i + q_i} m \cdot f(m(p_i + q_i)) \leq \sum_{i \in A} \frac{(p_i - q_i)^2}{p_i + q_i} m \leq \frac{m}{\delta} \sum_{i \in A} (p_i - q_i)^2 = \frac{m}{\delta} d_{\ell_2}^2(p_A, q_A). \qquad \square$$

Proposition 7. $\mathbf{E}[\boldsymbol{Z}] \leq m^2 d_{\ell_2}^2(p, q)$.

Proof. Let L be the set of i such that $m(p_i + q_i) \geq 1$. Then $\mathbf{E}[\boldsymbol{Z}] = \mathbf{E}[\boldsymbol{Z}_L] + \mathbf{E}[\boldsymbol{Z}_{\overline{L}}]$, and by Proposition 6, $\mathbf{E}[\boldsymbol{Z}_L] \leq m^2 d_{\ell_2}^2(p_L, q_L)$. On the other hand, by Fact 1, $f(x) \leq x$, and therefore

$$\mathbf{E}[\boldsymbol{Z}_{\overline{L}}] = \sum_{i \in \overline{L}} \frac{(p_i - q_i)^2}{p_i + q_i} m \cdot f(m(p_i + q_i)) \leq \sum_{i \in \overline{L}} (p_i - q_i)^2 m^2 = m^2 d_{\ell_2}^2(p_{\overline{L}}, q_{\overline{L}}).$$

The proof is completed by noting that $d_{\ell_2}^2(p_L, q_L) + d_{\ell_2}^2(p_{\overline{L}}, q_{\overline{L}}) = d_{\ell_2}^2(p, q)$. $\square$

Now we give a pair of lower bounds.

Proposition 8. *Suppose for every $i \in A$, $m(p_i + q_i) \geq 1$. Then*

$$\mathbf{E}[\boldsymbol{Z}_A] \geq \frac{2m}{3} d_{\mathrm{H}}^2(p_A, q_A).$$

Proof. Because $f(x)$ is monotonically increasing and $f(1) = 1/e$,

$$\mathbf{E}[\boldsymbol{Z}_A] = m \sum_{i \in A} \frac{(p_i - q_i)^2}{p_i + q_i} f(m(p_i + q_i)) \geq m \sum_{i \in A} \frac{(p_i - q_i)^2}{p_i + q_i} f(1) \geq \frac{2m}{e} d_{\mathrm{H}}^2(p_A, q_A),$$

where the first step is by Proposition 5 and the last is by Proposition 3. The result follows from $e \leq 3$. $\square$

The next proposition is essentially the second half of the proof of Lemma 4 from [CDVV14].

Proposition 9. *For any subset A,*

$$\mathbf{E}[\boldsymbol{Z}_A] \geq \left(\frac{4m^2}{2|A| + m \cdot (p(A) + q(A))} \right) \cdot d_{\mathrm{TV}}^2(p_A, q_A),$$

where we write $p(A) = \sum_{i \in A} p(i)$ and likewise for $q(A)$.

Proof. Consider the function $g(x) = x f(x)^{-1}$. Then $g(x) \leq 2 + x$ for nonnegative x. Furthermore,

$$\frac{(p_i - q_i)^2}{g(m(p_i + q_i))} = \frac{(p_i - q_i)^2}{m(p_i + q_i)} \left(1 - \frac{1 - e^{-m(p_i + q_i)}}{m(p_i + q_i)} \right),$$

which, from Proposition 5, is $\frac{1}{m^2} \cdot \mathbf{E}[\boldsymbol{Z}_{\{i\}}]$. As a result,

$$\begin{aligned} d_{\mathrm{TV}}^2(p_A, q_A) &= \frac{1}{4} \left(\sum_{i \in A} |p_i - q_i| \right)^2 = \frac{1}{4} \left(\sum_{i \in A} |p_i - q_i| \cdot \frac{\sqrt{g(m(p_i + q_i))}}{\sqrt{g(m(p_i + q_i))}} \right)^2 \\ &\leq \frac{1}{4} \left(\sum_{i \in A} \frac{(p_i - q_i)^2}{g(m(p_i + q_i))} \right) \cdot \left(\sum_{i \in A} g(m(p_i + q_i)) \right) \leq \frac{1}{4m^2} \cdot \mathbf{E}[\boldsymbol{Z}_A] \cdot (2|A| + m \cdot (p(A) + q(A))), \end{aligned}$$

where the first inequality is Cauchy-Schwarz. Rearranging finishes the proof. $\square$

4.2 Equivalence Testing with Total Variation Distance

In this section, we prove Theorem 4. We will take the number of samples to be

$$m = \max\left\{C \cdot \frac{n^{2/3}}{\varepsilon^{4/3}}, C^{3/2} \cdot \frac{n^{1/2}}{\varepsilon^2}\right\}, \tag{6}$$

where C is some constant which can be taken to be 10^{10}.

