Electronic Colloquium on Computational Complexity, Report No. 132 (2017)

Sharp Bounds for Generalized Uniformity Testing

Ilias Diakonikolas[*]
University of Southern California
diakonik@usc.edu

Daniel M. Kane[†]
University of California, San Diego
dakane@cs.ucsd.edu

Alistair Stewart
University of Southern California
alistais@usc.edu

September 6, 2017

Abstract

We study the problem of *generalized uniformity testing* [BC17] of a discrete probability distribution: Given samples from a probability distribution p over an *unknown* discrete domain $\mathbf{\Omega}$, we want to distinguish, with probability at least 2/3, between the case that p is uniform on some *subset* of $\mathbf{\Omega}$ versus ϵ-far, in total variation distance, from any such uniform distribution.

We establish tight bounds on the sample complexity of generalized uniformity testing. In more detail, we present a computationally efficient tester whose sample complexity is optimal, up to constant factors, and a matching information-theoretic lower bound. Specifically, we show that the sample complexity of generalized uniformity testing is $\Theta\left(1/(\epsilon^{4/3}\|p\|_3) + 1/(\epsilon^2\|p\|_2)\right)$.

1 Introduction

Consider the following statistical task: Given independent samples from a distribution over an *unknown* discrete domain $\mathbf{\Omega}$, determine whether it is uniform on some *subset* of the domain versus significantly different from any such uniform distribution. Formally, let $\mathcal{C}_U \stackrel{\text{def}}{=} \{\mathbf{u}_S : S \subseteq \mathbf{\Omega}\}$ denote the set of uniform distributions $\mathbf{u}_S$ over subsets S of $\mathbf{\Omega}$. Given sample access to an unknown distribution p on $\mathbf{\Omega}$ and a proximity parameter $\epsilon > 0$, we want to correctly distinguish between the case that $p \in \mathcal{C}_U$ versus $d_{\mathrm{TV}}(p, \mathcal{C}_U) \stackrel{\text{def}}{=} \min_{S \subseteq \mathbf{\Omega}} d_{\mathrm{TV}}(p, \mathbf{u}_S) \geq \epsilon$, with probability at least 2/3. Here, $d_{\mathrm{TV}}(p, q) = (1/2)\|p - q\|_1$ denotes the total variation distance between distributions p and q. This natural problem, termed *generalized uniformity testing*, was recently studied by Batu and Canonne [BC17], who gave the first upper and lower bounds on its sample complexity.

Generalized uniformity testing bears a strong resemblance to the familiar task of *uniformity testing*, where one is given samples from a distribution p on an *explicitly known* domain of size n and the goal is to determine, with probability at least 2/3, whether p is the uniform distribution $\mathbf{u}_n$ on this domain versus $d_{\mathrm{TV}}(p, \mathbf{u}_n) \geq \epsilon$. Uniformity testing is arguably *the* most extensively studied problem in distribution property testing [GR00, Pan08, VV14, DKN15b, Gol16, DGPP16, DGPP17] and its sample complexity is well understood. Specifically, it is known [Pan08, CDVV14, VV14, DKN15b] that $\Theta(n^{1/2}/\epsilon^2)$ samples are necessary and sufficient for this task.

[*]Supported by NSF Award CCF-1652862 (CAREER) and a Sloan Research Fellowship.

[†]Supported by NSF Award CCF-1553288 (CAREER) and a Sloan Research Fellowship.

ISSN 1433-8092

The field of *distribution property testing* [BFR$^+$00] has seen substantial progress in the past decade, see [Rub12, Can15] for two recent surveys. A large body of the literature has focused on characterizing the sample size needed to test properties of arbitrary distributions of a *given* support size. This regime is fairly well understood: for many properties of interest there exist sample-efficient testers [Pan08, CDVV14, VV14, DKN15b, ADK15, CDGR16, DK16, DGPP16, CDS17, DGPP17]. Moreover, an emerging body of work has focused on leveraging a priori structure of the underlying distributions to obtain significantly improved samples complexities [BKR04, DDS$^+$13, DKN15b, DKN15a, CDKS17, DP17, DDK16, DKN17].

