Optimal Algorithms for Testing Closeness of Discrete Distributions

Siu-On Chan*
MSR New England
siuon@cs.berkeley.edu.

Ilias Diakonikolas†
University of Edinburgh
ilias.d@ed.ac.uk.

Gregory Valiant‡
Stanford University
valiant@stanford.edu.

Paul Valiant
Brown University
pvaliant@gmail.com.

October 10, 2013

Abstract

We study the question of closeness testing for two discrete distributions. More precisely, given samples from two distributions p and q over an n-element set, we wish to distinguish whether $p = q$ versus p is at least ε-far from q, in either ℓ_1 or ℓ_2 distance. Batu et al [BFR$^+$00, BFR$^+$13] gave the first sub-linear time algorithms for these problems, which matched the lower bounds of [Val11] up to a logarithmic factor in n, and a polynomial factor of ε.

In this work, we present simple testers for both the ℓ_1 and ℓ_2 settings, with sample complexity that is information-theoretically optimal, to constant factors, both in the dependence on n, and the dependence on ε; for the ℓ_1 testing problem we establish that the sample complexity is $\Theta(\max\{n^{2/3}/\varepsilon^{4/3}, n^{1/2}/\varepsilon^2\})$.

1 Introduction

Consider the following natural statistical task: Given independent samples from a pair of unknown distributions p, q, determine whether the two distributions are *the same* versus significantly different. We focus on the most basic (and well-studied) setting in which both p and q are discrete distributions supported on a set of size n. For a parameter $0 < \varepsilon < 1$, we want to distinguish (with probability at least 2/3, say) between the case that $p = q$ and the case that p and q are ε-*far* from each other, i.e., the ℓ_1 distance between p and q is at least ε. We will henceforth refer to this task as the problem of *closeness testing* for p and q.

We would like to design an algorithm (tester) for this task that uses as few samples as possible and is computationally efficient (i.e., has running time polynomial in its sample size). One natural way to solve this problem would be to get sufficiently many samples from p, q in order to *learn* each distribution to accuracy $O(\varepsilon)$, and then check closeness of the corresponding hypothesis distributions. As natural as it may be, this testing-via-learning approach is quite naive and gives suboptimal results. We note that learning an arbitrary distribution over support of size n to ℓ_1 distance ε requires $\Theta(n/\varepsilon^2)$ samples (i.e., there is an upper bound of $O(n/\varepsilon^2)$ and a matching information-theoretic lower bound of $\Omega(n/\varepsilon^2)$). One might hope that a better sample size bound could be achieved for the closeness testing problem, since this task is, in some sense, more specific than the general task of learning. Indeed, this is known to be the case: previous work [BFR$^+$00] gave a tester for this problem with sample complexity *sub-linear* in n and polynomial in $1/\varepsilon$.

Despite its long history in both statistics and computer science, the sample complexity of this basic task has not been resolved to date. While the dependence on n in the previous bound [BFR$^+$00] was subsequently shown [Val08, Val11] to be tight to within logarithmic factors of n, there was a polynomial gap between the upper and lower bounds in the dependence on ε. Due to its fundamental nature, we believe it is of interest from a theoretical standpoint to obtain an *optimal* sample (and time) algorithm for the problem. From a practical perspective, we note that in an era of "big data" it is critical to use data efficiently. In particular, in such a context, even modest asymptotic differences in the sample complexity can play a big role.

In this paper, we resolve the complexity of the closeness testing problem, up to a constant factor, by

*Supported by NSF award DMS-1106999, DOD ONR grant N000141110140 and NSF award CCF-1118083.

†Supported in part by a SICSA PECE grant. Part of this work was done while the author was at UC Berkeley supported by a Simons Postdoctoral Fellowship.

‡The majority of this work was done while the author was at Microsoft Research.

designing a sample-optimal algorithm (tester) for it whose running time is linear in the sample size. Our tester has a different structure from the one in [BFR+00] and is also much simpler. We also study the closeness testing problem with respect to the ℓ_2 distance metric between distributions. This problem, interesting in its own right, has been explicitly studied in previous work [GR00, BFR+00].

As our second contribution, we design a similarly optimal algorithm for closeness testing in the ℓ_2 norm. In this ℓ_2 setting, we show that the *same* sample complexity allows one to "robustly" test closeness; namely, the same sample complexity allows one to distinguish the case that $||p-q||_2 \leq \varepsilon$ from the case that $||p-q||_2 \geq 2\varepsilon$. This correspondence between the robust and non-robust closeness testing in the ℓ_2 setting does not hold for the ℓ_1 setting: the lower bounds of [VV11b] show that robust ℓ_1 testing for distributions of support size n requires $\Theta(\frac{n}{\log n})$ samples (for constant ε), as opposed to the $\Theta(n^{2/3})$ for the non-robust testing problem. One may alternately consider "robust" closeness testing under the ℓ_2 norm as essentially the problem of *estimating* the ℓ_2 distance, and the results of Proposition 3.1 are presented from this perspective.

Algorithmic ideas developed for the closeness testing problem have typically been useful for related testing questions, including the independence of bivariate distributions (see e.g. [BFF+01, BKR04]). It is plausible that our techniques may be used to obtain similarly optimal algorithms for these problems, but we have not pursued this direction.

Before we formally state our results, we start by providing some background in the area of distribution property testing.

Related Work. Estimating properties of distributions using samples is a classical topic in statistics that has received considerable attention in the theoretical CS community during the past decade; see [GR00, BFR+00, BFF+01, Bat01, BDKR02, BKR04, Pan08, Val08, Ona09, Val11, VV11a, VV11b, DDS+13, Rub12, BNNR11, ADJ+11, ADJ+12, LRR11, ILR12, AIOR09] for a sample of works and [Rub12] for a recent survey on the topic. In addition to closeness testing, various properties of distributions have been considered, including independence [BFF+01, Ona09], entropy [BDKR02], and the more general class of "symmetric" properties [Val08, VV11a, VV11b], monotonicity [BKR04], etc.

One of the first theoretical CS papers that explicitly studied such questions is the work of Batu et al [BFR+00] (see [BFR+13] for the journal version). In this work, the authors formally pose the closeness testing problem and give a tester for the problem with sub-linear sample complexity. In particular, the sample complexity of their algorithm under the ℓ_1 norm is $O(\frac{n^{2/3}\log n}{\varepsilon^{8/3}})$. A related (easier) problem is that of *uniformity testing*, i.e., distinguishing between the case that an unknown distribution p (accessible via samples) is uniform versus ε-far from uniform. Goldreich and Ron [GR00], motivated by a connection to testing expansion in graphs, obtained a uniformity tester using $O(\sqrt{n}/\varepsilon^4)$ samples. Subsequently, Paninski gave the tight bound of $\Theta(\sqrt{n}/\varepsilon^2)$ [Pan08]. (Similar results are obtained for both testing problems under the ℓ_2 norm.)

Acharya et al. [ADJ+12] also considered the problem of ℓ_1 closeness testing, but from a rather different "competitive analysis" perspective, constructing a single tester that is competitive against all testers from a broad class, even those with, essentially, knowledge of the underlying distributions built-in. The form of our ℓ_1 tester is very similar to that proposed in their work, and we discuss this connection and the intuition behind such an estimator in Section 2.

