Sample-Optimal Identity Testing with High Probability

Ilias Diakonikolas[1]
USC, Los Angeles, CA, USA
diakonik@usc.edu

Themis Gouleakis[2]
CSAIL,MIT, Cambridge, MA, USA
tgoule@mit.edu

John Peebles[3]
CSAIL,MIT, Cambridge, MA, USA
jpeebles@mit.edu

Eric Price
UT Austin, Austin, TX, USA
ecprice@cs.utexas.edu

Abstract

We study the problem of testing identity against a given distribution with a focus on the high confidence regime. More precisely, given samples from an unknown distribution p over n elements, an explicitly given distribution q, and parameters $0 < \varepsilon, \delta < 1$, we wish to distinguish, *with probability at least* $1 - \delta$, whether the distributions are identical versus ε-far in total variation distance. Most prior work focused on the case that $\delta = \Omega(1)$, for which the sample complexity of identity testing is known to be $\Theta(\sqrt{n}/\varepsilon^2)$. Given such an algorithm, one can achieve arbitrarily small values of δ via black-box amplification, which multiplies the required number of samples by $\Theta(\log(1/\delta))$.

We show that black-box amplification is suboptimal for any $\delta = o(1)$, and give a new identity tester that achieves the optimal sample complexity. Our new upper and lower bounds show that the optimal sample complexity of identity testing is

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

for any n, ε, and δ. For the special case of uniformity testing, where the given distribution is the uniform distribution U_n over the domain, our new tester is surprisingly simple: to test whether $p = U_n$ versus $d_{\mathrm{TV}}(p, U_n) \geq \varepsilon$, we simply threshold $d_{\mathrm{TV}}(\widehat{p}, U_n)$, where $\widehat{p}$ is the empirical probability distribution. The fact that this simple "plug-in" estimator is sample-optimal is surprising, even in the constant δ case. Indeed, it was believed that such a tester would not attain sublinear sample complexity even for constant values of ε and δ.

An important contribution of this work lies in the analysis techniques that we introduce in this context. First, we exploit an underlying strong convexity property to bound from below the expectation gap in the completeness and soundness cases. Second, we give a new, fast method for obtaining provably correct empirical estimates of the true worst-case failure probability for a broad class of uniformity testing statistics over all possible input distributions – including all previously studied statistics for this problem. We believe that our novel analysis techniques will be useful for other distribution testing problems as well.

[1] Supported by NSF Award CCF-1652862 (CAREER) and a Sloan Research Fellowship.
[2] Supported by NSF Awards, including No. CCF-1650733, CCF-1733808, and IIS-1741137
[3] Supported by NSF Awards, including No. CCF-1650733, CCF-1733808, and IIS-1741137

45th International Colloquium on Automata, Languages, and Programming (ICALP 2018).
Editors: Ioannis Chatzigiannakis, Christos Kaklamanis, Dániel Marx, and Donald Sannella;
Article No. 41; pp. 41:1–41:14

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

2012 ACM Subject Classification Mathematics of computing → Probability and statistics

Keywords and phrases distribution testing, property testing, sample complexity

Digital Object Identifier 10.4230/LIPIcs.ICALP.2018.41

Related Version The full version of this paper can be found at `https://eccc.weizmann.ac.il/report/2017/133/`.

1 Introduction

Distribution property testing [15, 4, 5], originating in statistical hypothesis testing [18, 17], studies problems of the form: given sample access to one or more unknown distributions, determine whether they satisfy some global property or are "far" from satisfying the property. (See Section 1.1 for a formal definition.) During the past two decades problems of this form have received significant attention within the computer science community. See [20, 6] for two recent surveys.

Research in this field has primarily centered on determining tight bounds on the sample complexity of testing various properties *in the constant probability of success regime*. That is, the testing algorithm must succeed with a probability of (say) at least 2/3. This constant confidence regime is fairly well understood. For a range of fundamental properties [19, 7, 22, 12, 11, 1, 10, 9] we now have *sample-optimal* testers that use provably optimal number of samples (up to constant factors) in this regime.

In sharp contrast, the high confidence regime – i.e., the case where the desired failure probability is subconstant – is poorly understood even for the most basic properties. For essentially all distribution property testing problems studied in the literature, the standard amplification method is the only way known to achieve a high confidence success probability. Amplification is a black-box method that can boost the success probability to any desired accuracy. However, using it increases the number of required samples beyond what is necessary to obtain constant confidence. Specifically, to achieve a high confidence success probability of $1 - \delta$ via amplification, the number of samples required increases by a factor of $\Theta(\log(1/\delta))$ compared to the constant confidence regime.

This discussion raises the following natural questions: *For a given distribution property testing problem, does black-box amplification give sample-optimal testers for obtaining a high confidence success probability? Specifically, is the* $\Theta(\log(1/\delta))$ *multiplicative increase in the sample size the best possible? If not, can we design testers that have optimal sample complexity in terms of all relevant problem parameters, including the error probability* δ*?*

We believe that these are fundamental questions that merit theoretical investigation in their own right. As Goldreich notes [14], "eliminating the error probability as a parameter does not allow to ask whether or not one may improve over the straightforward error reduction". From a practical perspective, understanding this high confidence regime is important to applications of hypothesis testing (e.g., in biology), because the failure probability δ of the test can be reported as a p-value. (The family of distribution testing algorithms with success probability $1 - \delta$ for a given problem is equivalent to the family of statistical tests whose p-value (probability of Type I error) and probability of Type II error are both at most δ.) Standard techniques for addressing the problem of multiple comparisons, such as Bonferroni correction, require vanishingly small p-values.