Perhaps surprisingly, the natural setting where the distribution is arbitrary on a discrete but unknown domain (of unknown size) does not seem to have been explicitly studied before the recent work of Batu and Canonne [BC17]. Returning to the specific problem studied here, at first sight it might seem that generalized uniformity testing and uniformity testing are essentially the same task. However, as shown in [BC17], the sample complexities of these two problems are significantly different. Specifically, [BC17] gave a generalized uniformity tester with expected sample complexity $O(1/(\epsilon^6\|p\|_3))$ and showed a lower bound of $\Omega(\|p\|_3)$. Since generalized uniformity is a symmetric property, any tester should essentially rely on the empirical moments (collision statistics) of the distribution [RRSS09, Val11]. The algorithm in [BC17] uses sufficiently accurate approximations of the second and third moments of the unknown distribution. Their lower bound formalizes the intuition that an approximation of the third norm is in some sense necessary to solve this problem.

1.1 Our Results and Techniques

An immediate open question arising from the work of [BC17] is to precisely characterize the sample complexity of generalized uniformity testing, as a function of all relevant parameters. The main result of this paper provides an answer to this question. In particular, we show the following:

Theorem 1.1 (Main Result)**.** *There is an algorithm with the following performance guarantee: Given sample access to an arbitrary distribution p over an unknown discrete domain $\mathbf{\Omega}$ and a parameter $0 < \epsilon < 1$, the algorithm uses $O\left(1/(\epsilon^{4/3}\|p\|_3) + 1/(\epsilon^2\|p\|_2)\right)$ independent samples from p in expectation, and distinguishes between the case $p \in \mathcal{C}_U$ versus $d_{\mathrm{TV}}(p, \mathcal{C}_U) \geq \epsilon$ with probability at least $2/3$. Moreover, for every $0 < \epsilon < 1/0$ and $n > 1$, any algorithm that distinguishes between $p \in \mathcal{C}_U$ and $d_{\mathrm{TV}}(p, \mathcal{C}_U) \geq \epsilon$ requires at least $\Omega(n^{2/3}/\epsilon^{4/3} + n^{1/2}/\epsilon^2)$ samples, where p is guaranteed to have $\|p\|_3 = \Theta(n^{-2/3})$ and $\|p\|_2 = \Theta(n^{-1/2})$.*

In the following paragraphs, we provide an intuitive explanation of our algorithm and our matching sample size lower bound, in tandem with a comparison to the prior work [BC17].

Sample-Optimal Generalized Uniformity Tester. Our algorithm requires considering two cases based on the relative size of ϵ and $\|p\|_2^2$. This case analysis seems somewhat intrinsic to the problem as the correct sample complexity branches into these cases.

For large ϵ, we use the same overall technique as [BC17], noting that p is uniform if and only if $\|p\|_3 = \|p\|_2^{4/3}$, and that for p far from uniform, $\|p\|_3$ must be substantially larger. The basic idea from here is to first obtain rough approximations to $\|p\|_2$ and $\|p\|_3$ in order to ascertain the correct number of samples to use, and then use standard unbiased estimators of $\|p\|_2^2$ and $\|p\|_3^3$ to approximate them to appropriate precision, so that their relative sizes can be compared with appropriate accuracy.

We improve upon the work of [BC17] in this parameter regime in a couple of ways. First, we obtain more precise lower bounds on the difference $\|p\|_3^3 - \|p\|_2^4$ in the case where p is far from uniform (Lemma 2.4). This allows us to reduce the accuracy needed in estimating $\|p\|_2$ and $\|p\|_3$.