Notation. We write $[n]$ to denote the set $\{1, \ldots, n\}$. We consider discrete probability distributions over $[n]$, which are functions $p : [n] \to [0, 1]$ such that $\sum_{i=1}^n p_i = 1$. We will typically use the notation p_i to denote the probability of element i in distribution p. The ℓ_1 (resp. ℓ_2) norm of a distribution is identified with the ℓ_1 (resp. ℓ_2) norm of the corresponding n-vector, i.e., $\|p\|_1 = \sum_{i=1}^n |p_i|$ and $\|p\|_2 = \sqrt{\sum_{i=1}^n p_i^2}$. The ℓ_1 (resp. ℓ_2) distance between distributions p and q is defined as the the ℓ_1 (resp. ℓ_2) norm of the vector of their difference, i.e., $\|p - q\|_1 = \sum_{i=1}^n |p_i - q_i|$ and $\|p - q\|_2 = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$. For $\lambda \geq 0$, we denote by $\mathrm{Poi}(\lambda)$ the Poisson distribution with parameter λ.

Our Results. Our main result is an optimal algorithm for the ℓ_1-closeness testing problem:

THEOREM 1.1. *Given $\varepsilon > 0$ and sample access to distributions p and q over $[n]$, there is an algorithm which uses $O(\max\{n^{2/3}/\varepsilon^{4/3}, n^{1/2}/\varepsilon^2\})$ samples, runs in time linear in its sample size and with probability at least $2/3$ distinguishes whether $p = q$ versus $\|p - q\|_1 \geq \varepsilon$. Additionally, $\Omega(\max\{n^{2/3}/\varepsilon^{4/3}, n^{1/2}/\varepsilon^2\})$ samples are information-theoretically necessary.*

The lower bound is obtained by leveraging the techniques of [Val11] to show that $\Omega(n^{2/3}/\varepsilon^{4/3})$ is a lower bound, as long as $\varepsilon = \Omega(n^{-1/4})$ (see Section 4 for the proof). On the other hand, the sample complexity of ℓ_1-closeness testing is bounded from below by the sample complexity of uniformity testing (for all values of n and $\varepsilon > 0$), since knowing that one distribution is exactly the uniform distribution can only make the

testing problem easier.

Hence, by the result of Paninski [Pan08], it follows that $\Omega(\sqrt{n}/\varepsilon^2)$ is also a lower bound. The tight lower bound of $\Omega(\max\{n^{2/3}/\varepsilon^{4/3}, n^{1/2}/\varepsilon^2\})$ follows from the fact that the two functions intersect for $\varepsilon = \Theta(n^{-1/4})$. Hence, our algorithm of Theorem 1.1 is optimal (up to constant factors) for all $\varepsilon > 0$.

Our second result is an algorithm for "robustly" testing the closeness of a pair of distributions with respect to ℓ_2 distance, which is also information theoretically optimal for all parameters, to constant factors. The parameter b in the following theorem upper-bounds the ℓ_2 norm-squared of each distribution, which allows the theorem to be more finely tuned to the cases when testing should be easier or harder.

THEOREM 1.2. *For two distributions p, q, over $[n]$ with $b \geq ||p||_2^2, ||q||_2^2$, there is an algorithm which distinguishes the case that $||p-q||_2 \leq \varepsilon$ from the case that $||p-q||_2 \geq 2\varepsilon$ when given $O(\sqrt{b}/\varepsilon^2)$ samples from p and q with probability at least $2/3$. This is information theoretically optimal, as distinguishing the case that $p = q$ from the case that $||p-q||_2 > 2\varepsilon$ requires $\Omega(\sqrt{b}/\varepsilon^2)$ samples.*

We note that both the upper and lower bounds of the above theorem continue to hold if b is defined to be an upper bound on $||p||_\infty, ||q||_\infty$; the upper bound trivially holds because, for all p, $||p||_2^2 \leq \max_i p_i$, and the lower bound holds because the specific lower bound instance we construct consists of nearly uniform distributions for which $||p||_2^2 \geq \max_i p_i/2$. See Proposition 3.1 and the discussion following it for analysis of our algorithm as an *estimator* for ℓ_2 distance.

The $\ell_2 \to \ell_1$ testing approach. Recall that the ℓ_1 closeness tester in [BFR+00] proceeds in two steps: In the first step, it "filters" the elements of p and q that are "b-heavy", i.e., have probability mass at least b – for an appropriate value of b. (This step essentially amounts to *learning* the heavy parts of p and q.) In the second step, it uses an ℓ_2 closeness tester applied to the "light" parts of p and q. The ℓ_2 tester used in [BFR+00] is a generalization of a tester proposed in [GR00].

Using such a two step approach, Theorem 1.2 can be used as a black-box to obtain an ℓ_1 closeness tester with sample complexity $O(n^{2/3}\log n/\varepsilon^2)$. This can further be improved to $O(n^{2/3}/\varepsilon^2)$ by improving the "filtering" algorithm of [BFR+00]; in Appendix A we describe an optimal "filtering" algorithm, which might be applicable in other settings. Curiously, since the sample complexity of both the improved filtering algorithm, and the ℓ_2 tester are optimal, the corresponding sample complexity of $O(n^{2/3}/\varepsilon^2)$ for the ℓ_1 testing problem seems to be the best that could possibly be achieved via this reduction-based approach. This suggests that, in some sense, our novel (and more direct) approach underlying Theorem 1.1 is necessary to achieve the optimal ε-dependence for the ℓ_1 testing problem.

Structure of the paper. In Section 2 we present our ℓ_1 tester, and in Section 3 we present our ℓ_2 tester. In Section 4 we prove the information theoretic lower bounds, establishing the optimality of both testers. The details of the reduction–based (though suboptimal) ℓ_1 closeness tester can be found in the appendix.

Remark. Throughout our technical sections, we employ the standard "Poissonization" approach: namely, we assume that, rather than drawing k independent samples from a distribution, we first select k' from $\mathrm{Poi}(k)$, and then draw k' samples. This Poissonization makes the number of times different elements occur in the sample independent, simplifying the analysis. As $\mathrm{Poi}(k)$ is tightly concentrated about k, we can carry out this Poissonization trick without loss of generality at the expense of only subconstant factors in the sample complexity, and adding only $o(\frac{1}{\mathrm{poly}})$ probability of failure.

2 Closeness testing in ℓ_1 norm

We begin by describing our ℓ_1 closeness testing algorithm:

> Input: A constant C and m samples from distributions p, q, with X_i, Y_i denoting the number of occurrences of the ith domain elements in the samples from p and q, respectively.
>
> 1. Define
>
> $$Z = \sum_i \frac{(X_i - Y_i)^2 - X_i - Y_i}{X_i + Y_i}. \tag{2.1}$$
>
> 2. If $Z \leq C \cdot \sqrt{m}$ then output EQUAL, else output DIFFERENT.

In Equation 2.1, we interpret terms where $X_i = Y_i = 0$ to be 0.

The following proposition characterizes the performance of the above tester, establishing the algorithmic portion of Theorem 1.1.

PROPOSITION 2.1. *There exist absolute constants C, C' such that the above algorithm, on input C and a set of $\mathrm{Poi}(m)$ samples drawn from two distributions, p, q, supported on $[n]$, will correctly distinguish the case that*

$p = q$ from the case that $||p-q||_1 \geq \varepsilon$, with probability at least $2/3$ provided that $m \geq C' \max\{n^{2/3}/\varepsilon^{4/3}, n^{1/2}/\varepsilon^2\}$.