Perhaps surprisingly, with one exception [16], this basic problem has not been previously investigated in the finite sample regime. A conceptual contribution of this work is to raise this problem as a fundamental goal in distribution property testing. We note here that

the analogous question in the context of *distribution learning* has been intensely studied in statistics and probability theory (see, e.g., [23, 8]) and tight bounds are known in a range of settings.

1.1 Formal Framework

The focus of this work is on the task of identity testing, which is arguably *the* most fundamental distribution testing problem.

▶ **Definition 1** (Distribution Identity Testing Problem). Given a target distribution q with domain D of size n, parameters $0 < \varepsilon, \delta < 1$, and sample access to an unknown distribution p over the same domain, we want to distinguish *with probability at least* $1 - \delta$ between the following cases:

- Completeness: $p = q$.
- Soundness: $d_{\mathrm{TV}}(p, q) \geq \varepsilon$.

We call this the problem of (ε, δ) *testing identity to* q. The special case of q being uniform is known as *uniformity testing*. An algorithm that solves one of these problems will be called an (ε, δ)-*tester* for identity/uniformity.

Note that $d_{\mathrm{TV}}(p, q)$ denotes the total variation distance or statistical distance between distributions p and q, i.e., $d_{\mathrm{TV}}(p, q) \stackrel{\text{def}}{=} \frac{1}{2} \cdot \|p - q\|_1$. The goal is to characterize the sample complexity of the problem: i.e., the number of samples that are necessary and sufficient to correctly distinguish between the completeness and soundness cases with probability $1 - \delta$.

1.2 Our Results

Our main result is a complete characterization of the worst-case sample complexity of identity testing in the high confidence regime. For this problem, we show that black-box amplification is suboptimal for *any* $\delta = o(1)$, and give a new identity tester that achieves the optimal sample complexity:

▶ **Theorem 2** (Main Result). *There exists a computationally efficient* (ε, δ)*-identity tester for discrete distributions of support size* n *with sample complexity*

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

Moreover, this sample size is information-theoretically optimal, up to a constant factor, for all n, ε, δ.

As we explain in Section 1.3, [16] gave a tester that achieves the optimal sample complexity when the sample size is $o(n)$. However this tester *completely* fails with $\Omega(n)$ samples, as may be required when either ε or δ are sufficiently small. Theorem 2 provides a complete characterization of the worst-case sample complexity of the problem with a *single* statistic for all settings of parameters n, ε, δ.

Brief Overview of Techniques. To analyze our tester, we introduce two new techniques for the analysis of distribution testing statistics, which we describe in more detail in Section 1.4. Our techniques leverage a simple common property of numerous distribution testing statistics which does not seem to have been previously exploited in their analysis: their convexity. Our first technique crucially exploits an underlying strong convexity property to bound from below the expectation gap between the completeness and soundness cases.

We remark that this is a contrast to most known distribution testers where bounding the expectation gap is easy, and the challenge is in bounding the variance of the statistic.

Our second technique implies a new, fast method for obtaining empirical estimates of the true worst-case failure probability of any member of a broad class of uniformity testing statistics. This class includes all uniformity testing statistics studied in the literature. Critically, these estimates come with provable guarantees about the worst-case failure probability of the statistic over all possible input distributions, and have tunable additive error. We elaborate in Section 1.4.

1.3 Discussion and Prior Work

Uniformity testing is the first and one of the most well-studied problems in distribution testing [15, 19, 22, 12, 9]. As already mentioned, the literature has almost exclusively focused on the case of constant error probability δ. The first uniformity tester, introduced by Goldreich and Ron [15], counts the number of collisions among the samples and was shown to work with $O(\sqrt{n}/\varepsilon^4)$ samples [15]. A related tester proposed by Paninski [19], which relies on the number of distinct elements in the set of samples, was shown to have the optimal $m = \Theta(\sqrt{n}/\varepsilon^2)$ sample complexity, as long as $m = o(n)$. Recently, a chi-squared based tester was shown in [22, 12] to achieve the optimal $\Theta(\sqrt{n}/\varepsilon^2)$ sample complexity without any restrictions. Finally, the original collision-based tester of [15] was very recently shown to also achieve the optimal $\Theta(\sqrt{n}/\varepsilon^2)$ sample complexity [9]. Thus, the situation for constant values of δ is well understood.

The problem of identity testing against an arbitrary (explicitly given) distribution was studied in [3], who gave an $(\varepsilon, 1/3)$-tester with sample complexity $\tilde{O}(n^{1/2})/\mathrm{poly}(\varepsilon)$. The tight bound of $\Theta(n^{1/2}/\varepsilon^2)$ was first given in [22] using a chi-squared type tester (inspired by [7]). In subsequent work, a similar chi-squared tester that also achieves the same sample complexity bound was given in [1]. (We note that the [22, 1] testers have sub-optimal sample complexity in the high confidence regime, even for the case of uniformity.) In a related work, [12] obtained a reduction of identity to uniformity that preserves the sample complexity, up to a constant factor, in the constant error probability regime. More recently, Goldreich [13], building on [10], gave a different reduction of identity to uniformity that preserves the error probability. We use the latter reduction in this paper to obtain an optimal identity tester starting from our new optimal uniformity tester.