Rather than drawing samples from p or q, our algorithm draws samples from $p^{+1/2}$ and $q^{+1/2}$. By Proposition 4, we have the following guarantees in the two cases:

$$\text{(Case 1): } d_{\ell_2}(p^{+1/2}, q^{+1/2}) \le \frac{\varepsilon}{4\sqrt{n}}, \qquad \text{(Case 2): } d_{\mathrm{TV}}(p^{+1/2}, q^{+1/2}) \ge \frac{\varepsilon}{2}.$$

Furthermore, for any $i \in [n]$, we know the i-th coordinates of $p^{+1/2}$ and $q^{+1/2}$ are both at least $\frac{1}{2n}$. Henceforth, we will write p' and q' for $p^{+1/2}$ and $q^{+1/2}$, respectively.

In Case 1, if we apply Proposition 6 with $A = [n]$ and $\delta = \frac{1}{n}$ and Proposition 7,

$$\mathbf{E}[\boldsymbol{Z}] \le \min\{m^2, mn\} \cdot d_{\ell_2}^2(p', q') \le \min\{m^2, mn\} \cdot \frac{\varepsilon^2}{16n} \le \frac{m^2}{4(2m+2n)} \cdot \varepsilon^2.$$

On the other hand, in Case 2, applying Proposition 9 with $A = [n]$,

$$\mathbf{E}[\boldsymbol{Z}] \ge \frac{4m^2}{2m+2n} \cdot d_{\mathrm{TV}}(p', q')^2 \ge \frac{m^2}{2m+2n} \cdot \varepsilon^2.$$

Our algorithm therefore thresholds $\boldsymbol{Z}$ on the value $\frac{5m^2}{8(2m+2n)}\varepsilon^2$, outputting "close" if it's below this value and "far" otherwise.

The two bounds in (6) meet when $C^3\varepsilon^{-4} = n$, which is exactly when $m = n$. When $m \le n$, the first bound applies, and when $m > n$ the second bound applies. As a result, we will split our analysis into the two cases.

Lemma 3. *The tester succeeds in the $m \le n$ case of Theorem 4.*

Proof. By Corollary 1

$$\mathbf{Var}[\boldsymbol{Z}] \le 2\min\{m, n\} + 20md_{\mathrm{H}}(p', q')^2 \le 22m,$$

where we used the fact that $d_{\mathrm{H}}(p', q') \le 1$. In Case 1, by Chebyshev's inequality,

$$\Pr\left[\boldsymbol{Z} \ge \frac{5m^2}{8(2m+2n)}\varepsilon^2\right] \le \frac{\mathbf{Var}[\boldsymbol{Z}]}{\left(\frac{3m^2}{8(2m+2n)}\varepsilon^2\right)^2} = O\left(\frac{m}{\frac{m^4}{n^2}\varepsilon^4}\right) = O\left(\frac{n^2}{m^3\varepsilon^4}\right).$$

In Case 2,

$$\Pr\left[\boldsymbol{Z} \le \frac{5m^2}{8(2m+2n)}\varepsilon^2\right] \le \frac{64\mathbf{Var}[\boldsymbol{Z}]}{9\mathbf{E}[\boldsymbol{Z}]^2} = O\left(\frac{m}{\frac{m^4}{n^2}\varepsilon^4}\right) = O\left(\frac{n^2}{m^3\varepsilon^4}\right).$$

Both of these bounds can be made arbitrarily small constants by setting C sufficiently large. □

Lemma 4. *The tester succeeds in the $m \ge n$ case of Theorem 4.*

Proof. We first consider Case 1. By Proposition 5,

$$\mathbf{Var}[\boldsymbol{Z}] \le 2\min\{m, n\} + \sum_{i=1}^{n} 5m\frac{(p_i' - q_i')^2}{p_i' + q_i'} \le 2n + 5mnd_{\ell_2}^2(p', q') \le 2n + \tfrac{5}{16}m\varepsilon^2.$$

Then, we have that

$$\Pr\left[\boldsymbol{Z} \ge \frac{5m^2}{8(2m+2n)}\varepsilon^2\right] \le \frac{\mathbf{Var}[\boldsymbol{Z}]}{\left(\frac{3m^2}{8(2m+2n)}\varepsilon^2\right)^2} = O\left(\frac{n}{m^2\varepsilon^4} + \frac{m\varepsilon^2}{m^2\varepsilon^4}\right) = O\left(\frac{n}{m^2\varepsilon^4} + \frac{1}{m\varepsilon^2}\right).$$

Next, we focus on Case 2. Write L for the set of $i \in [n]$ such that $m(p'_i + q'_i) \geq 1$. Then $d^2_{\mathrm{H}}(p'_{\overline{L}}, q'_{\overline{L}}) \leq \frac{1}{2}\sum_{i\in\overline{L}}(p'_i + q'_i) \leq n/2m$. As a result, by Corollary 1

$$\mathbf{Var}[\boldsymbol{Z}] \leq 2\min\{m,n\} + 20md^2_{\mathrm{H}}(p',q') \leq 12n + 20md^2_{\mathrm{H}}(p'_L, q'_L).$$