Second, we refine the method used for performing the approximations to these moments (ℓ_r-norms). In particular, we observe that using the generic estimators for these quantities yields sub-optimal bounds for the following reason: The error of the unbiased estimators is related to their variance, which in turn can be expressed in terms of the higher moments of p (Fact 2.1). This implies for example that the worst case sample complexity for estimating $\|p\|_3$ comes when the fourth and fifth moments of p are large. However, since we are trying to test for the case of uniformity (where these higher moments are minimal), we do not need to worry about this worst case. In particular, after applying sample efficient tests to ensure that the higher moments of p are not much larger than expected (Lemma 2.2 (ii)), the standard estimators for the second and third moments of p can be shown to converge more rapidly than they would in the worst case (Lemma 2.5).

The above algorithm is not sufficient for small values of ϵ. For ϵ sufficiently small, we employ a different, perhaps more natural, algorithm. Here we take m samples (for m appropriately chosen based on an approximation to $\|p\|_2$) and consider the subset S of the domain that appears in the sample. We then test whether the conditional distribution p on S is uniform, and output the answer of this tester. The number of samples m drawn in the first step is sufficiently large so that $p(S)$, the probability mass of S under p, is relatively high. Hence, it is easy to sample from the conditional distribution using rejection sampling. Furthermore, we can use a standard uniformity testing algorithm requiring $O(\sqrt{|S|}/\epsilon^2)$ samples.

To establish correctness of this algorithm, we need to show that if p is far from uniform, then the conditional distribution p on S is far from uniform as well. To prove this statement, we distinguish two further subcases. If ϵ is "very small", then we can afford to set m sufficiently large so that $p(S)$ is at least $1-\epsilon/10$. In this case, our claim follows straightforwardly. For the remaining values of ϵ, we can only guarantee that $p(S) = \Omega(1)$, hence we require a more sophisticated argument. Specifically, we show (Lemma 2.6) that for any x in an appropriate interval, with high constant probability, the random variable $Z(x) = \sum_{i \in S} |p_i - x|$ is large. It is not hard to show that this holds with high probability for each fixed x, as p being far from uniform implies that $\sum_{i \in \mathbf{\Omega}} \min(p_i, |p_i - x|)$ is large. This latter condition can be shown to provide a clean lower bound for the expectation of $Z(x)$. To conclude the argument, we show that $Z(x)$ is tightly concentrated around its expectation.

Sample Complexity Lower Bound. The lower bound of $\Omega(1/(\epsilon^2\|p\|_2))$ follows directly from the standard lower bound of $\Omega(n^{1/2}/\epsilon^2)$ [Pan08] for uniformity testing on a given domain of size n. Specifically, it is implied from the fact that the hard instances satisfy $\|p\|_2 = \Theta(n^{-1/2})$. The other branch of the lower bound, namely $\Omega(1/(\epsilon^{4/3}\|p\|_3))$, is more involved. To prove this lower bound, we use the shared information method [DK16] for the following family of hard instances: In the "YES" case, we consider the distribution over (pseudo-)distributions on N bins, where each p_i is $(1+\epsilon^2)/N$ with probability $n/(N(1+\epsilon^2))$, and 0 otherwise. (Here we assume that the parameter N is sufficiently large compared to the other parameters.) In the "NO" case, we consider the distribution over (pseudo-)distributions on N bins, where each p_i is $(1+\epsilon)/N$ with probability $n/(2N)$, $(1-\epsilon)/N$ with probability $n/(2N)$, and 0 otherwise.