We will show that the error probability of the above algorithm is $O(\frac{1}{C^2})$, hence for a suitable constant C the tester succeeds with probability $\frac{2}{3}$. (Repeating the tester and taking the majority answer results in an exponential decrease in the error probability.)

The form of the right hand side of Eq. (2.1) is rather similar to our ℓ_2 distance tester (given in the next section), though the difference in normalization is crucial. However, though we do not prove corresponding theorems here, the right hand side of Eq. (2.1) can have a variety of related forms while yielding similar results, with possibly improved constants. For example, one could use $\sum_i |X_i - Y_i| - f(X_i + Y_i)$, where $f(j)$ is the expected absolute difference between the number of heads and the number of tails in j fair coin flips, which is $\binom{j-1}{\lfloor (j-1)/2 \rfloor} \frac{j}{2^{j-1}}$.

Previously, Acharya et al. had analyzed a very similar tester, of the form $\sum_i \frac{(X_i - Y_i)^2 - X_i - Y_i}{X_i + Y_i + 1}$ in a different setting of the closeness testing problem [ADJ+12]. The numerator of this expression is naturally motivated as a variant of a chi-squared test—the sum of the squares of several independent expressions, each of which has expectation 0. It is standard to analyze such tests by bounding the expectation and variance of the computed sum. In both our case and theirs, we must rescale the terms in the sum so as to avoid the following case: despite p and q being far from each other, there is a single domain element i with most of the probability mass, $p_i = q_i = \frac{1}{2}$, and the contribution of the ith term, $(X_i - Y_i)^2 - X_i - Y_i$, adds so much variance to the resulting sum that we lose the ability to distinguish what is happening on the rest of the domain. To avoid this, we divide by $(X_i + Y_i)^\alpha$, where any α between $\frac{1}{2}$ and 1 solves this problem. Thus, while the tester itself is a natural choice, the very tight analysis is the surprising part of this work, in particular since previous attempts at tight analysis used rather more complicated testers [BFR+00].

In the remainder of this section we prove Proposition 2.1. First, letting p_i, q_i respectively denote the probabilities of the ith elements in each distribution, note that if $p_i = q_i$ then the expectation of the sum in Eq. (2.1) is 0, as can be seen by conditioning the summand for each i on the value of $X_i + Y_i$: subject to this, X_i, Y_i can be seen as the number of heads and tails respectively found in $X_i + Y_i$ fair coin flips, and $\mathbb{E}[(X_i - Y_i)^2]$ is 4 times the variance of X_i alone, which is a quarter of the number of coin flips, and thus the expression in total has expectation 0.

When $p \neq q$, we use the following lemma to bound from below the expected value of our estimator in terms of $\|p - q\|_1$.

LEMMA 2.1. *For Z as defined in Equation 2.1, $\mathbb{E}[Z] \geq \frac{m^2}{4n+2m} \|p - q\|_1^2$.*

Proof. Conditioned on $X_i + Y_i = j$, for some j, we have that X_i is distributed as the number of heads in the distribution $\text{Binom}(j, \frac{p_i}{p_i+q_i})$. For the distribution $\text{Binom}(j, \alpha)$, the expected value of the square of the difference between the number of heads and tails can be easily seen to be $4j^2(\frac{1}{2} - \alpha)^2 + 4j\alpha(1 - \alpha)$; we subtract j from this because of the $-X_i - Y_i$ term in the numerator of Eq. (2.1) to yield $4(j^2 - j)(\frac{1}{2} - \alpha)^2$, and divide by j because of the denominator of Eq. (2.1) to yield $4(j - 1)(\frac{1}{2} - \alpha)^2$. Plugging in $\alpha = \frac{p_i}{p_i+q_i}$ yields $(j - 1)(\frac{p_i - q_i}{p_i + q_i})^2$. Thus the expected value of the summand of Eq. (2.1), for a given i, conditioned on $X_i + Y_i = j$ is this last expression, if $j \neq 0$, and 0 otherwise. Thus the expected value of the summand across all j, since $\mathbb{E}[j] = m(p_i + q_i)$, equals

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

where we have used the fact that $\Pr[X_i + Y_i = 0] = e^{-m(p_i+q_i)}$. Gathering terms, we conclude that the expectation of each term of Eq. (2.1) equals

$$\frac{(p_i - q_i)^2}{p_i + q_i} m \left(1 - \frac{1 - e^{-m(p_i+q_i)}}{m(p_i + q_i)}\right) \tag{2.2}$$

Defining the function $g(\alpha) = \alpha \Big/ \left(1 - \frac{1-e^{-\alpha}}{\alpha}\right)$, this expression becomes $m^2 \frac{(p_i - q_i)^2}{g(m(p_i+q_i))}$, and we bound its sum via Cauchy-Schwarz as

$$\begin{aligned} & m^2 \left(\sum_i \frac{(p_i - q_i)^2}{g(m(p_i + q_i))} \right) \left(\sum_i g(m(p_i + q_i)) \right) \\ & \geq m^2 \left(\sum_i |p_i - q_i| \right)^2 \end{aligned}$$

It is straightforward to bound $g(\alpha) \leq 2 + \alpha$, leading to $\sum_i g(m(p_i+q_i)) \leq 4n+2m$, since the support of each distribution is at most n and each has total probability mass 1. Thus the expected value of the left hand side of Eq. (2.1) is at least $\frac{m^2}{4n+2m} \left(\sum_i |p_i - q_i|\right)^2$.

We now bound the variance of the ith term of Z.

LEMMA 2.2. *For Z as defined in Equation Eq. (2.1), $\text{Var}[Z] \leq 2 \min\{n, m\} + \sum_i 5m \frac{(p_i - q_i)^2}{p_i + q_i}$.*

Proof. To bound the variance of the ith term of Z, we will split this variance calculation into two parts: the variance conditioned on $X_i+Y_i = j$, and the component of the variance due to the variation in j. Letting

$$f(X_i, Y_i) = \frac{(X_i - Y_i)^2 - X_i - Y_i}{X_i + Y_i},$$

we have that

$$\begin{aligned}\mathrm{Var}[f(X,Y)] \leq \max_j \left(\mathrm{Var}[f(X,Y)|X+Y=j]\right) \\ + \mathrm{Var}_j[\mathbb{E}[f(X,Y)|X+Y=j]].\end{aligned}$$

We now bound the first term; since $(X_i - Y_i)^2 = (j-2Y_i)^2$, and Y_i is distributed as $\mathrm{Binom}(j; \frac{q_i}{p_i+q_i})$ where for convenience we let $\alpha = \frac{q_i}{p_i+q_i}$ we can compute the variance of $(j-2Y_i)^2$ from standard expressions for the moments of the Binomial distribution as

$$\mathrm{Var}[(j-2Y_i)^2] = 16j(j-1)\alpha(1-\alpha)\left((j-\frac{3}{2})(1-2\alpha)^2 + \frac{1}{2}\right)$$