By Proposition 8, $\mathbf{E}[\boldsymbol{Z}] \geq \frac{2m}{3}d^2_{\mathrm{H}}(p'_L, q'_L)$. Hence,

$$\begin{aligned}\Pr\left[\boldsymbol{Z} \leq \frac{5m^2}{8(2m+2n)}\varepsilon^2\right] &\leq \frac{64\mathbf{Var}[\boldsymbol{Z}]}{9\mathbf{E}[\boldsymbol{Z}]^2} = O\left(\frac{n}{\mathbf{E}[\boldsymbol{Z}]^2} + \frac{md^2_{\mathrm{H}}(p'_L, q'_L)}{\mathbf{E}[\boldsymbol{Z}]^2}\right)\\ &= O\left(\frac{n}{\mathbf{E}[\boldsymbol{Z}]^2} + \frac{1}{\mathbf{E}[\boldsymbol{Z}]}\right) = O\left(\frac{n}{m^2\varepsilon^4} + \frac{1}{m\varepsilon^2}\right).\end{aligned}$$

Both of these bounds can be made arbitrarily small constants by setting C sufficiently large. □

4.3 Equivalence Testing with Hellinger Distance

In this section, we prove Theorem 5. We will take the number of samples to be

$$m = \min\left\{C \cdot \frac{n^{2/3}}{\varepsilon^{8/3}}, C^{3/4} \cdot \frac{n^{3/4}}{\varepsilon^2}\right\},$$

where C is some constant which can be taken to be 10^{10}.

Rather than drawing samples from p or q, our algorithm draws samples from $p^{+\delta}$ and $q^{+\delta}$ for $\delta = \varepsilon^2/32$. By Proposition 4, we have the following guarantees in the two cases:

$$\text{(Case 1): } d_{\ell_2}(p,q) \leq \frac{\varepsilon^2}{32\sqrt{n}}, \qquad \text{(Case 2): } d_{\mathrm{H}}(p,q) \geq \frac{1}{2}\varepsilon.$$

Furthermore, for any $i \in [n]$, we know the i-th coordinates of $p^{+\delta}$ and $q^{+\delta}$ are both at least $\frac{\varepsilon^2}{32n}$. Henceforth, we will write p' and q' for $p^{+\delta}$ and $q^{+\delta}$, respectively.

The two bounds meet when $C^{3/4}\varepsilon^{-2} = n^{1/4}$, which is exactly when $m = n$. When $m \leq n$, the first bound applies, and when $m > n$ the second bound applies. As a result, we will split our analysis into the two cases.

Lemma 5. *The tester succeeds in the $m \leq n$ case of Theorem 5.*

Proof. In Case 1, if we apply Proposition 7,

$$\mathbf{E}[\boldsymbol{Z}] \leq m^2 \cdot d^2_{\ell_2}(p',q') \leq \frac{m^2\varepsilon^4}{32^2 n}.$$

On the other hand, in Case 2, applying Proposition 9 with $A = [n]$,

$$\mathbf{E}[\boldsymbol{Z}] \geq \left(\frac{4m^2}{2n+2m}\right) \cdot d_{\mathrm{TV}}(p',q')^2 \geq \left(\frac{4m^2}{2n+2m}\right) \cdot d_{\mathrm{H}}(p',q')^4 \geq \frac{m^2\varepsilon^4}{16n}.$$

Our algorithm therefore thresholds $\boldsymbol{Z}$ on the value $\frac{m^2\varepsilon^4}{128n}$, outputting "close" if it's below this value and "far" otherwise.

By Corollary 1

$$\mathbf{Var}[\boldsymbol{Z}] \leq 2\min\{m,n\} + 20md_{\mathrm{H}}(p',q')^2 \leq 22m,$$

where we used the fact that $d_{\mathrm{H}}(p',q') \leq 1$. In Case 1,

$$\Pr\left[\boldsymbol{Z} \geq \frac{m^2\varepsilon^4}{128n}\right] \leq \frac{\mathbf{Var}[\boldsymbol{Z}]}{\left(\frac{m^2\varepsilon^4}{256n}\right)^2} = O\left(\frac{m}{\frac{m^4}{n^2}\varepsilon^8}\right) = O\left(\frac{n^2}{m^3\varepsilon^8}\right).$$

In Case 2,

$$\Pr\left[\boldsymbol{Z} \leq \frac{m^2\varepsilon^4}{128n}\right] \leq \frac{64\mathbf{Var}[\boldsymbol{Z}]}{49\mathbf{E}[\boldsymbol{Z}]^2} = O\left(\frac{m}{\frac{m^4}{n^2}\varepsilon^8}\right) = O\left(\frac{n^2}{m^3\varepsilon^8}\right).$$

Both of these bounds can be made arbitrarily small constants by setting C sufficiently large. □

Lemma 6. *The tester succeeds in the $m > n$ case of Theorem 5.*

Proof. In Case 1, if we apply Proposition 6 with $A = [n]$ and $\delta = \frac{\varepsilon^2}{16n}$ and Proposition 7,

$$\mathbf{E}[\boldsymbol{Z}] \le \min\left\{m^2, 16\frac{mn}{\varepsilon^2}\right\} \cdot d_{\ell_2}^2(p', q') \le \min\left\{m^2, 16\frac{mn}{\varepsilon^2}\right\} \cdot \frac{\varepsilon^4}{32^2 n} = \min\left\{\frac{m^2\varepsilon^4}{32^2 n}, \frac{m\varepsilon^2}{64}\right\}.$$