Since the sample complexity of identity testing is $\Theta(\sqrt{n}/\varepsilon^2)$ for $\delta = 1/3$ [22, 12], standard amplification gives a sample upper bound of $\Theta(\sqrt{n}\log(1/\delta)/\varepsilon^2)$ for this problem. It is not hard to observe that this naive bound cannot be optimal for all values of δ. For example, in the extreme case that $\delta = 2^{-\Theta(n)}$, this gives a sample complexity of $\Theta(n^{3/2}/\varepsilon^2)$. On the other hand, one can *learn* the underlying distribution (and therefore test for identity) with $O(n/\varepsilon^2)$ samples for such values of δ[4].

The case where $1 \gg \delta \gg 2^{-\Theta(n)}$ is more subtle, and it is not a priori clear how to improve upon naive amplification. Theorem 2 provides a smooth transition between the extremes of $\Theta(\sqrt{n}/\varepsilon^2)$ for constant δ and $\Theta(n/\varepsilon^2)$ for $\delta = 2^{-\Theta(n)}$. It thus provides a quadratic improvement in the dependence on δ over the naive bound for all $\delta \geq 2^{-\Theta(n)}$, and shows that this is the best possible. For $\delta < 2^{-\Theta(n)}$, it turns out that the additive $\Theta(\log(1/\delta)/\varepsilon^2)$ term is necessary, as outlined in Section 1.4, so learning the distribution is optimal for such values of δ.

[4] This follows from the fact that, for any distribution p over n elements, the empirical probability distribution $\widehat{p}_m$ obtained after $m = \Omega((n + \log(1/\delta))/\varepsilon^2)$ samples drawn from p is ε-close to p in total variation distance with probability at least $1 - \delta$.

We obtain the first sample-optimal *uniformity tester* for the high confidence regime. Our sample-optimal identity tester follows from our uniformity tester by applying the recent result of Goldreich [13], which provides a black-box reduction of identity to uniformity. We also show a matching information-theoretic lower bound on the sample complexity.

The sample-optimal uniformity tester we introduce is remarkably simple: to distinguish between the cases that p is the uniform distribution U_n over n elements versus $d_{\mathrm{TV}}(p, U_n) \geq \varepsilon$, we simply compute $d_{\mathrm{TV}}(\widehat{p}, U_n)$ for the empirical distribution $\widehat{p}$. The tester accepts that $p = U_n$ if the value of this statistic is below some well-chosen threshold, and rejects otherwise.

It should be noted that such a tester was not previously known to work with sub-learning sample complexity – i.e., fewer than $\Theta(n/\varepsilon^2)$ samples – even in the constant confidence regime. Surprisingly, in a literature with several different uniformity testers [15, 19, 22, 12], no one has previously used the empirical total variation distance. On the contrary, it would be natural to assume – as was suggested in [4, 5] – that this tester cannot possibly work. A likely reason for this is the following observation: When the sample size m is smaller than the domain size n, the empirical total variation distance is very far from the true distance to uniformity. This suggests that the empirical distance statistic gives little, if any, information in this setting.

Despite the above intuition, we prove that the natural "plug-in" estimator relying on the empirical distance from uniformity actually works for the following reason: the empirical distance from uniformity is *noticeably smaller* for the uniform distribution than for "far from uniform" distributions, *even with a sub-linear sample size*. Moreover, we obtain the stronger statement that the "plug-in" estimator is a sample-optimal uniformity tester for all parameters n, ε and δ.

In [16], it was shown that the distinct-elements tester of [19] achieves the optimal sample complexity of $m = \Theta(\sqrt{n \log(1/\delta)}/\varepsilon^2)$, as long as $m = o(n)$. When $m = \Omega(n)$, as is the case in many practically relevant settings (see, e.g., the Polish lottery example in [21] with $n < \sqrt{n}/\varepsilon^2 \ll n/\varepsilon^2$), this tester is known to fail completely even in the constant confidence regime. On the other hand, in such settings the sample size is *not* sufficiently large so that we can actually *learn* the underlying distribution.

It is important to note that *all* previously considered uniformity testers [15, 19, 22, 12] do not achieve the optimal sample complexity (as a function of all parameters, including δ), and this is *inherent*, i.e., not just a failure of previous analyses. Roughly speaking, since the collision statistic [15] and the chi-squared based statistic [22, 12] are not Lipschitz, it can be shown that their high-probability performance is poor. Specifically, in the completeness case ($p = U_n$), if many samples happen to land in the same bucket (domain element), these test statistics become quite large, leading to their suboptimal behavior *for all* $\delta = o(1)$. (For a formal justification, the reader is referred to Section V of [16]). On the other hand, the distinct-elements tester [19] does not work for $m = \omega(n)$. For example, if ε or δ are sufficiently small to necessitate $m \gg n \log n$, then typically all n domain elements will appear in both the completeness and soundness cases, hence the test statistic provides no information.

1.4 Our Techniques

1.4.1 Upper Bound for Uniformity Testing

We would like to show that the test statistic $d_{\mathrm{TV}}(\widehat{p}, U_n)$ is with high probability larger when $d_{\mathrm{TV}}(p, U_n) \geq \varepsilon$ than when $p = U_n$. We start by showing that among all possible alternative distributions p with $d_{\mathrm{TV}}(p, U_n) \geq \varepsilon$, it suffices to consider those in a very simple family. We then show that the test statistic is highly concentrated around its expectation, and that the expectations are significantly different in the two cases. The main technical components of our paper are our techniques for accomplishing these tasks.

To simplify the structure of p, it can be shown that if p majorizes another distribution q, then the test statistic $d_{\mathrm{TV}}(\widehat{p}, U_n)$ stochastically dominates $d_{\mathrm{TV}}(\widehat{q}, U_n)$. (In fact, this statement holds for *any* test statistic that is a convex symmetric function of the empirical histogram.) We defer this proof to the full version. Therefore, for any p, if we average out the large and small entries of p, the test statistic becomes harder to distinguish from uniform.