We bound this expression, since $\alpha(1-\alpha) \leq \frac{1}{4}$ and $j - \frac{3}{2} < j - 1 < j$ as $j^2(2 + 4j(1-2\alpha)^2)$. Because the denominator of the ith term of Eq. (2.1) is $X_i + Y_i = j$, we must divide this by j^2, make it 0 when $j = 0$, and take its expectation as j is distributed as $\mathrm{Poi}(m(p_i+q_i))$, yielding:

$$\mathrm{Var}[f(X_i,Y_i)|X_i+Y_i=j] \leq 2(1-e^{-m(p_i+q_i)})+4m\frac{(p_i-q_i)^2}{p_i+q_i}.$$

We now consider the second component of the variance—the contribution to the variance due to the variation in the sum $X_i + Y_i$. Since for fixed j, as noted above, we have Y_i distributed as $\mathrm{Binom}(j; \frac{q_i}{p_i+q_i})$, where for convenience we let $\alpha = \frac{q_i}{p_i+q_i}$, we have

$$\begin{aligned}\mathbb{E}[(X_i - Y_i)^2] =& \mathbb{E}[j^2 - 4jY_i + 4Y_i^2] \\ =& j^2 - 4j^2\alpha + 4(j\alpha - j\alpha^2 + j^2\alpha^2) \\ =& j^2(1-2\alpha)^2 + 4j\alpha(1-\alpha).\end{aligned}$$

As in Eq. (2.1), we finally subtract j and divide by j to yield $(j-1)(1-2\alpha)^2$, except with a value of 0 when $j = 0$ by definition; however, note that replacing the value at $j = 0$ with 0 can only lower the variance. Since the sum $j = X_i+Y_i$ is drawn from a Poisson distribution with parameter $m(p_i + q_i)$, we thus have:

$$\begin{aligned}\mathrm{Var}\left[\mathbb{E}[f(X_i,Y_i)|X_i+Y_i=j]\right] &\leq m(p_i+q_i)(1-2\alpha)^4 \\ &\leq m(p_i+q_i)(1-2\alpha)^2 \\ &= m\frac{(p_i-q_i)^2}{p_i+q_i}.\end{aligned}$$

Summing the final expressions of the previous two paragraphs yields a bound on the variance of the ith term of Eq. (2.1) of

$$2(1-e^{-m(p_i+q_i)}) + 5m\frac{(p_i-q_i)^2}{p_i+q_i}.$$

We note that since $1 - e^{-m(p_i+q_i)}$ is bounded by both 1 and $m(p_i + q_i)$, the sum of the first part is bounded as

$$\sum_i 2(1-e^{-m(p_i+q_i)}) \leq 2\min\{n, m\}.$$

This completes the proof.

We now complete our proof of Proposition 2.1, establishing the upper bound of Theorem 1.1.

Proof. [Proof of Proposition 2.1] With a view towards applying Chebyshev's inequality, we compare the square of the expectation of Z to its variance. From Lemma 2.1, the expectation equals

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

which we showed is at least $\frac{m^2}{4n+2m}\|p-q\|_1^2$; from Lemma 2.2, the variance is at most

$$2\min\{n,m\} + \sum_i 5m\frac{(p_i-q_i)^2}{p_i+q_i}.$$

We consider the second part of the variance expression. It is clearly bounded by $10m$, so when $m < n$ the first expression dominates. Otherwise, assume that $m \geq n$. Consider the case when our bound on the expectation, $\frac{m^2}{4n+2m}\|p-q\|_1^2$, is at least 2, namely that $m = \Omega(\|p-q\|_1^{-2})$. Thus, with a view towards applying Chebyshev's inequality, we can bound the square of the expectation by:

$$\begin{aligned}&\left(\sum_i \frac{(p_i-q_i)^2}{p_i+q_i} m\left(1 - \frac{1-e^{-m(p_i+q_i)}}{m(p_i+q_i)}\right)\right)^2 \\ &\geq \sum_i \frac{(p_i-q_i)^2}{p_i+q_i} m\left(1 - \frac{1-e^{-m(p_i+q_i)}}{m(p_i+q_i)}\right)\cdot 2.\end{aligned}$$

For those i for which the multiplier $\left(1 - \frac{1-e^{-m(p_i+q_i)}}{m(p_i+q_i)}\right)\cdot 2$ is greater than 1, we have that the ith term here is greater than the ith term of the expression for the variance, $\sum_i \frac{(p_i-q_i)^2}{p_i+q_i}m$; otherwise, we have $1 - \frac{1-e^{-m(p_i+q_i)}}{m(p_i+q_i)} \leq \frac{1}{2}$ which implies $m(p_i + q_i) \leq 2$, and thus the sum of the remaining

terms is bounded by $2n$, which is dominated by the first expression in the variance, $2\min\{n,m\}$ in the case under consideration, where $m \geq n$. Thus we need only compare the square of the expectation, which is at least $\frac{m^2}{4n+2m}\|p-q\|_1^2 = \frac{m^2}{O(\max\{n,m\})}\|p-q\|_1^2$, to $O(\min\{n,m\})$, yielding, when $m < n$ a bound $m = \Omega(n^{2/3}/\|p-q\|_1^{4/3})$, and when $m \geq n$ a bound $m = \Omega(n^{1/2}/\|p-q\|_1^2)$; note that in the latter case, this implies $m = \Omega(\|p-q\|_1^{-2})$, which we needed in the derivation above.

3 Robust ℓ_2 testing

In this section, we give an optimal algorithm for robust closeness testing of distributions with respect to ℓ_2 distance. For distributions p and q over $[n]$ with ℓ_2^2 norm at most b (i.e., $\sum_i p_i^2 \leq b$, and $\sum_i q_i^2 \leq b$), the algorithm when given $O(\sqrt{b}/\varepsilon^2)$ samples will distinguish the case that $||p-q||_2 \leq \varepsilon$ from the case that $||p-q|| \geq 2\varepsilon$, with high probability. Since $||p||_2^2 \leq \max_i p_i$, this sample complexity is also bounded by the corresponding expression with b replaced by a bound on the maximum probability of an element of p or q. As we show in Section 4, this sample complexity is optimal even for the easier testing problem of distinguishing the case that the ℓ_2 distance is 0 versus at least ε.

Our algorithm is a very natural linear estimator and is similar to the ℓ_2 tester of [BFR+00].

> Input: m samples from distributions p, q, with X_i, Y_i denoting the number of occurrences of the ith domain elements in the samples from p and q, respectively.
> Output: an estimate of $||p-q||_2$.
>
> 1. Define $Z = \sum_i (X_i - Y_i)^2 - X_i - Y_i$.
> 2. Return $\frac{\sqrt{Z}}{m}$.

The following proposition characterizes the performance of the above estimator, establishing the algorithmic portion of Theorem 1.2 from the observation that $||p-q||_4^2 \leq ||p-q||_2^2$.

PROPOSITION 3.1. *There exists an absolute constant c such that the above estimator, when given* Poi(m) *samples drawn from two distributions, p, q will, with probability at least 3/4, output an estimate of $||p-q||_2$ that is accurate to within $\pm\varepsilon$ provided that $m \geq c\left(\frac{\sqrt{b}}{\varepsilon^2} + \frac{\sqrt{b}||p-q||_4^2}{\varepsilon^4}\right)$, where b is an upper bound on $||p||_2^2, ||q||_2^2$.*

Proof. Letting X_i, Y_i denote the number of occurrences of the ith domain elements in the samples from p and q, respectively. Define $Z_1 = (X_i - Y_i)^2 - X_i - Y_i$. Since X_i is distributed as Poi$(m \cdot p_i)$, $\mathbb{E}[Z_i] = m^2 \cdot |p_i - q_i|^2$, hence Z is an unbiased estimator for $m^2||p-q||_2^2$.