Case 2 is more complicated. We will need to define the set of "large" coordinates $L = \{i : m(p'_i + q'_i) \ge 1\}$ and the set of "small" coordinates $S = [n] \setminus L$. Applying Proposition 9 to S, we have

$$\mathbf{E}[\boldsymbol{Z}_S] \ge \left(\frac{4m^2}{2|S| + m \cdot (p'(S) + q'(S))}\right) \cdot d_{\mathrm{TV}}^2(p'_S, q'_S) \ge \frac{4m^2}{3n} d_{\mathrm{TV}}^2(p'_S, q'_S),$$

where $m \cdot (p'(S) + q'(S)) \le n$ by the definition of S. If we also apply Proposition 8 to L, we get

$$\mathbf{E}[\boldsymbol{Z}] = \mathbf{E}[\boldsymbol{Z}_S] + \mathbf{E}[\boldsymbol{Z}_L] \ge \frac{4m^2}{3n} d_{\mathrm{TV}}^2(p'_S, q'_S) + \frac{2m}{3} d_{\mathrm{H}}^2(p'_L, q'_L) \ge \min\left\{\frac{m^2\varepsilon^4}{48n}, \frac{m\varepsilon^2}{12}\right\},$$

where the last step follows because $d_{\mathrm{H}}^2(p'_S, q'_S) + d_{\mathrm{H}}^2(p'_L, q'_L) = d_{\mathrm{H}}^2(p', q')$ and $d_{\mathrm{TV}}^2(p'_S, q'_S) \ge d_{\mathrm{H}}^4(p'_S, q'_S)$. As a result, we threshold $\boldsymbol{Z}$ on the value

$$\frac{1}{2} \cdot \min\left\{\frac{m^2\varepsilon^4}{48n}, \frac{m\varepsilon^2}{12}\right\},$$

outputting "close" if it's below this value and "far" otherwise.

In Case 1, by Proposition 5,

$$\mathbf{Var}[\boldsymbol{Z}] \le 2\min\{m, n\} + \sum_{i=1}^{m} 5m\frac{(p'_i - q'_i)^2}{p'_i + q'_i} \le 2n + \frac{80mn}{\varepsilon^2}\|p' - q'\|_2^2 \le 2n + \frac{5}{64}m\varepsilon^2.$$

Hence, by Chebyshev's inequality,

$$\begin{aligned}\Pr\left[\boldsymbol{Z} \ge \frac{1}{2} \cdot \min\left\{\frac{m^2\varepsilon^4}{48n}, \frac{m\varepsilon^2}{12}\right\}\right] &\le \frac{\mathbf{Var}[\boldsymbol{Z}]}{(\frac{1}{8} \cdot \min\left\{\frac{m^2\varepsilon^4}{48n}, \frac{m\varepsilon^2}{12}\right\})^2} \\ &\le O\left(\frac{n}{(\frac{m^2\varepsilon^4}{n})^2} + \frac{n}{(m\varepsilon^2)^2} + \frac{m\varepsilon^2}{(\frac{m^2\varepsilon^4}{n})^2} + \frac{m\varepsilon^2}{(m\varepsilon^2)^2}\right) \\ &= O\left(\frac{n^3}{m^4\varepsilon^8} + \frac{n}{m^2\varepsilon^4} + \frac{n^2}{m^3\varepsilon^6} + \frac{1}{m\varepsilon^2}\right).\end{aligned}$$

This can be made an arbitrarily small constant by setting C sufficiently large.

In Case 2, by Corollary 1,

$$\Pr\left[\boldsymbol{Z} \le \frac{\mathbf{E}[\boldsymbol{Z}]}{2}\right] \le \frac{4\mathbf{Var}[\boldsymbol{Z}]}{\mathbf{E}[\boldsymbol{Z}]^2} \le \frac{8n + 80md_{\mathrm{H}}(p', q')^2}{\mathbf{E}[\boldsymbol{Z}]^2}. \tag{7}$$

Because $d_{\mathrm{H}}(p', q')^2 = d_{\mathrm{H}}^2(p'_S, q'_S) + d_{\mathrm{H}}^2(p'_L, q'_L)$, either $d_{\mathrm{H}}^2(p'_S, q'_S)$ or $d_{\mathrm{H}}^2(p'_L, q'_L)$ is at least $\frac{1}{2}d_{\mathrm{H}}^2(p', q')$. Suppose that $d_{\mathrm{H}}^2(p'_S, q'_S) \ge \frac{1}{2}d_{\mathrm{H}}^2(p', q')$. We note that

$$md_{\mathrm{H}}^2(p'_S, q'_S) = \frac{m}{2}\sum_{i \in S}(\sqrt{p'_i} - \sqrt{q'_i})^2 \le \frac{m}{2}\sum_{i \in S}|p'_i + q'_i| \le \frac{n}{2},$$

by the definition of S. Thus,

$$(7) \le \frac{8n + 160md_{\mathrm{H}}^2(p'_S, q'_S)}{(\frac{4m^2}{3n}d_{\mathrm{TV}}^2(p'_S, q'_S))^2} \le \frac{88n}{(\frac{4m^2}{3n}d_{\mathrm{TV}}^2(p'_S, q'_S))^2} = O\left(\frac{n^3}{m^4 d_{\mathrm{TV}}^4(p'_S, q'_S)}\right) \le O\left(\frac{n^3}{m^4\varepsilon^8}\right),$$

where the last step used the fact that $d_{\mathrm{TV}}(p'_S, q'_S) \ge d_{\mathrm{H}}^2(p'_S, q'_S) \ge \frac{1}{2}d_{\mathrm{H}}^2(p', q') \ge \frac{1}{2}\varepsilon^2$.