We remark as a matter of independent interest that this stochastic domination lemma immediately implies a fast algorithm for performing rigorous empirical comparisons of test statistics. A major difficulty in empirical studies of distribution testing is that it is not possible to directly check the failure probability of a tester over every possible distribution as input, because the space of such distributions is quite large. Our structural lemma reduces the search space dramatically for uniformity testing: for any convex symmetric test statistic (which includes all existing ones), the worst case distribution will have αn coordinates of value $(1+\varepsilon/\alpha)/n$, and the rest of value $(1-\varepsilon/(1-\alpha))/n$, for some α. Hence, there are only n possible worst-case distributions for any ε. Notably, this reduction does not lose anything, so it could be used to identify the non-asymptotic optimal constants that a distribution testing statistic achieves for a given set of parameters.

Returning to our uniformity tester, at the cost of a constant factor in ε we can assume $\alpha = 1/2$. As a result, we only need to consider p to be either U_n or of the form $\frac{1\pm\varepsilon}{n}$ in each coordinate. We now need to separate the expectation of the test statistic in these two situations. The challenge is that both expectations are large, and we do not have a good analytic handle on them. We therefore introduce a new technique for showing a separation between the completeness and soundness cases that utilizes the strong convexity of the test statistic. Specifically, we obtain an explicit expression for the Hessian of the expectation, as a function of p. The Hessian is diagonal, and for our two situations of $p_i \approx 1/n$ each entry is within constant factors of the same value, giving a lower bound on its eigenvalues. Since the expectation is minimized at $p = U_n$, strong convexity implies an expectation gap. Specifically, we prove that this gap is $\varepsilon^2 \cdot \min(m^2/n^2, \sqrt{m/n}, 1/\varepsilon)$.

Finally, we need to show that the test statistic concentrates about its expectation. For $m \geq n$, this follows from McDiarmid's inequality: since the test statistic is $1/m$-Lipschitz in the m samples, with probability $1-\delta$ it lies within $\sqrt{\log(1/\delta)/m}$ of its expectation. When m is larger than the desired sample complexity given in (1), this is less than the expectation gap above. The concentration is trickier when $m < n$, since the expectation gap is smaller, so we need to establish tighter concentration. We get this by using a Bernstein variant of McDiarmid's inequality, which is stronger than the standard version of McDiarmid in this context. We note that the use of the stochastic domination is also crucial here. Since our statistic is a symmetric convex function of the histogram values, we show (details deferred to the full version) that for each soundness case distribution, there is a different distribution that has possible probability mass values exclusively in the set $\{\frac{1+\varepsilon'}{n}, \frac{1}{n}, \frac{1-\varepsilon'}{n}\}$, for some $\varepsilon' = O(\varepsilon)$, and is harder to distinguish from the uniform distribution. However, we could show that this distribution has a stronger Lipschitz-type property than the other soundness case distributions. Therefore, we are able to use a stronger concentration bound via McDiarmid's inequality and argue that even though other soundness case distributions may have weaker concentration, they still have smaller error due to our stochastic domination argument.

1.4.2 Upper Bound for Identity Testing

In [13], it was shown how to reduce ε-testing of an arbitrary distribution q over $[n]$ to $\varepsilon/3$-testing of U_{6n}. This reduction preserves the error probability δ, so applying it gives an identity tester with the same sample complexity as our uniformity tester, up to constant

factors.

1.4.3 Sample Complexity Lower Bound

To match our upper bound (1), we need two lower bounds. The lower bound of $\Omega(\frac{1}{\varepsilon^2}\log(1/\delta))$ is straightforward from the same lower bound as for distinguishing a fair coin from an ε-biased coin, while the $\sqrt{n\log(1/\delta)}/\varepsilon^2$ bound is more challenging.

For intuition, we start with a $\sqrt{n\log(1/\delta)}$ lower bound for constant ε. When $p = U_n$, the chance that all m samples are distinct is at least $(1-m/n)^m \approx e^{-m^2/n}$. Hence, if $m \ll \sqrt{n\log(1/\delta)}$, this would happen with probability significantly larger than 2δ. On the other hand, if p is uniform over a random subset of $n/2$ coordinates, the m samples will also all be distinct with probability $(1-2m/n)^m > 2\delta$. The two situations thus look the same with 2δ probability, so no tester could have accuracy $1-\delta$.

This intuition can easily be extended to include a $1/\varepsilon$ dependence, but getting the desired $1/\varepsilon^2$ dependence requires more work. First, we Poissonize the number of samples, so we independently see $\mathrm{Poi}(mp_i)$ samples of each coordinate i; with exponentially high probability, this Poissonization only affects the sample complexity by constant factors. Then, in the alternative hypothesis, we set each p_i independently at random to be $\frac{1\pm\varepsilon}{n}$. This has the unfortunate property that p no longer sums to 1, so it is a "pseudo-distribution" rather than an actual distribution. Still, it is exponentially likely to sum to $\Theta(1)$, and using techniques from [24, 10] this is sufficient for our purposes.

At this point, we are considering a situation where the number of times we see each coordinate is either $\mathrm{Poi}(m/n)$ or $\frac{1}{2}(\mathrm{Poi}((1-\varepsilon)\frac{m}{n}) + \mathrm{Poi}((1+\varepsilon)\frac{m}{n}))$, and every coordinate is independent of the others. These two distributions have Hellinger distance at least $\varepsilon^2 m/n$ in each coordinate. Then the composition property for Hellinger distance over n independent coordinates implies $m \geq \sqrt{n\log(1/\delta)}/\varepsilon^2$ is necessary for success probability $1-\delta$.