We compute the variance of Z_i via a straightforward calculation involving standard expressions for the moments of a Poisson distribution[1] : $\mathrm{Var}[Z_i] = 4(p_i - q_i)^2(p_i+q_i)m^3 + 2(p_i+q_i)^2m^2$.

Hence

$$\begin{aligned}\mathrm{Var}[Z] &= \sum_i \mathrm{Var}[Z_i] \\ &= \sum_i \left(4m^3(p_i-q_i)^2(p_i+q_i) + 2m^2(p_i+q_i)^2\right).\end{aligned}$$

By Cauchy-Schwarz, and since $\sum_i (p_i+q_i)^2 \leq 4b$, we have

$$\begin{aligned}\sum_i (p_i-q_i)^2(p_i+q_i) &\leq \sqrt{\sum_i (p_1-q_i)^4 \sum_i (p_i+q_i)^2} \\ &\leq 2||p-q||_4^2\sqrt{b}.\end{aligned}$$

Hence

$$\mathrm{Var}[Z] \leq 8m^3\sqrt{b}||p-q||_4^2 + 8m^2b.$$

By Chebyshev's inequality, the returned estimate of $||p-q||_2$ will be accurate to within $\pm\varepsilon$ with probability at least 3/4 provided $\varepsilon^2 m^2 \geq 2\sqrt{8m^3\sqrt{b}||p-q||_4^2 + 8m^2b}$, which holds whenever

$$m \geq 6\frac{\sqrt{b}}{\varepsilon^2} + 32\frac{\sqrt{b}||p-q||_4^2}{\varepsilon^4},$$

since $m \geq x+y$ implies $m^2 \geq mx + y^2$, for any $x, y \geq 0$.

A slightly different kind of result is obtained if we parameterize by $B = \max\{\max_i p_i, \max_i q_i\}$ instead of b—where we note that $B \geq b$. We can replace the Cauchy Schwarz inequality of the proof above with $\sum_i (p_i-q_i)^2(p_i+q_i) \leq 2B\sum_i (p_i-q_i)^2 = 2B||p-q||_2^2$, yielding, analogously to above, that the tester is accurate to $\pm\epsilon$ when given $c(\frac{\sqrt{B}}{\varepsilon^2} + \frac{B||p-q||_2^2}{\varepsilon^4})$ samples. This matches the lower-bound of $\Omega(\frac{\sqrt{B}}{\varepsilon^2})$ provided the second term is not much larger than the first, namely when $\frac{||p-q||_2}{\varepsilon} = O(B^{-1/2})$. Thus our algorithm approximates ℓ_2 distance to within ϵ using the optimal number of samples, provided the ℓ_2 distance is not a $B^{-1/2}$ factor greater than ϵ. For greater distances, we have not shown optimality.

[1] This calculation can be performed in Mathematica, for example, via the expression Variance[TransformedDistribution[(X - Y)^2 - X - Y, {X \[Distributed] PoissonDistribution[m p], Y \[Distributed] PoissonDistribution[m q]}]]

An $O(n^{2/3}/\varepsilon^2)$ ℓ_1-tester. As noted in the introduction, Theorem 1.2 combined with the two step approach of [BFR+00], immediately leads to an ℓ_1 tester for distinguishing the case that $p = q$ from $||p - q||_1 \geq \varepsilon$ with sample complexity $O(n^{2/3} \log n/\varepsilon^2)$. One can use Theorem 1.2 to obtain an ℓ_1 tester with sample complexity $O(n^{2/3}/\varepsilon^2)$ – i.e., saving a factor of $\log n$ in the sample complexity. While this does not match the $O(\max\{n^{2/3}/\varepsilon^{4/3}, n^{1/2}/\varepsilon^2\})$ performance of the ℓ_1 tester described in Section 2, the ideas used to remove the $\log n$ factor might be applicable to other problems, and we give the details in Appendix A.

4 Lower bounds

In this section, we present our lower bounds for closeness testing under ℓ_1 and ℓ_2 norms. We derive the results of this section as applications of the machinery developed in [Val11] and [VV13].

The lower bounds for ℓ_1 testing require the following definition:

DEFINITION 4.1. *The (k, k)-based moments $m(r, s)$ of a distribution pair (p, q) are $k^{r+s} \sum_{i=1}^{n} p_i^r q_i^s$.*

THEOREM 4.1. ([VAL11], THEOREM 4.18) *If distributions p_1^+, p_2^+, p_1^-, p_2^- have probabilities at most $1/1000k$, and their (k, k)-based moments m^+, m^- satisfy*

$$\sum_{r+s\geq 2} \frac{|m^+(r,s) - m^-(r,s)|}{\lfloor \frac{r}{2} \rfloor! \lfloor \frac{s}{2} \rfloor! \sqrt{1 + \max\{m^+(r,s), m^-(r,s)\}}} < \frac{1}{360},$$

then the distribution pair (p_1^+, p_2^+) cannot be distinguished with probability $13/24$ from (p_1^-, p_2^-) by tester that takes $Poi(k)$ samples from each distribution.

The optimality of our ℓ_1 tester, establishing the lower bound of Theorem 1.1, follows from the following proposition together with the lower bound of $\sqrt{n}/\varepsilon^2$ for testing uniformity given in [Pan08].

PROPOSITION 4.1. *If $\varepsilon \geq 4^{3/4} n^{-1/4}$, then $\Omega(n^{2/3}\varepsilon^{-4/3})$ samples are needed for 0-vs-ε closeness testing under the ℓ_1 norm.*

Proof. Let $b = \varepsilon^{4/3}/n^{2/3}$ and $a = 4/n$, where the restriction on ε yields that $b \geq a$. Let p and q be the distributions

$$p = b\mathbf{1}_A + \varepsilon a \mathbf{1}_B \qquad q = b\mathbf{1}_A + \varepsilon a \mathbf{1}_C$$

where A, B and C are disjoint subsets of size $(1 - \varepsilon)/b$, $1/a$ and $1/a$—where the notation $\mathbf{1}_A$ denotes the indicator function that is 1 on the set A. Then $\|p - q\|_1 = 2\varepsilon$. Let $k = cn^{2/3}\varepsilon^{-4/3}$ for a small enough constant $0 < c < 1$, so that $\|p\|_\infty = \|q\|_\infty = b \leq \frac{1}{1000k}$, since $b \geq a$.