In the case when $d_{\mathrm{H}}^2(p'_L, q'_L) \ge \frac{1}{2}d_{\mathrm{H}}^2(p', q')$,

$$(7) \le \frac{8n + 160md_{\mathrm{H}}^2(p'_L, q'_L)}{(\frac{2m}{3}d_{\mathrm{H}}^2(p'_L, q'_L))^2} = O\left(\frac{n}{m^2 d_{\mathrm{H}}^4(p'_L, q'_L)} + \frac{1}{md_{\mathrm{H}}^2(p'_L, q'_L)}\right) \le O\left(\frac{n}{m^2\varepsilon^4} + \frac{1}{m\varepsilon^2}\right).$$

This can be made an arbitrarily small constant by setting C sufficiently large. □

5 Upper Bounds Based on Estimation

We start by showing a simple meta-algorithm – in short, it says that if a testing problem is well-defined (i.e., has appropriate separation between the cases) and we can estimate one of the distances, it can be converted to a testing algorithm.

Theorem 6. *Suppose there exists an $m(n,\varepsilon)$-sample algorithm which, given sample access to distributions p and q over $[n]$ and parameter ε, estimates some distance $d(p,q)$ up to an additive ε with probability at least $2/3$. Consider distances $d_X(\cdot,\cdot), d_Y(\cdot,\cdot)$ and $\varepsilon_1,\varepsilon_2 > 0$ such that $d_Y(p,q) \geq \varepsilon_2 \to d_X(p,q) > 3\varepsilon_1/2$ and $d_X(p,q) \leq \varepsilon_1 \to d_Y(p,q) < 2\varepsilon_2/3$, and $d(\cdot,\cdot)$ is either $d_X(\cdot,\cdot)$ or $d_Y(\cdot,\cdot)$.*

Then there exists an algorithm for equivalence testing between p and q distinguishing the cases:

- *$d_X(p,q) \leq \varepsilon_1$;*
- *$d_Y(p,q) \geq \varepsilon_2$.*

The algorithm uses either $m(n, O(\varepsilon_1))$ or $m(n, O(\varepsilon_2))$ samples, depending on whether $d = d_X$ or d_Y.

Proof. Suppose that $d = d_X$, the other case follows similarly. Using the $m(n,\varepsilon_1/4)$ samples, obtain an estimate $\hat{\tau}$ of $d_X(p,q)$, accurate up to an additive $\varepsilon_1/4$. If $\hat{\tau} \leq 5\varepsilon_1/4$, output that $d_X(p,q) \leq \varepsilon_1$, else output that $d_Y(p,q) \geq \varepsilon_2$. Conditioning on the correctness of the estimation algorithm, correctness for the case when $d_X(p,q) \leq \varepsilon_1$ is immediate, and correctness for the case when $d_Y(p,q) \geq \varepsilon_2$ follows from the separation between the cases. □

It is folklore that a distribution over $[n]$ can be ε-learned in ℓ_2-distance with $O(1/\varepsilon^2)$ samples (see, i.e., [CDVV14, Wag15] for a reference). By triangle inequality, this implies that we can estimate the ℓ_2 distance between p and q up to an additive $O(\varepsilon)$ with $O(1/\varepsilon^2)$ samples, leading to the following corollary.

Corollary 2. *There exists an algorithm for equivalence testing between p and q distinguishing the cases:*

- *$d(p,q) \leq f(n,\varepsilon)$;*
- *$d_{\ell_2}(p,q) \geq \varepsilon$,*

where $d(\cdot,\cdot)$ is a distance and $f(n,\varepsilon)$ is such that $d_{\ell_2}(p,q) \geq \varepsilon \to d(p,q) \geq 3f(n,\varepsilon)/2$ and $d(p,q) \leq f(n,\varepsilon) \to d_{\ell_2}(p,q) \leq 2\varepsilon/3$. The algorithm uses $O(1/\varepsilon^2)$ samples.

Finally, we note that total variation distance between p and q can be additively estimated up to a constant using $O(n/\log n)$ samples [LC06, VV11b, JHW16], leading to the following corollary:

Corollary 3. *For constant $\varepsilon > 0$, there exists an algorithm for equivalence testing between p and q distinguishing the cases:*

- *$d_{\mathrm{TV}}(p,q) \leq \varepsilon^2/4$;*
- *$d_{\mathrm{H}}(p,q) \geq \varepsilon/\sqrt{2}$.*

The algorithm uses $O(n/\log n)$ samples.