1.5 Notation

We write $[n]$ to denote the set $\{1,\ldots,n\}$. We consider discrete distributions over $[n]$, which are functions $p : [n] \to [0,1]$ such that $\sum_{i=1}^n p_i = 1$. We use the notation p_i to denote the probability of element i in distribution p. For $S \subseteq [n]$, we will denote $p(S) = \sum_{i\in S} p_i$. We will also sometimes think of p as an n-dimensional vector. We will denote by U_n the uniform distribution over $[n]$.

For $r \geq 1$, the ℓ_r-norm of a distribution is identified with the ℓ_r-norm of the corresponding vector, i.e., $\|p\|_r = (\sum_{i=1}^n |p_i|^r)^{1/r}$. The ℓ_r-distance between distributions p and q is defined as the the ℓ_r-norm of the vector of their difference. The total variation distance between distributions p and q is defined as $d_{\mathrm{TV}}(p,q) \stackrel{\mathrm{def}}{=} \max_{S\subseteq[n]} |p(S)-q(S)| = (1/2)\cdot\|p-q\|_1$. The Hellinger distance between p and q is $H(p,q) \stackrel{\mathrm{def}}{=} (1/\sqrt{2})\cdot\|\sqrt{p}-\sqrt{q}\|_2 = (1/\sqrt{2})\cdot\sqrt{\sum_{i=1}^n(\sqrt{p_i}-\sqrt{q_i})^2}$. We denote by $\mathrm{Poi}(\lambda)$ the Poisson distribution with parameter λ.

2 Sample-Optimal Uniformity Testing

In this section, we describe and analyze our optimal uniformity tester. Given samples from an unknown distribution p over $[n]$, our tester returns "YES" with probability $1-\delta$ if $p = U_n$, and "NO" with probability $1-\delta$ if $d_{\mathrm{TV}}(p,U_n) \geq \varepsilon$.

2.1 Our Test Statistic

We define a very natural statistic that yields a uniformity tester with optimal dependence on the domain size n, the proximity parameter ε, and the error probability δ. Our statistic is a thresholded version of the empirical total variation distance between the unknown distribution p and the uniform distribution. Our tester Test-Uniformity is described in the following pseudocode:

Algorithm Test-Uniformity$(p, n, \varepsilon, \delta)$
Input: sample access to a distribution p over $[n]$, $\varepsilon > 0$, and $\delta > 0$.
Output: "YES" if $p = U_n$; "NO" if $d_{\mathrm{TV}}(p, U_n) \geq \varepsilon$.

1. Draw $m = \Theta\left((1/\varepsilon^2) \cdot \left(\sqrt{n \log(1/\delta)} + \log(1/\delta)\right)\right)$ i.i.d. samples from p.
2. Let $X = (X_1, X_2, \ldots, X_n) \in \mathbb{Z}^n_{>0}$ be the histogram of the samples. That is, X_i is the number of times domain element i appears in the (multi-)set of samples.
3. Define the random variable $S = \frac{1}{2} \sum_{i=1}^{n} \left| \frac{X_i}{m} - \frac{1}{n} \right|$ and set the threshold

$$t = \mu(U_n) + C \cdot \begin{cases} \varepsilon^2 \cdot \frac{m^2}{n^2} & \text{for } m \leq n \\ \varepsilon^2 \cdot \sqrt{\frac{m}{n}} & \text{for } n < m \leq \frac{n}{\varepsilon^2} \\ \varepsilon & \text{for } \frac{n}{\varepsilon^2} \leq m \end{cases},$$

   where C is a universal constant (derived from the analysis of the algorithm), and $\mu(U_n)$ is the expected value of the statistic in the completeness case. (We can compute $\mu(U_n)$ in $O(m)$ time using the procedure in Appendix A of the full version.)
4. If $S \geq t$ return "NO"; otherwise, return "YES".

The main part of this section is devoted to the analysis of Test-Uniformity, establishing the following theorem:

▶ **Theorem 3.** *There exists a universal constant $C > 0$ such that the following holds: Given*

$$m \geq C \cdot (1/\varepsilon^2) \left(\sqrt{n \log(1/\delta)} + \log(1/\delta)\right)$$

samples from an unknown distribution p, Algorithm Test-Uniformity *is an (ε, δ)-tester for uniformity of distribution p.*

As we point out in Appendix A of the full version, the value $\mu(U_n)$ can be computed efficiently, hence our overall tester is computationally efficient. To prove correctness of the above tester, we need to show that the expected value of the statistic in the completeness case is sufficiently separated from the expected value in the soundness case, and also that the value of the statistic is highly concentrated around its expectation in both cases. In Section 2.2, we bound from below the difference in the expectation of our statistic in the completeness and soundness cases. The proof the desired concentration, which completes the proof of Theorem 3, is deferred to the full version.

2.2 Bounding the Expectation Gap

The expectation of the statistic in algorithm Test-Uniformity can be viewed as a function of the n variables $p_1, \ldots, p_n$. We denote this expectation by $\mu(p) \stackrel{\text{def}}{=} \mathbb{E}[S(X_1, \ldots, X_n)]$ when the samples are drawn from distribution p.

Our analysis has a number of complications for the following reason: the function $\mu(p) - \mu(U_n)$ is a linear combination of sums that have no indefinite closed form, even if the distribution p assigns only two possible probabilities to the elements of the domain. This statement is made precise in Appendix B of the full version. As such, we should only hope to obtain an approximation of this quantity.