1.2 Notation

Let $\mathbf{\Omega}$ denote the unknown discrete domain. Each probability distribution over $\mathbf{\Omega}$ can be associated with a probability mass function $p : \mathbf{\Omega} \to \mathbb{R}_+$ such that $\sum_{i \in \mathbf{\Omega}} p_i = 1$. We will use p_i, instead of $p(i)$, to denote the probability of element $i \in \mathbf{\Omega}$ in p. For a distribution (with mass function) p and a set $S \subseteq \mathbf{\Omega}$, we denote by $p(S) \stackrel{\text{def}}{=} \sum_{i \in S} p_i$ and by $(p|S)$ the conditional distribution of p on S. For $r \geq 1$, the ℓ_r-norm of a function $p : \mathbf{\Omega} \to \mathbb{R}$ is $\|p\|_r \stackrel{\text{def}}{=} \left(\sum_{i \in \mathbf{\Omega}} |p_i|^r\right)^{1/r}$. For convenience, we will denote $\mathbf{F}_r(p) \stackrel{\text{def}}{=} \|p\|_r^r = \sum_{i\in\mathbf{\Omega}} |p_i|^r$. For $\emptyset \neq S \subseteq \mathbf{\Omega}$, let $\mathbf{u}_S$ be the uniform distribution over S. Let $\mathcal{C}_U \stackrel{\text{def}}{=} \{\mathbf{u}_S : \emptyset \neq S \subseteq \mathbf{\Omega}\}$ be the set of uniform distributions over subsets of $\mathbf{\Omega}$. The total variation distance between distributions p, q on $\mathbf{\Omega}$ is defined as $d_{\mathrm{TV}}(p,q) \stackrel{\text{def}}{=} \max_{S\subseteq\mathbf{\Omega}} |p(S)-q(S)| = (1/2)\cdot\|p-q\|_1$. Finally, we denote by $\mathrm{Poi}(\lambda)$ the Poisson distribution with parameter λ.

2 Generalized Uniformity Tester

In this section, we give our sample-optimal generalized uniformity tester, Gen-Uniformity-Test. Before we describe our algorithm, we summarize a few preliminary results on estimating the power sums $\mathbf{F}_r(p) = \sum_{i\in\mathbf{\Omega}} |p_i|^r$ of an unknown distribution p. We present these results in Section 2.1. In Section 2.2, we give a detailed pseudo-code for our algorithm. In Section 2.3, we analyze the sample complexity, and in Section 2.4 we provide the proof of correctness.

2.1 Estimating the Power Sums of a Discrete Distribution

We will require various notions of approximation for the power sums of a discrete distribution. We start with the following fact:

Fact 2.1 ([AOST17]). *Let p be a probability distribution on an unknown discrete domain. For any $r \geq 1$, there exists an estimator $\widehat{\mathbf{F}}_r(p)$ for $\mathbf{F}_r(p)$ that draws $\mathrm{Poi}(m)$ samples from p and satisfies the following: $\mathbf{E}\left[\widehat{\mathbf{F}}_r(p)\right] = \mathbf{F}_r(p)$ and $\mathbf{Var}\left[\widehat{\mathbf{F}}_r(p)\right] = m^{-2r}\sum_{t=0}^{r-1} m^{r+t}\binom{r}{t} r^{r-t}\mathbf{F}_{r+t}(p)$.*

The estimator $\widehat{\mathbf{F}}_r(p)$ is standard: It draws $\mathrm{Poi}(m)$ samples from p and $m^r \cdot \widehat{\mathbf{F}}_r(p)$ equals the number of r-wise collisions, i.e., ordered r-tuples of samples that land in the same bin. Using Fact 2.1, we get the following lemma which will be crucial for our generalized uniformity tester:

Lemma 2.2. *Let p be a probability distribution on an unknown discrete domain and $r \geq 1$. We have the following:*

(i) There exists an algorithm that, given a parameter $0 < \delta < 1$ and sample access to p, draws $O(\frac{1}{\delta^2\|p\|_r})$ samples from p in expectation and outputs an estimate $\widehat{\gamma}_r$ that with probability at least $19/20$ satisfies: $|\widehat{\gamma}_r - \mathbf{F}_r(p)| \leq \delta\cdot\mathbf{F}_r(p)$.

(ii) For any $c \geq 1$, there exist an algorithm that draws $\mathrm{Poi}\,(O(m))$ samples from p and correctly distinguishes with probability at least $19/20$ between the case that $m^r\mathbf{F}_r(p) \geq 20c$ versus $m^r\mathbf{F}_r(p) \leq c/20$.