6 Lower Bounds

We start with a simple lower bound, showing that identity testing with respect to KL divergence is impossible. A similar observation was made in [BFR+00].

Theorem 7. *No finite sample test can perform identity testing between p and q distinguishing the cases:*

- *$p = q$;*
- *$d_{\mathrm{KL}}(p,q) \geq \varepsilon^2$.*

Proof. Simply take $q = (1, 0)$ and let p be either $(1, 0)$ or $(1 - \delta, \delta)$, for $\delta > 0$ tending to zero. Then $p = q$ in the first case and $d_{\mathrm{KL}}(p, q) = \infty$ in the second, but distinguishing between these two possibilities for p takes $\Omega(\delta^{-1}) \to \infty$ samples. □

Next, we prove our lower bound for KL tolerant identity testing.

Theorem 8. *There exist constants $0 < s < c$, such that any algorithm for identity testing between p and q distinguishing the cases:*

- $d_{\mathrm{KL}}(p, q) \leq s$;
- $d_{\mathrm{TV}}(p, q) \geq c$;

requires $\Omega(n/\log n)$ samples.

Proof. Let $q = (\frac{1}{n}, \ldots, \frac{1}{n})$ be the uniform distribution. Let $R(\cdot, \cdot)$ denote the *relative earthmover distance* (see [VV10a] for the definition). By Theorem 1 of [VV10a], for any $\delta < \frac{1}{4}$ there exist sets of distributions $\mathcal{C}$ and $\mathcal{F}$ (for *close* and *far*) such that:

- For every $p \in \mathcal{C}$, $R(p, q) = O(\delta |\log \delta|)$.
- For every $p \in \mathcal{F}$ there exists a distribution r which is uniform over $n/2$ elements such that $R(p, r) = O(\delta |\log \delta|)$.
- Distinguishing between $p \in \mathcal{C}$ and $p \in \mathcal{F}$ requires $\Omega(\frac{\delta n}{\log(n)})$ samples.

Now, if $p \in \mathcal{C}$ then

$$d_{\mathrm{KL}}(p, q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{1/n}\right) = \log(n) - H(p) \leq O(\delta |\log(\delta)|),$$

where $H(p)$ is the Shannon entropy of p, and here we used the fact that $|H(p) - H(q)| \leq R(p, q)$, which follows from Fact 5 of [VV10a]. On the other hand, if $q \in \mathcal{F}$, let r be the corresponding distribution which is uniform over $n/2$ elements. Then

$$\frac{1}{2} = d_{\mathrm{TV}}(p, q) \leq d_{\mathrm{TV}}(q, p) + d_{\mathrm{TV}}(p, r) \leq d_{\mathrm{TV}}(q, p) + O(\delta |\log \delta|),$$

where we used the triangle inequality and the fact that $d_{\mathrm{TV}}(p, r) \leq R(p, r)$ (see [VV10a] page 4). As a result, if we set δ to be some small constant, $s = O(\delta |\log(\delta)|)$, and $c = \frac{1}{2} - O(\delta |\log \delta|)$, then this argument shows that distinguish $d_{\mathrm{KL}}(p, q) \leq s$ versus $d_{\mathrm{TV}}(p, q) \geq c$ requires $\Omega(n/\log n)$ samples. □

Finally, we conclude with our lower bound for χ^2-tolerant equivalence testing.

Theorem 9. *There exists a constant $\varepsilon > 0$ such that any algorithm for equivalence testing between p and q distinguishing the cases:*

- $d_{\chi^2}(p, q) \leq \varepsilon^2/4$;
- $d_{\mathrm{TV}}(p, q) \geq \varepsilon$;

requires $\Omega(n/\log n)$ samples.

Proof. We reduce the problem of distinguishing $d_{\mathrm{H}}(p, q) \leq \frac{1}{\sqrt{48}}\varepsilon$ from $d_{\mathrm{TV}}(p, q) \geq 3\varepsilon$ to this. Define the distributions

$$p' = \frac{2}{3}p + \frac{1}{3}q, \qquad q' = \frac{1}{3}p + \frac{2}{3}q.$$

Then m samples from p' and q' can be simulated by m samples from p and q. Furthermore,

$$d_{\mathrm{H}}(p', q') \leq \frac{1}{\sqrt{48}}\varepsilon, \qquad d_{\mathrm{TV}}(p', q') = \frac{1}{3}d_{\mathrm{TV}}(p, q) \geq \varepsilon,$$

where we used the fact that Hellinger distance satisfies the data processing inequality. But then, in the "close" case,

$$d_{\chi^2}(p', q') = \sum_{i=1}^{n} \frac{(p'_i - q'_i)^2}{q'_i} \leq 3 \sum_{i=1}^{n} \frac{(p'_i - q'_i)^2}{p'_i + q'_i} \leq 12 d_{\mathrm{H}}^2(p', q') \leq \frac{1}{4}\varepsilon^2,$$

where we used the fact that $p'_i \leq 2q'_i$ and Proposition 3. Hence, this problem, which requires $\Omega(n/\log n)$ samples (by the relationship between total variation and Hellinger distance, and the lower bound for testing total variation-close versus -far of [VV10a]), reduces to the problem in the proposition, and so that requires $\Omega(n/\log n)$ samples as well. □

References

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 3577–3598. Curran Associates, Inc., 2015.