A natural approach to try and obtain such an approximation would be to produce separate closed form approximations for $\mu(p)$ and $\mu(U_n)$, and combine these quantities to obtain an approximation for their difference. However, one should not expect such an approach to work in our context. The reason is that the difference $\mu(p) - \mu(U_n)$ can be much smaller than $\mu(p)$ and $\mu(U_n)$; it can even be arbitrarily small. As such, obtaining separate approximations of $\mu(p)$ and $\mu(U_n)$ to any fixed accuracy would contribute too much error to their difference.

To overcome these difficulties, we introduce the following technique, which is novel in this context. We directly bound from below the difference $\mu(p) - \mu(U_n)$ using *strong convexity*. Specifically, we show that the function μ is strongly convex with appropriate parameters and use this fact to bound the desired expectation gap. The main result of this section is the following lemma:

▶ **Lemma 4.** *Let p be a distribution over $[n]$ and $\varepsilon = d_{\mathrm{TV}}(p, U_n)$. For all $m \geq 6$ and $n \geq 2$, we have that:*

$$\mu(p) - \mu(U_n) \geq \Theta(1) \cdot \begin{cases} \varepsilon^2 \cdot \frac{m^2}{n^2} & \text{for } m \leq n \\ \varepsilon^2 \cdot \sqrt{\frac{m}{n}} & \text{for } n < m \leq \frac{n}{\varepsilon^2} \\ \varepsilon & \text{for } \frac{n}{\varepsilon^2} \leq m \end{cases} .$$

We note that the bounds in the right hand side above are tight, up to constant factors. Any asymptotic improvement would yield a uniformity tester with sample complexity that violates our tight information-theoretic lower bounds.

The proof of Lemma 4 (which will be deferred to the full version) requires a couple of important intermediate lemmas. Our starting point is as follows: By the intermediate value theorem, we have the quadratic expansion

$$\mu(p) = \mu(U_n) + \nabla\mu(U_n)^\intercal (p - U_n) + \frac{1}{2}(p - U_n)^\intercal H_{p'} (p - U_n) \,,$$

where $H_{p'}$ is the Hessian matrix of the function μ at some point p' which lies on the line segment between U_n and p. This expression can be simplified as follows: First, we show that our μ is minimized over all probability distributions on input U_n (see the full version of the paper for details). Thus, the gradient $\nabla\mu(U_n)$ must be orthogonal to being a direction in the space of probability distributions. In other words, $\nabla\mu(U_n)$ must be proportional to the all-ones vector. More formally, since μ is symmetric its gradient is a symmetric function, which implies it will be symmetric when given symmetric input. Moreover, $(p - U_n)$ is a direction within the space of probability distributions, and therefore sums to 0, making it orthogonal to the all-ones vector. Thus, we have that $\nabla\mu(U_n)^\intercal (p - U_n) = 0$, and we obtain

$$\mu(p) - \mu(U_n) = \frac{1}{2}(p - U_n)^\intercal H_{p'} (p - U_n) \geq \frac{1}{2}\|p - U_n\|_2^2 \cdot \sigma \geq \frac{1}{2}\|p - U_n\|_1^2 / n \cdot \sigma \,, \tag{2}$$

where σ is the minimum eigenvalue of the Hessian of μ on the line segment between U_n and p.

The majority of this section is devoted to proving a lower bound for σ. Before doing so, however, we must first address a technical consideration. Because we are considering a function over the space of probability distributions – which is not full-dimensional – the

Hessian and gradient of μ with respect to $\mathbb{R}^n$ depend not only on the definition of our statistic S, but also its parameterization. For the purposes of this subsection, we parameterize S as $S(x) = \sum_{i=1}^{n} \max\left\{\frac{x_i}{m} - \frac{1}{n}, 0\right\} = \frac{1}{m}\sum_{i=1}^{n} \max\left\{x_i - \frac{m}{n}, 0\right\}$.

In the analysis we are about to perform, it will be helpful to replace $\frac{m}{n}$ with a free parameter t which we will eventually set back to roughly m/n. Thus, we define

$$S_t(x) \triangleq \frac{1}{m}\sum_{i=1}^{n} \max\{x_i - t, 0\}$$

and

$$\mu_t(p) \triangleq \mathbb{E}_{x\sim\text{Multinomial}(m,p)}[S_t(x)] = \frac{1}{m}\sum_{i=1}^{n}\sum_{k=\lceil t\rceil}^{m}\binom{m}{k}p_i^k(1-p_i)^{m-k}(k-t)\ . \tag{3}$$

Note that when $t = m/n$ we have $S_t = S$ and $\mu_t = \mu$. Also note that when we compute the Hessian of $\mu_t(p)$, we are treating $\mu_t(p)$ as a function of p and not of t. In the following lemma, we derive an exact expression for the entries of the Hessian. This result is perhaps surprising in light of the likely nonexistence of a closed form expression for $\mu(p)$. That is, while the expectation $\mu(p)$ may have no closed form, we prove that the Hessian of $\mu(p)$ does in fact have a closed form.