Proof. Using Fact 2.1, it is shown in [AOST17] that if we draw $m = O(\frac{1}{\delta^2\|p\|_r})$ samples from p, then with high constant probability we have that $|\widehat{\mathbf{F}}_r(p) - \mathbf{F}_r(p)| \leq \delta\cdot\mathbf{F}_r(p)$. Since the value of $\|p\|_r$ is unknown, this guarantee does not quite suffice for (i). We instead start by approximating $1/\|p\|_r^r$ within a constant factor. We do this by counting the number of samples we need to draw from p until we see the first r-wise collision. By Fact 2.1 and Chebyshev's inequality, this gives a constant factor approximation to $1/\|p\|_r^r$ with expected sample size of $O(1/\|p\|_r)$. We thus get (i).

We now proceed to show (ii). The algorithm is straightforward: Draw $\mathrm{Poi}\,(O(m))$ samples from p and calculate $\widehat{\mathbf{F}}_r(p)$. If $m^r\widehat{\mathbf{F}}_r(p) > c$, output "large"; otherwise output "small". Suppose that $m^r\mathbf{F}_r(p) \leq c/20$. By Markov's inequality, with probability at least $19/20$ we will have that $m^r\widehat{\mathbf{F}}_r(p) \leq c$, in which case we output "small". Now suppose that $m^r\mathbf{F}_r(p) \geq 20c$. Since $c \geq 1$, this gives that $\|p\|_r \geq 1/m$. Therefore, after we draw $\mathrm{Poi}(O(m))$ samples from p, with probability at least $19/20$ we have that $\widehat{\mathbf{F}}_r(p)$ is a factor 2 approximation to $\mathbf{F}_r(p)$. In other words, $m^r\widehat{\mathbf{F}}_r(p) \geq 10c$ and the algorithm outputs "large". □

2.2 Pseudo-code for Gen-Uniformity-Test Algorithm

The algorithm is given in the following pseudo-code:

Algorithm 1 Sample-Optimal Algorithm for Generalized Uniformity Testing

```
 1: procedure Gen-Uniformity-Test($(p, \epsilon)$)
input: Sample access to arbitrary distribution $p$ on unknown discrete domain $\mathbf{\Omega}$ and $\epsilon > 0$.
output: "YES" with probability 2/3 if $p \in \mathcal{C}_U$, "NO" with probability 2/3 if $d_{\mathrm{TV}}(p, \mathcal{C}_U) \geq \epsilon$.
 2:     Compute an estimate $\widehat{\gamma}_2$ satisfying $|\widehat{\gamma}_2 - \mathbf{F}_2(p)| \leq (1/2) \cdot \mathbf{F}_2(p)$ with probability 19/20.
 3:     $n \leftarrow \lceil 2/\gamma_2 \rceil$.
 4:     if ($\epsilon \geq n^{-1/4}$) then
 5:         Compute an estimate $\widehat{\gamma}_3$ satisfying $|\widehat{\gamma}_3 - \mathbf{F}_3(p)| \leq (1/2) \cdot \mathbf{F}_3(p)$ with probability 19/20.
 6:         if ($\widehat{\gamma}_3 \geq 8/n^2$ or $\widehat{\gamma}_3 \leq 1/(8n^2)$) then return "NO".
 7:         Let $m \leftarrow \Theta(n^{2/3}/\epsilon^{4/3})$, for a sufficiently large constant in the $\Theta()$.
 8:         Let $c_4 = \Theta(1 + m^4/n^3)$, for a sufficiently large constant in the $\Theta()$.
 9:         Draw $\mathrm{Poi}(O(m))$ samples from $p$ and let $\widehat{\gamma}_4$ denote the value of $\widehat{\mathbf{F}}_4(p)$ on this sample.
10:         if $m^4 \widehat{\gamma}_4 > 20c_4$ then return "NO".
11:         Let $c_5 = \Theta(1 + m^5/n^4)$, for a sufficiently large constant in the $\Theta()$.
12:         Draw $\mathrm{Poi}(O(m))$ samples from $p$ and let $\widehat{\gamma}_5$ denote the value of $\widehat{\mathbf{F}}_5(p)$ on this sample.
13:         if $m^5 \widehat{\gamma}_5 > 20c_5$ then return "NO".
14:         Compute the estimates $\widehat{\mathbf{F}}_2(p)$, $\widehat{\mathbf{F}}_3(p)$ on two separate sets of $\mathrm{Poi}(m)$ samples.
15:         if $\left(\widehat{\mathbf{F}}_3(p) - \widehat{\mathbf{F}}_2(p)^2 \leq \epsilon^2/(300n^2)\right)$ then return "YES".
16:         else return "NO".
17:     if ($n^{-1/4} \log^{-1}(n) \leq \epsilon < n^{-1/4}$) then
18:         Let $m_1 \leftarrow \Theta(n)$, for an appropriately large constant in the $\Theta()$.
19:         Draw $\mathrm{Poi}(m_1)$ samples from $p$. Let $S$ be the subset of $\mathbf{\Omega}$ that appears in the sample.
20:         Verify the following conditions: (i) Each $i \in S$ appears $O(\log n)$ times;
21:         (ii) $|S| \geq n/2$; (iii) $p(S) \geq 1/2$.
22:         if (any of conditions (20), (21) is violated) then return "NO".
23:         Using rejection sampling, draw $m_2 \leftarrow O(n^{1/2}/\epsilon^2)$ samples from $(p|S)$.
24:         Test whether $(p|S) = \mathbf{u}_S$ versus $\epsilon/10$-far from $\mathbf{u}_S$ with confidence probability 19/20.
25:         return the answer of the tester in Step 24.
26:     if ($\epsilon < n^{-1/4} \log^{-1}(n)$) then
27:         $m_1 \leftarrow \Theta(n \log n)$, for an appropriately large constant in the $\Theta()$.
28:         Draw $\mathrm{Poi}(m_1)$ samples from $p$. Let $S$ be the subset of $\mathbf{\Omega}$ that appears in the sample.
29:         Draw $m_2 \leftarrow O(n^{1/2}/\epsilon^2)$ samples from $p$.
30:         if (any of the above samples lands outside $S$) then return "NO".
31:         Test whether $(p|S) = \mathbf{u}_S$ versus $\epsilon/2$-far from $\mathbf{u}_S$ with confidence probability 19/20.
32:         return the answer of the tester in Step 31.
```

2.3 Bounding the Sample Complexity

We start by analyzing the sample complexity of the algorithm. We claim that the expected sample complexity is $O\left(1/\left(\epsilon^{4/3}\|p\|_3\right)\right)$ for $\epsilon \geq n^{-1/4}$ and $O\left(1/\left(\epsilon^2\|p\|_2\right)\right)$ for $\epsilon < n^{-1/4}$.

By Lemma 2.2 (i), Step 2 can be implemented with expected sample complexity $O(1/\|p\|_2)$ and Step 5 with expected sample complexity $O(1/\|p\|_3)$.

We start with the case $\epsilon \geq n^{-1/4}$. If Steps 2, 5, and 6 succeed, then we have that $\mathbf{F}_2(p) = \Theta(1/n)$ and $\mathbf{F}_3(p) = \Theta(1/n^2)$. Also note that no further steps are executed unless the condition of Step 6 holds. Note that all subsequent steps that draw samples (Steps 9, 12, and 14) by definition use at most $\mathrm{Poi}(O(m))$ additional samples. Since Step 14 is executed only if $\widehat{\gamma}_3 = \Theta(1/n^2)$, we have that $m = O(\widehat{\gamma}_3^{-1/3}/\epsilon^{4/3}) = O(1/(\epsilon^{4/3}\|p\|_3))$. Therefore, for $\epsilon \geq n^{-1/4}$, the expected sample complexity of the algorithm is bounded by

$$O\left(1/\|p\|_2\right) + O\left(1/\|p\|_3\right) + O\left(1/\left(\epsilon^{4/3}\|p\|_3\right)\right) = O\left(1/\left(\epsilon^{4/3}\|p\|_3\right)\right) .$$