[ADOS17] Jayadev Acharya, Hirakendu Das, Alon Orlitsky, and Ananda Theertha Suresh. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 11–21. JMLR, Inc., 2017.

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

[BCG17] Eric Blais, Clément L. Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. In *Proceedings of the 32nd Computational Complexity Conference*, CCC '17, pages 28:1–28:40, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[BFF+01] Tuğkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, FOCS '01, pages 442–451, Washington, DC, USA, 2001. IEEE Computer Society.

[BFR+00] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '00, pages 259–269, Washington, DC, USA, 2000. IEEE Computer Society.

[BFR+13] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *Journal of the ACM*, 60(1):4:1–4:25, 2013.

[BKR04] Tuğkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of the 36th Annual ACM Symposium on the Theory of Computing*, STOC '04, New York, NY, USA, 2004. ACM.

[BV15] Bhaswar Bhattacharya and Gregory Valiant. Testing closeness with unequal sized samples. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 2611–2619. Curran Associates, Inc., 2015.

[Can15] Clément L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22(63), 2015.

[CDGR16] Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. In *Proceedings of the 33rd Symposium on Theoretical Aspects of Computer Science*, STACS '16, pages 25:1–25:14, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[CDKS17] Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing Bayesian networks. In *Proceedings of the 30th Annual Conference on Learning Theory*, COLT '17, pages 370–448, 2017.

[CDVV14] Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 1193–1203, Philadelphia, PA, USA, 2014. SIAM.

[DBNNR11] Khanh Do Ba, Huy L. Nguyen, Huy N. Nguyen, and Ronitt Rubinfeld. Sublinear time algorithms for earth movers distance. *Theory of Computing Systems*, 48(2):428–442, 2011.

[DDK18] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing Ising models. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, Philadelphia, PA, USA, 2018. SIAM.

[DGPP16] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Collision-based testers are optimal for uniformity and closeness. *arXiv preprint arXiv:1611.03579*, 2016.

[DK16] Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 685–694, Washington, DC, USA, 2016. IEEE Computer Society.

[DKN15] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 1841–1854, Philadelphia, PA, USA, 2015. SIAM.

[DP17] Constantinos Daskalakis and Qinxuan Pan. Square Hellinger subadditivity for Bayesian networks and its applications to identity testing. In *Proceedings of the 30th Annual Conference on Learning Theory*, COLT '17, pages 697–703, 2017.

[Gol16] Oded Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. *Electronic Colloquium on Computational Complexity (ECCC)*, 23(15), 2016.

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 7(20), 2000.

[GS02] Alison L. Gibbs and Francis E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, dec 2002.

[HJW16] Yanjun Han, Jiao Jiantao, and Tsachy Weissman. Minimax rate-optimal estimation of divergences between discrete distributions. *arXiv preprint arXiv:1605.09124*, 2016.

[JHW16] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the ℓ_1 distance. In *Proceedings of the 2016 IEEE International Symposium on Information Theory*, ISIT '16, pages 750–754, Washington, DC, USA, 2016. IEEE Computer Society.

[JVHW17] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2017.

[LC06] Erich Leo Lehmann and George Casella. *Theory of Point Estimation*. Springer, 2006.

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

[Val11] Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011.

[VV10a] Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 17(179), 2010.

[VV10b] Gregory Valiant and Paul Valiant. Estimating the unseen: A sublinear-sample canonical estimator of distributions. *Electronic Colloquium on Computational Complexity (ECCC)*, 17(180), 2010.

[VV11a] Gregory Valiant and Paul Valiant. Estimating the unseen: An $n/\log n$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the 43rd Annual ACM Symposium on the Theory of Computing*, STOC '11, pages 685–694, New York, NY, USA, 2011. ACM.

[VV11b] Gregory Valiant and Paul Valiant. The power of linear estimators. In *Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science*, FOCS '11, pages 403–412, Washington, DC, USA, 2011. IEEE Computer Society.

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

[Wag15] Bo Waggoner. l_p testing and learning of discrete distributions. In *Proceedings of the 6th Conference on Innovations in Theoretical Computer Science*, ITCS '15, pages 347–356, New York, NY, USA, 2015. ACM.

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

A Proof of Proposition 1

Recall that we will prove this for restrictions of probability distributions to subsets of the support – in other words, we do not assume $\sum_{i\in S} p_i = \sum_{i\in S} q_i = 1$, we only assume that $\sum_{i\in S} p_i \leq 1$ and $\sum_{i\in S} q_i \leq 1$.