▶ **Lemma 5.** *The Hessian of $\mu_t(p)$ viewed as a function of p is a diagonal matrix whose ith diagonal entry is given by*

$$h_{ii} = s_{t,i}\ ,$$

where we define $s_{t,i}$ as follows: Let Δt be the distance of t from the next largest integer, i.e., $\Delta t \triangleq \lceil t\rceil - t$. Then, we have that

$$s_{t,i} = \begin{cases} 0 & \text{for } t = 0 \\ (m-1)\binom{m-2}{t-1}p_i^{t-1}(1-p_i)^{m-t-1} & \text{for } t \in \mathbb{Z}_{>0} \\ \Delta t\cdot s_{\lfloor t\rfloor,i} + (1-\Delta t)\cdot s_{\lceil t\rceil,i} & \text{for } t \geq 0 \text{ and } t \notin \mathbb{Z} \end{cases}.$$

In other words, we will derive the formula for integral $t \geq 1$ and then prove that the value for nonintegral $t \geq 0$ can be found by linearly interpolating between the closest integral values of t.

Proof. Note that because $S_t(x)$ is a separable function of x, $\mu_t(p)$ is a separable function of p, and hence the Hessian of $\mu_t(p)$ is a diagonal matrix. By Equation 3, the i-th diagonal entry of this Hessian can be written explicitly as the following expression:

$$s_{t,i} = \frac{\partial^2}{\partial p_i^2}\mu_t(p) = \frac{d^2}{dp_i^2}\frac{1}{m}\sum_{k=\lceil t\rceil}^{m}\binom{m}{k}p_i^k(1-p_i)^{m-k}(k-t)\ .$$

Notice that if we sum starting from $k = 0$ instead of $k = \lceil t\rceil$, then the sum equals the expectation of $\text{Bin}(m, p_i)$ minus t. That is, notice that:

$$\frac{\mathrm{d}^2}{\mathrm{d}p_i^2}\frac{1}{m}\sum_{k=0}^{m}\binom{m}{k}p_i^k(1-p_i)^{m-k}(k-t) = \frac{\mathrm{d}^2}{\mathrm{d}p_i^2}\frac{1}{m}(p_i m - t) = 0\ .$$

By this observation and the fact that the summand is 0 for integer t when $k = t$, we can switch which values of k we are summing over to k from 0 through $\lfloor t \rfloor$ if we negate the expression:

$$s_{t,i} = \frac{\partial^2}{\partial^2 p_i}\mu_t(p) = \frac{1}{m}\frac{\mathrm{d}^2}{\mathrm{d}p_i^2}\sum_{k=0}^{\lfloor t \rfloor}\binom{m}{k}p_i^k(1-p_i)^{m-k}(t-k) \; .$$

We first prove the case when $t \in \mathbb{Z}_+$. In this case, we view $s_{t,i}$ as a sequence with respect to t (where i is fixed), which we denote s_t. We now derive a generating function for this sequence.[5] Observe that derivatives that are not with respect to the formal variable commute with taking generating functions. Then, the generating function for the sequence $\{s_t\}$ is

$$\frac{\mathrm{d}^2}{\mathrm{d}p_i^2}\frac{1}{m}\left(x\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{(p_i x + 1 - p_i)^m}{1-x}\right) - \frac{x\frac{\mathrm{d}}{\mathrm{d}x}(p_i x + 1 - p_i)^m}{1-x}\right) = (m-1)(p_i x + 1 - p_i)^{m-2}x \; .$$

Note that the coefficient on x^0 is 0, so $s_{0,i} = 0$ as claimed. For $t \in \mathbb{Z}_{>0}$, the right hand side is the generating function of

$$(m-1)\binom{m-2}{t-1}p^{t-1}(1-p)^{m-t-1} \; .$$

Thus, this expression gives the i-th entry Hessian in the $t \in \mathbb{Z}_{\geq 0}$, as claimed.

Now consider the case when t is not an integer. In this case, we have:

$$\begin{aligned}
s_{t,i} &\triangleq \frac{\mathrm{d}^2}{\mathrm{d}p_i^2}\frac{1}{m}\sum_{k=\lceil t \rceil}^{m}\binom{m}{k}p_i^k(1-p_i)^{m-k}(k-t) \\
&= \frac{\mathrm{d}^2}{\mathrm{d}p_i^2}\frac{1}{m}\sum_{k=\lceil t \rceil}^{m}\binom{m}{k}p_i^k(1-p_i)^{m-k}(k-\lceil t \rceil + \Delta t) \\
&= s_{\lceil t \rceil,i} + \Delta t\frac{\mathrm{d}^2}{\mathrm{d}p_i^2}\frac{1}{m}\sum_{k=\lceil t \rceil}^{m}\binom{m}{k}p_i^k(1-p_i)^{m-k}. \\
&= s_{\lceil t \rceil,i} - \Delta t\frac{\mathrm{d}^2}{\mathrm{d}p_i^2}\frac{1}{m}\sum_{k=0}^{\lceil t \rceil - 1}\binom{m}{k}p_i^k(1-p_i)^{m-k} \; .
\end{aligned}$$

The last equality is because if we change bounds on the sum so they are from 0 through m, we get 1 which has partial derivative 0. Thus, we can flip which terms we are summing over if we negate the expression.

Note that this expression we are subtracting above can be alternatively written as:

$$\Delta t\frac{\mathrm{d}^2}{\mathrm{d}p_i^2}\frac{1}{m}\sum_{k=0}^{\lceil t \rceil - 1}\binom{m}{k}p_i^k(1-p_i)^{m-k} = \Delta t \cdot (s_{\lceil t \rceil,i} - s_{\lfloor t \rfloor,i}) \; .$$

Thus, we have

$$s_{t,i} = s_{\lceil t \rceil,i} - \Delta t \cdot (s_{\lceil t \rceil,i} - s_{\lfloor t \rfloor,i}) = \Delta t \cdot s_{\lfloor t \rfloor,i} + (1 - \Delta t) \cdot s_{\lceil t \rceil,i} \; ,$$

as desired. This completes the proof of Lemma 5. ◀

[5] To avoid potential convergence issues, we view generating functions as formal polynomials from the ring of infinite formal polynomials. Under this formalism, there is no need to deal with convergence at all.