For the case of $n^{-1/4}\log^{-1}(n) \leq \epsilon < n^{-1/4}$, the additional sample size drawn on top of Step 2 is $O(n + n^{1/2}/\epsilon^2) = O(n^{1/2}/\epsilon^2)$. Since $n = \Theta(1/\|p\|_2^2)$, the total sample complexity in this case is

$$O\left(1/\|p\|_2\right) + O\left(1/\left(\epsilon^2\|p\|_2\right)\right) = O\left(1/\left(\epsilon^2\|p\|_2\right)\right) .$$

Finally, for $\epsilon < n^{-1/4}\log^{-1}(n)$, the sample size drawn on top of Step 2 is $O(n\log n + n^{1/2}/\epsilon^2) = O(n^{1/2}/\epsilon^2)$. Since $n = \Theta(1/\|p\|_2^2)$, the total sample complexity in this case is $O\left(1/\left(\epsilon^2\|p\|_2\right)\right)$, as before. This completes the analysis of the sample complexity.

2.4 Correctness Proof

This section is devoted to the correctness proof of Gen-Uniformity-Test. In particular, we will show that if $p \in \mathcal{C}_U$, the algorithm outputs "YES" with probability at least 2/3 (completeness); and if $d_{\mathrm{TV}}(p, \mathcal{C}_U) \geq \epsilon$, the algorithm outputs "NO" with probability at least 2/3 (soundness).

We start with the following simple claim giving a useful condition for the soundness case:

Claim 2.3. *If $d_{\mathrm{TV}}(p, \mathcal{C}_U) \geq \epsilon$, then for all $x \in [0,1]$ we have that $\sum_{i\in\mathbf{\Omega}} \min\{p_i, |x - p_i|\} \geq \epsilon/2$.*

Proof. Let S_h be the set of $i \in \mathbf{\Omega}$ on which $p_i > x/2$. Let $\delta = \sum_{i\in\mathbf{\Omega}} \min\{p_i, |x - p_i|\}$. Note that $\delta = \|p - c_{x,S_h}\|_1$, where c_{x,S_h} is the pseudo-distribution that is x on S_h on 0 elsewhere. If $\|c_{x,S_h}\|_1$ were 1, c_{x,S_h} would be the uniform distribution $\mathbf{u}_{S_h}$ and we would have $\delta \geq \epsilon$. However, this need not be the case. That said, it is easy to see that $\|\mathbf{u}_{S_h} - c_{x,S_h}\|_1 = |1 - \|c_{x,S_h}\|_1| \leq \|p - c_{x,S_h}\|_1 = \delta$. Therefore, by the triangle inequality

$$2\delta \geq \|p - c_{x,S_h}\|_1 + \|\mathbf{u}_{S_h} - c_{x,S_h}\|_1 \geq \|p - \mathbf{u}_{S_h}\|_1 \geq \epsilon .$$

This completes the proof of Claim 2.3. □

We now proceed to analyze correctness for the various ranges of ϵ.

Case I: [$\epsilon \geq n^{-1/4}$]. The following structural lemma provides a reformulation of generalized uniformity testing in terms of the second and third norms of the unknown distribution:

Lemma 2.4. *We have the following:*

(i) If $p \in \mathcal{C}_U$, then $\mathbf{F}_3(p) = \mathbf{F}_2^2(p)$.

(ii) If $d_{\mathrm{TV}}(p, \mathcal{C}_U) \geq \epsilon$, then $\mathbf{F}_3(p) - \mathbf{F}_2^2(p) > \epsilon^2\mathbf{F}_2^2(p)/64$.

Proof. The proof of (i) is straightforward. Suppose that $p = \mathbf{u}_S$ for some $\emptyset \neq S \subseteq \mathbf{\Omega}$. It then follows that $\mathbf{F}_2(p) = 1/|S|$ and $\mathbf{F}_3(p) = 1/|S|^2$, yielding part (i) of the lemma.