$d^2_{\mathrm{H}}(p_S, q_S) \leq d_{\mathrm{TV}}(p_S, q_S)$:

$$\begin{aligned} d^2_{\mathrm{H}}(p_S, q_S) &= \frac{1}{2}\sum_{i\in S}(\sqrt{p_i} - \sqrt{q_i})^2 \\ &\leq \frac{1}{2}\sum_{i\in S}|\sqrt{p_i} - \sqrt{q_i}|(\sqrt{p_i} + \sqrt{q_i}) \\ &= \frac{1}{2}\sum_{i\in S}|p_i - q_i| \\ &= d_{\mathrm{TV}}(p_S, q_S). \end{aligned}$$

$d_{\mathrm{TV}}(p_S, q_S) \le \sqrt{2} d_{\mathrm{H}}(p_S, q_S)$:

$$\begin{aligned}
d_{\mathrm{TV}}^2(p_S, q_S) &= \frac{1}{4}\left(\sum_{i\in S} |p_i - q_i|\right)^2 \\
&= \frac{1}{4}\left(\sum_{i\in S} |\sqrt{p_i} - \sqrt{q_i}|\,(\sqrt{p_i} + \sqrt{q_i})\right)^2 \\
&\le \frac{1}{4}\left(\sum_{i\in S} |\sqrt{p_i} - \sqrt{q_i}|^2\right)\left(\sum_{i\in S} (\sqrt{p_i} + \sqrt{q_i})^2\right) \\
&\le d_{\mathrm{H}}^2(p_S, q_S) \cdot \frac{1}{2}\left(\sum_{i\in S} (\sqrt{p_i} + \sqrt{q_i})^2\right) \\
&= d_{\mathrm{H}}^2(p_S, q_S) \cdot \left(\sum_{i\in S} p_i + \sum_{i\in S} q_i - d_{\mathrm{H}}^2(p_S, q_S)\right) \\
&\le d_{\mathrm{H}}^2(p_S, q_S) \cdot \left(2 - d_{\mathrm{H}}^2(p_S, q_S)\right) \\
&\le 2 d_{\mathrm{H}}^2(p_S, q_S).
\end{aligned}$$

Taking the square root of both sides gives the result. The second inequality is Cauchy-Schwarz.

$2d_{\mathrm{H}}^2(p_S, q_S) \le \sum_{i\in S}(q_i - p_i) + d_{\mathrm{KL}}(p_S, q_S)$:

$$\begin{aligned}
2d_{\mathrm{H}}^2(p_S, q_S) &= \sum_{i\in S}(q_i + p_i) - 2\sum_{i\in S}\sqrt{p_i q_i} \\
&= \sum_{i\in S}(q_i + p_i) - 2\left(\left(\sum_{j\in S} p_j\right)\sum_{i\in S}\frac{p_i}{\sum_{j\in S} p_j}\sqrt{\frac{q_i}{p_i}}\right) \\
&\le \sum_{i\in S}(q_i + p_i) - 2\left(\left(\sum_{j\in S} p_j\right)\exp\left(\frac{1}{2}\sum_{i\in S}\frac{p_i}{\sum_{j\in S} p_j}\log\frac{q_i}{p_i}\right)\right) \\
&\le \sum_{i\in S}(q_i + p_i) - 2\left(\left(\sum_{j\in S} p_j\right)\left(1 + \frac{1}{2}\sum_{i\in S}\frac{p_i}{\sum_{j\in S} p_j}\log\frac{q_i}{p_i}\right)\right) \\
&= \sum_{i\in S}(q_i - p_i) - \left(\sum_{i\in S} p_i \log\frac{q_i}{p_i}\right) \\
&= \sum_{i\in S}(q_i - p_i) + d_{\mathrm{KL}}(p_S, q_S).
\end{aligned}$$

The first inequality is Jensen's, and the second is $1 + x \le \exp(x)$.

$d_{\mathrm{KL}}(p_S, q_S) \leq \sum_{i\in S}(p_i - q_i) + d_{\chi^2}(p_S, q_S)$:

$$\begin{aligned}
d_{\mathrm{KL}}(p_S, q_S) &= \left(\sum_{j\in S} p_j\right)\left(\sum_{i\in S} \frac{p_i}{\sum_{j\in S} p_j} \log \frac{p_i}{q_i}\right) \\
&\leq \left(\sum_{j\in S} p_j\right)\left(\log \frac{1}{\sum_{j\in S} p_j} \sum_{i\in S} \frac{p_i^2}{q_i}\right) \\
&= \left(\sum_{j\in S} p_j\right)\left(\log\left(\frac{1}{\sum_{j\in S} p_j}\left(d_{\chi^2}(p_S, q_S) + 2\sum_{i\in S} p_i - \sum_{i\in S} q_i\right)\right)\right) \\
&= \left(\sum_{j\in S} p_j\right)\left(\log\left(2 + \frac{1}{\sum_{j\in S} p_j}\left(d_{\chi^2}(p_S, q_S) - \sum_{i\in S} q_i\right)\right)\right) \\
&\leq \left(\sum_{j\in S} p_j\right)\left(1 + \frac{1}{\sum_{j\in S} p_j}\left(d_{\chi^2}(p_S, q_S) - \sum_{i\in S} q_i\right)\right) \\
&= \sum_{i\in S}(p_i - q_i) + d_{\chi^2}(p_S, q_S).
\end{aligned}$$

The first inequality is Jensen's, and the second is $1 + x \leq \exp(x)$.

ECCC ISSN 1433-8092
https://eccc.weizmann.ac.il