Let $(p_1^+, p_2^+) = (p, p)$ and $(p_1^-, p_2^-) = (p, q)$, so that they have (k, k)-based moments

$$m^+(r,s) = k^t(1-\varepsilon)b^{t-1} + k^t\varepsilon^t a^{t-1} \quad \text{and}$$
$$m^-(r,s) = k^t(1-\varepsilon)b^{t-1},$$

for $r, s \geq 1$, where $t = r + s$. We have the inequality

$$\frac{|m^+(r,s) - m^-(r,s)|}{\sqrt{1 + \max\{m^+(r,s), m^-(r,s)\}}} \leq \frac{k^t\varepsilon^t a^{t-1}}{\sqrt{k^t(1-\varepsilon)b^{t-1}}}.$$

For $t \geq 2$, it is at most $k^{t/2}\varepsilon^t a^{t-1}/b^{(t-1)/2} \leq c^{t/2}4^{(2t-1)/3}$ (where we used that $\varepsilon \geq 4/n$). Further, when one of r or s is 0, the moments are equal, since p and q are permutations of each other, yielding a contribution of 0 to the expression of Theorem 4.1. Thus the expression in Theorem 4.1 is bounded by $O(c)$ as the sum of a geometric series (in two dimensions), and thus the distribution pairs (p, p) and (p, q) are indistinguishable by Theorem 4.1.

The optimality of our ℓ_2 tester will follow from the following result from [VV13]:

THEOREM 4.2. ([VV13], THEOREM 3) *Given a distribution p, and associated values $\varepsilon_i \in [0, p_i]$, define the distribution over distributions, $Q_{p,\varepsilon}$ by the following process: for each domain element i, randomly choose $q_i = p_i \pm \varepsilon_i$, and then normalize q to be a distribution. There exists a constant c such that it takes at least $c\left(\sum_i \frac{\varepsilon_i^4}{p_i^2}\right)^{-1/2}$ samples to distinguish p from a sample drawn from a random element of $Q_{p,\varepsilon}$ with success probability at least $2/3$.*

The following proposition establishes the lower bound of Theorem 1.2, showing the optimality of our ℓ_2 tester. Note that if $\max_i p_i \leq b$ and $\max_i q_i \leq b$, then $||p - q||_2 \leq \sqrt{2b}$, hence the testing problem is trivial unless $\varepsilon \leq \sqrt{2b}$.

PROPOSITION 4.2. *For any $b \in [0, 1]$, and $\varepsilon \leq \sqrt{b}$, there exists a distribution p_b and a family of distributions $T_{p,\varepsilon}$ such that for a $q \leftarrow T$ chosen uniformly at random, the following hold:*

- *$||p||_2^2 \in [b/2, b]$ and $\max_i p_i \in [b/2, b]$ and with probability at least $1 - o(1)$, $||q||_2^2 \in [b/2, b]$ and $\max_i q_i \in [b/2, b]$.*
- *With probability at least $1 - o(1)$, $||p - q||_2 \geq \varepsilon/2$.*
- *No algorithm can distinguish a set of $k = c\frac{\sqrt{b}}{\varepsilon^2}$ samples from q from a set drawn from p with*

probability of success greater than $3/4$*, hence no algorithm can distinguish sets of* k *samples drawn from the pair* (p,p) *versus drawn from* (p,q) *with this probability.*

Proof. Assume for the sake of clarity that $1/b$ is an integer. The proof follows from applying Theorem 4.2 to the distribution p consisting of $1/b$ domain elements that each occur with probability b, and setting $\varepsilon_i = \varepsilon\sqrt{b}$. Letting Q be the family of distributions defined in Theorem 4.2 associated to p and the ε_i's, note that with probability $1 - o(1)$ it is the case that the first and second conditions in the proposition statement are satisfied. Additionally, the theorem guarantees that p cannot be distinguished with probability $> 2/3$ from such a q given a sample of size m provided that $m < c\left(\sum_i \frac{\varepsilon_i^4}{p_i^2}\right)^{-1/2} = c\frac{\sqrt{b}}{\varepsilon^2}$.

Given an algorithm that could distinguish, with probability at least $3/4 > 2/3 + o(1)$, whether $||p' - q'||_2 = 0$ versus $||p' - q'||_2 \geq \varepsilon/2$, using $m = O(\sqrt{b}/\varepsilon^2)$ samples drawn from each of p', q', one could use it to perform the above (impossible) task of distinguishing with probability greater than $2/3$ whether a set of samples was drawn from p, versus a random $q \leftarrow Q$ by running the hypothetical ℓ_2 tester on the set of samples, and a set drawn from p.

References

[ADJ+11] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan. Competitive closeness testing. *Journal of Machine Learning Research - Proceedings Track*, 19:47–68, 2011.

[ADJ+12] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, S. Pan, and A. Suresh. Competitive classification and closeness testing. In *COLT*, 2012.

[AIOR09] A. Andoni, P. Indyk, K. Onak, and R. Rubinfeld. External sampling. In *ICALP (1)*, pages 83–94, 2009.

[Bat01] T. Batu. *Testing Properties of Distributions*. PhD thesis, Cornell University, 2001.

[BDKR02] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating entropy. In *ACM Symposium on Theory of Computing*, pages 678–687, 2002.

[BFF+01] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *Proc. 42nd IEEE Symposium on Foundations of Computer Science*, pages 442–451, 2001.

[BFR+00] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000.

[BFR+13] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *J. ACM*, 60(1):4, 2013.

[BKR04] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *ACM Symposium on Theory of Computing*, pages 381–390, 2004.

[BNNR11] K. D. Ba, H. L. Nguyen, H. N. Nguyen, and R. Rubinfeld. Sublinear time algorithms for earth mover's distance. *Theory Comput. Syst.*, 48(2):428–442, 2011.

[DDS+13] C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *SODA*, 2013.

[GR00] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, Electronic Colloquium on Computational Complexity, 2000.

[ILR12] P. Indyk, R. Levi, and R. Rubinfeld. Approximating and Testing k-Histogram Distributions in Sublinear Time. In *PODS*, pages 15–22, 2012.

[LRR11] R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. In *ICS*, pages 179–194, 2011.

[Ona09] K. Onak. Testing distribution identity efficiently. *CoRR*, abs/0910.3243, 2009.

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

[Rub12] R. Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

[Val08] P. Valiant. *Testing Symmetric Properties of Distributions*. PhD thesis, M.I.T., 2008.

[Val11] P. Valiant. Testing symmetric properties of distributions. *SIAM J. Comput.*, 40(6):1927–1968, 2011.

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

[VV11b] G. Valiant and P. Valiant. The power of linear estimators. In *FOCS*, 2011.

[VV13] G. Valiant and P. Valiant. Instance-by-instance optimal identity testing. *ECCC*, http://eccc.hpi-web.de/report/2013/111/, 2013.

Appendix

A An $O(n^{2/3}/\varepsilon^2)$ ℓ_1-tester

In this section, we show how we to obtain an ℓ_1 closeness tester with sample complexity $O(n^{2/3}/\varepsilon^2)$, by using essentially the same approach as [BFR+00].

Recall that the ℓ_1 closeness tester in [BFR+00] proceeds in two steps: In the first step, it "filters" the elements of p and q that are "b-heavy", i.e., have probability mass at least b – for an appropriate value of

b. (This step essentially amounts to *learning* the heavy parts of p and q.) In the second step, it uses an ℓ_2 closeness tester to test closeness of the "light" parts of p and q. (Note that in the second step the ℓ_2 tester needs to be called with error parameter $\varepsilon/\sqrt{n}$.)

Our improvement over [BFR+00] is two fold: First, we perform the first step (learning) in a more efficient way (using a different algorithm). Roughly, this improvement allows us to save a $\log n$ factor in the sample complexity. Second, we apply our optimal ℓ_2 tester in the second step.