It will be convenient to simplify the exact expressions of Lemma 5 into something more manageable. This is done in the following lemma:

▶ **Lemma 6.** *Fix any constant $c > 0$. The Hessian of $\mu(p)$, viewed as a function of p, is a diagonal matrix whose i-th diagonal entry is given by*

$$h_{ii} = s_{t:=m/n,i} \geq \Theta(1) \cdot \begin{cases} \frac{m^2}{n} & \text{for } m \leq n \\ \sqrt{mn} & \text{for } n < m \leq c \cdot \frac{n}{\varepsilon^2} \end{cases},$$

assuming $p_i = \frac{1 \pm \varepsilon}{n}$, $m \geq 6$, $n \geq 2$, and $\varepsilon \leq 1/2$.

Similarly, these bounds are tight up to constant factors, as further improvements would violate our sample complexity lower bounds.

Proof. By Lemma 5, we have an exact expression $s_{t,i}$ for the ith entry of the Hessian of $\mu_t(p)$.

First, consider the case where $m \leq n$. Then we have

$$s_{t,i} = (1 - \Delta t) \cdot s_{\lceil t \rceil, i} \, .$$

Substituting $t = m/n$, $\lceil t \rceil = 1$, and $\Delta t = \lceil t \rceil - t = 1 - m/n$ gives

$$s_{t,i} = \frac{m}{n} \cdot (m-1)(1-p_i)^{m-2} = \Theta(1) \cdot \frac{m^2}{n} \, .$$

Now consider the case where $n < m \leq \Theta(1) \cdot \frac{n}{\varepsilon^2}$. Note that the case where $n < m < 2n$ follows from (i) the fact that $s_{t,i}$ for fractional t linearly interpolates between the value of $s_{t',i}$ the nearest two integral values of t' and (ii) the analyses of the cases where $m \leq n$ and $2n \leq m \leq \Theta(1)\frac{n}{\varepsilon^2}$. Thus, all we have left to do is prove the case where $2n \leq m \leq \Theta(1) \cdot \frac{n}{\varepsilon^2}$.

Since $s_{t,i}$ is a convex combination of $s_{\lceil t \rceil, i}$ and $s_{\lfloor t \rfloor, i}$, it suffices to bound from below these quantities for $t = m/n$. Both of these tasks can be accomplished simultaneously by bounding from below the quantity $s_{t=m/n+\gamma,i}$ for arbitrary $\gamma \in [-1, 1]$.

We do this as follows: Let $t = m/n + \gamma$. Using Stirling's approximation, we can show that for any $\gamma \in [-1, 1]$, we get:

$$s_{t,i} \geq \Theta(1) \cdot \sqrt{mn}$$

The calculations are deferred to the full version. ◀

3 Conclusions and Future Work

In this paper, we gave the first uniformity tester that is sample-optimal, up to constant factors, as a function of the confidence parameter. Our tester is remarkably simple and our novel analysis may be useful in other related settings. By using a known reduction of identity to uniformity, we also obtain the first sample-optimal identity tester in the same setting.

Our result is a step towards understanding the behavior of distribution testing problems in the high-confidence setting. We view this direction as one of fundamental theoretical and important practical interest. A number of interesting open problems remain. Perhaps the most appealing one is to design a *general technique* (see, e.g., [10]) that yields sample-optimal testers in the high confidence regime for a wide range of properties. From the practical standpoint, it would be interesting to perform a detailed experimental evaluation of the various algorithms (see, e.g., [16, 2]).

References

1 J. Acharya, C. Daskalakis, and G. Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3591–3599, 2015.

2 S. Balakrishnan and L. A. Wasserman. Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *CoRR*, abs/1706.10003, 2017. URL: `http://arxiv.org/abs/1706.10003`.

3 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.

4 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. URL: `citeseer.ist.psu.edu/batu00testing.html`.

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

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

7 S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014.

8 L. Devroye and G. Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics, Springer, 2001.

9 I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. Collision-based testers are optimal for uniformity and closeness. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:178, 2016.

10 I. Diakonikolas and D. M. Kane. A new approach for testing properties of discrete distributions. In *FOCS*, pages 685–694, 2016. Full version available at abs/1601.05557.

11 I. Diakonikolas, D. M. Kane, and V. Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015*, pages 1183–1202, 2015.

12 I. Diakonikolas, D. M. Kane, and V. Nikishkin. Testing identity of structured distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, pages 1841–1854, 2015.

13 O. Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. *ECCC*, 23, 2016.

14 O. Goldreich. Commentary on two works related to testing uniformity of distributions, 2017. URL: `http://www.wisdom.weizmann.ac.il/~oded/MC/229.html`.

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

16 D. Huang and S. Meyn. Generalized error exponents for small sample universal hypothesis testing. *IEEE Trans. Inf. Theor.*, 59(12):8157–8181, dec 2013.

17 E. L. Lehmann and J. P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, 2005.

18 J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933. `doi:10.1098/rsta.1933.0009`.

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

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

21 R. Rubinfeld. Taming probability distributions over big domains. *Talk given at STOC'14 Workshop on Efficient Distribution Estimation*, 2014. Available at http://www.iliasdiakonikolas.org/stoc14-workshop/rubinfeld.pdf.

22 G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *FOCS*, 2014.

23 A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

24 Y. Wu and P. Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, June 2016.