Regarding the first step, note that the heavy part of p and q has support size at most $2/b$. Roughly, we show that the heavy part can be learned to ℓ_1 error ε using $O((1/b)/\varepsilon^2)$ samples (which is the best possible) – without knowing a priori which elements are heavy versus light. The basic idea to achieve this is as follows: rather than inferring *all* the heavy elements (which inherently incurs an extra $\log(1/b)$ factor in sample complexity, due to coupon collector's problem), a small fraction of heavy elements are allowed to be undetected; this modification requires a more involved calculation for heavy elements and a relaxed definition for light elements.

The first step of our ℓ_1 test uses $s_1 = O((1/b)/\varepsilon^2)$ samples and the second step uses $s_2 = O(\sqrt{b}/\tilde{\varepsilon}^2)$ samples, where $\tilde{\varepsilon} = \varepsilon/\sqrt{n}$. The overall sample complexity is $s_1 + s_2$, which is minimized for $b = \Theta(n^{-2/3})$ for a total sample complexity of $O(n^{2/3}/\varepsilon^2)$. We remark that since the sample complexity of each step is individually optimal, our achieved bound seems to be the best that could possibly be achieved via this reduction-based approach, supporting the view that, in some sense, the more direct approach of Section 2 is necessary to achieve the optimal dependence on ε.

In the following subsections we provide the details of the algorithm and its analysis.

We start with the following definition:

DEFINITION A.1. *A distribution p is (b, C)-*bounded *if $\|p\|_2^2 \le Cb$.*

A.1 Heavy elements. We denote by $\hat{p}$ (resp. $\hat{q}$) the empirical distribution obtained after taking m independent samples from p (resp. q). We classify elements into the following subsets:

- Observed heavy $H(\hat{p}) = \{i \mid \hat{p}_i \ge b\}$ versus observed light $L(\hat{p}) = \{i \mid \hat{p}_i < b\}$.
- Truly heavy $\overline{H}(p) = \{i \mid p_i \ge b/2\}$ versus truly light $\overline{L}(p) = \{i \mid p_i < b/2\}$.

(Note the threshold for the observed distribution is b, while for the true distribution is $b/2$.)

Consider the random variables

$$D_i = |\hat{p}_i - \hat{q}_i| - |p_i - q_i|, \quad D(A) = \left|\sum_{i \in A} D_i\right|.$$

We sometimes write $D(AB)$ for $D(A \cap B)$.

We will also use the shorthand $\hat{H} = H(\hat{p}) \cup H(\hat{q})$. We want to show that

$$\|\hat{p} - \hat{q}\|_{H(\hat{p}) \cup H(\hat{q})} \approx_\varepsilon \|p - q\|_{H(\hat{p}) \cup H(\hat{q})}$$

with high probability. To do this, we use the bound

$$\begin{aligned} D(\hat{H}) \le & D(\hat{H}\overline{H}(p)\overline{H}(q)) + D(\hat{H}\overline{H}(p)\overline{L}(q)) \\ & + D(\hat{H}\overline{L}(p)\overline{H}(q)) + D(\hat{H}\overline{L}(p)\overline{L}(q)). \end{aligned} \tag{A.1}$$

The first three terms on RHS of Eq. (A.1) will be bounded by Corollary A.1 below. We start with the following simple claim:

CLAIM A.1. *For any $i \in [n]$,*

$$\mathbb{E}[D_i^2] \le \frac{p_i + q_i}{m}. \tag{A.2}$$

Proof. Expand the LHS of Eq. (A.2) as

$$\mathbb{E}(\hat{p}_i - \hat{q}_i)^2 - 2|p_i - q_i|\mathbb{E}|\hat{p}_i - \hat{q}_i| + |p_i - q_i|^2.$$

Since

$$\begin{aligned} \mathbb{E}(\hat{p}_i - \hat{q}_i)^2 &= \mathrm{Var}[\hat{p}_i - \hat{q}_i] + (\mathbb{E}[p_i - q_i])^2 \\ &= \frac{p_i(1-p_i) + q_i(1-q_i)}{m} + |p_i - q_i|^2, \end{aligned}$$

the LHS of Eq. (A.2) is at most

$$\frac{p_i + q_i}{m} - 2|p_i - q_i|(\mathbb{E}|\hat{p}_i - \hat{q}_i| - |p_i - q_i|).$$

The result follows by the elementary fact $\mathbb{E}|X| \ge |\mathbb{E}X|$ applied to $X = \hat{p}_i - \hat{q}_i$.

COROLLARY A.1. *If we use $m \ge 4/(\varepsilon^2 b\delta)$ samples, then for any (possibly random) $H \subseteq \overline{H}(p)$, we have*

$$D(H) \le \varepsilon$$

except with probability δ.

Proof. By Cauchy–Schwarz,

$$D(H)^2 = \left(\sum_{i \in H} D_i\right)^2 \le |\overline{H}(p)| \sum_{i \in \overline{H}(p)} D_i^2.$$

Now we take expectation on both sides. Since $\sum_i \mathbb{E}[D_i^2] \le \sum_i (p_i + q_i)/m \le 2/m$, and $|\overline{H}(p)| \le 2/b$, we have $\mathbb{E}[D(H)^2] \le \varepsilon^2\delta$. Hence

$$\Pr[D(H) \ge \varepsilon] = \Pr[D(H)^2 \ge \varepsilon^2] \le \delta$$

by Markov's inequality.

We bound the last term on the RHS of Eq. (A.1) by

(A.3)
$$D(\hat{H}L(p)L(q)) \le D(H(\hat{p})H(\hat{q})\overline{L}(p)\overline{L}(q)) + D(H(\hat{p})L(\hat{q})\overline{L}(p)\overline{L}(q)) + D(L(\hat{p})H(\hat{q})\overline{L}(p)\overline{L}(q)).$$

The RHS will be bounded by Corollaries A.2 and A.3 below.

CLAIM A.2. *For any $p_i \le b/2$, any $t \ge 1$, with $m = 1/(\varepsilon^2 b)$ samples,*

$$\Pr[\hat{p}_i \ge tb] \ll \frac{\varepsilon^2}{t^2}\frac{p_i}{b}. \tag{A.4}$$

Proof. Note that $\hat{p}_i$ has distribution $\mathrm{Binom}(m, p_i)/m$, so by Chebyshev's inequality,

$$\Pr[\hat{p}_i \ge tb] \le \Pr[|\hat{p}_i - p_i| \ge tb/2] \le \frac{\mathrm{Var}[\hat{p}_i]}{(tb/2)^2} \le \frac{4p_i}{m(tb)^2} = \frac{4\varepsilon^2}{t^2}\frac{p_i}{b}.$$

LEMMA A.1. *For any $\delta > 0$, for any $\varepsilon \ll \delta$, using $m \gg 1/(\varepsilon^2 b)$ samples,*

$$\|\hat{p}\|_{\overline{L}(p)H(\hat{p})} \le \varepsilon$$

except with probability δ.

Proof.

$$\begin{aligned}\mathbb{E}\|\hat{p}\|_{\overline{L}(p)\cap H(\hat{p})} &= \sum_{i\in\overline{L}(p)} \mathbb{E}[\hat{p}_i \cdot \mathbf{1}_{p\ge b}] \\ &= \sum_{i\in\overline{L}(p)}\sum_{j\ge 0} \mathbb{E}[\hat{p}_i \cdot \mathbf{1}_{2^j b \le p_i \le 2^{j+1}b}] \\ &\le \sum_{i\in\overline{L}(p)}\sum_{j\ge 0} 2^{j+1}b \cdot \Pr[\hat{p}_i \ge 2^j b] \\ &\overset{Eq.\ (A.4)}{\le} \sum_{i\in\overline{L}(p)} \frac{C\varepsilon^2 p_i}{b}\sum_{j\ge 0}\frac{2^{j+1}b}{2^{2j}} \le 4C\varepsilon^2.\end{aligned}$$

By Markov's inequality,

$$\Pr\left[\|\hat{p}\|_{\overline{L}(p)H(\hat{p})} \ge \varepsilon\right] \le \frac{4C\varepsilon^2}{\varepsilon} = 4C\varepsilon \le \delta.$$

COROLLARY A.2. *For any $\delta > 0$, any $\varepsilon \ll \delta$, using $m \gg 1/(\varepsilon^2 b)$ samples,*

$$D(H(\hat{p})H(\hat{q})\overline{L}(p)\overline{L}(q)) \le \varepsilon,$$

except with probability δ.

Proof. By the triangle inequality,

$$\begin{aligned}\|p-q\|_{H(\hat{p})H(\hat{q})\overline{L}(p)\overline{L}(q)} \le &\|\hat{p}-p\|_{\overline{L}(p)H(\hat{p})} \\ &+\|\hat{q}-q\|_{\overline{L}(q)H(\hat{q})} + \|\hat{p}-\hat{q}\|_{H(\hat{p})H(\hat{q})\overline{L}(p)\overline{L}(q)}.\end{aligned}$$

The first two terms on the RHS are dominated by $\|\hat{p}\|_{\overline{L}(p)H(\hat{p})}$ and $\|\hat{q}\|_{\overline{L}(q)H(\hat{q})}$. By Lemma A.1,

$$\|p-q\|_{H(\hat{p})H(\hat{q})\overline{L}(p)\overline{L}(q)} \le \|\hat{p}-\hat{q}\|_{H(\hat{p})H(\hat{q})\overline{L}(p)\overline{L}(q)} + \varepsilon$$

except with probability $\delta/2$. We also get the reverse inequality by swapping the roles of $p-q$ and $\hat{p}-\hat{q}$.

COROLLARY A.3. *For any $\delta > 0$, any $\varepsilon \ll \delta$, using $m \gg 1/(\varepsilon^2 b)$ samples,*

$$D(H(\hat{p})L(\hat{q})\overline{L}(p)\overline{L}(q)) \le \varepsilon$$

except with probabilty δ.

Proof. It is easy to see that $||\hat{p}_i - \hat{q}_i| - |p_i - q_i|| \le \hat{p}_i$ for $i \in H(\hat{p})L(\hat{q})\overline{L}(p)\overline{L}(q)$. Hence

$$D(H(\hat{p})L(\hat{q})\overline{L}(p)\overline{L}(q)) \le \|\hat{p}_i\|_{\overline{L}(p)H(\hat{p})},$$

and the result follows by Lemma A.1.

Applying Corollaries A.1, A.2 and A.3 to inequalities Eq. (A.1) and Eq. (A.3), we have thus shown the main theorem of this section.

THEOREM A.1. *For any $\delta > 0$, any $\varepsilon \ll \delta$, using $m \gg 1/(\varepsilon^2 b\delta)$ samples,*

$$\|\hat{p}-\hat{q}\|_{H(\hat{p})\cup H(\hat{q})} \approx_\varepsilon \|p-q\|_{H(\hat{p})\cup H(\hat{q})}$$

except with probability δ.

A.2 Light elements. We now deal with the light elements. Let p' be the low-frequency distribution constructed in Step 2 of the ℓ_1 tester (those elements with empirical frequency at least b have their weights redistributed evenly). It will be shown to be $(O(b), O(1))$-bounded in Theorem A.2 below.

THEOREM A.2. *p' is $(2b, O(1/\delta))$-bounded except with probability δ.*

Proof. Let $H = \{i \mid p_i \ge 2b\}$ and $\hat{L} = \{i \mid \hat{p}_i < b \text{ and } \hat{q}_i < b\}$. We wish to bound

$$\mathbb{E}\left[\sum_{i\in\hat{L}\cap H} p_i^t\right] = \sum_{i\in H} p_i^t \Pr[i \in \hat{L}] \tag{A.5}$$

by $O_t(b^{t-1})$. Indeed, writing $p_i = x_i b$, the summand

$$p_i^t \Pr[i \in \hat{L}] \le p_i^t \Pr[\hat{p}_i \le b] = p_i b^{t-1} \cdot x_i^{t-1}\,\mathrm{bin}(m, p_i, < bm).$$

The factor

$$x^{t-1}\,\mathrm{bin}(m, p_i, < bm) \le x_i^{t-1} \exp\left(-\frac{Cx_i}{8\varepsilon^2}\right)$$

by a Chernoff bound and equals $O_t(1)$ uniformly in x_i and ε. Hence Eq. (A.5) is $O_t(b^{t-1})$. By Markov's inequality,

$$\sum_{i \in \hat{L} \cap H} p_i^t \ll_t b^{t-1}/\delta \tag{A.6}$$

except with probability δ.

Note that

$$p_i' \le \left(p_i + \frac{1}{n}\right)\mathbf{1}_{i \in \hat{L}} + \frac{1}{n}\mathbf{1}_{i \notin \hat{L}},$$

thus

$$\|p'\|_t^t \le \sum_{\hat{i} \in \hat{L} \cap H} \left(p_i + \frac{1}{n}\right)^t + \sum_{i \notin H} \left(p_i + \frac{1}{n}\right)^t + \sum_{i \notin \hat{L}} \left(\frac{1}{n}\right)^t.$$

Together with $(r+s)^t \ll_t r^t + s^t$ and $\sum_i (1/n)^t \le 1/n^{t-1} \le b^{t-1}$, it follows that $\|p'\|_t^t \ll_t b^{t-1}/\delta$ whenever Eq. (A.6) holds.

THEOREM A.3. *There exists an algorithm ℓ_1-Distance-Test that, for $\varepsilon \ge 1/\sqrt{n}$, uses $O(n^{2/3}\varepsilon^{-2})$ samples from p, q and has the following behavior: it rejects with probability $2/3$ when $\|p - q\|_1 \ge \varepsilon$, and accepts with probability $2/3$ when $p = q$.*

Proof. (Sketch) The algorithm proceeds as follows: We pick $b = n^{-2/3}$. We check if the "b-heavy" parts $H(\hat{p}) \cup H(\hat{q})$ of p and q are $\varepsilon/2$-far using Theorem A.1. We then construct light versions p' and q' as in [BFR+00]; these distributions are $(b, O(1))$-bounded with high probability by Theorem A.2. Finally, we check whether they are $\varepsilon/2$-far using Proposition 3.1 (where we set $\tilde{\varepsilon} = \varepsilon/\sqrt{n}$). The number of samples we need for both Theorem A.1 and Proposition 3.1 is $O(n^{2/3}\varepsilon^{-2})$. This completes the proof.