We now proceed to prove part (ii). Suppose that $d_{\mathrm{TV}}(p, \mathcal{C}_U) \geq \epsilon$. First, it will be useful to rewrite the quantity $\mathbf{F}_3(p) - \mathbf{F}_2^2(p)$ as follows:

$$\mathbf{F}_3(p) - \mathbf{F}_2^2(p) = \sum_{i \in \mathbf{\Omega}} p_i (p_i - \mathbf{F}_2(p))^2 \ . \tag{1}$$

Note that (1) follows from the identity $p_i(p_i - \mathbf{F}_2(p))^2 = p_i^3 + p_i \mathbf{F}_2(p)^2 - 2p_i^2 \mathbf{F}_2(p)$ by summing over $i \in \mathbf{\Omega}$. Since $d_{\mathrm{TV}}(p, \mathcal{C}_U) \geq \epsilon$, an application of Claim 2.3 for $x = \mathbf{F}_2(p) \in [0, 1]$, gives that

$$\sum_{i \in \mathbf{\Omega}} \min\{p_i, |\mathbf{F}_2(p) - p_i|\} \geq \epsilon/2 \ .$$

We partition $\mathbf{\Omega}$ into the sets $S_l = \{i \in \mathbf{\Omega} \mid p_i < \mathbf{F}_2(p)/2\}$ and its complement $S_h = \mathbf{\Omega} \setminus S_l$. Note that $\sum_{i \in \mathbf{\Omega}} \min\{p_i, |\mathbf{F}_2(p) - p_i|\} = \sum_{i \in S_l} p_i + \sum_{i \in S_h} |\mathbf{F}_2(p) - p_i|$. It follows that either $\sum_{i \in S_l} p_i \geq \epsilon/4$ or $\sum_{i \in S_h} |\mathbf{F}_2(p) - p_i| \geq \epsilon/4$. We analyze each case separately. First, suppose that $\sum_{i \in S_l} p_i \geq \epsilon/4$. Using (1) we can now write

$$\mathbf{F}_3(p) - \mathbf{F}_2^2(p) \geq \sum_{i \in S_l} p_i (p_i - \mathbf{F}_2(p))^2 > (\mathbf{F}_2(p)/2)^2 \cdot \sum_{i \in S_l} p_i = \epsilon \mathbf{F}_2^2(p)/16 \ .$$

Now suppose that $\sum_{i \in S_h} |\mathbf{F}_2(p) - p_i| \geq \epsilon/4$. Note that $1 \leq |S_h| \leq 2/|\mathbf{F}_2(p)|$. In this case, using (1) we obtain

$$\begin{aligned}
\mathbf{F}_3(p) - \mathbf{F}_2^2(p) &\geq \sum_{i \in S_h} p_i (p_i - \mathbf{F}_2(p))^2 \\
&\geq (\mathbf{F}_2(p)/2) \cdot \sum_{i \in S_h} (p_i - \mathbf{F}_2(p))^2 \\
&\geq (\mathbf{F}_2(p)/2) \cdot \frac{(\sum_{i \in S_h} |\mathbf{F}_2(p) - p_i|)^2}{|S_h|} \\
&\geq (\mathbf{F}_2(p)/2)^2 \cdot (\epsilon/4)^2 \\
&= \epsilon^2 \mathbf{F}_2^2(p)/64 \ ,
\end{aligned}$$

where the second inequality uses the definition of S_h, and the third inequality is Cauchy-Schwarz. This completes the proof of Lemma 2.4. □

By Lemma 2.4, the proof in this case boils down to proving that our estimates for $\mathbf{F}_2(p)$ and $\mathbf{F}_3(p)$ obtained in Step 14 are sufficiently accurate to distinguish between the completeness and soundness cases. We note that since Steps (6), (10), and (13) have succeeded, with probability at least 19/20 each of the corresponding conditions is satisfied. Specifically, this implies that the following conditions hold: $\mathbf{F}_2(p) = \Theta(1/n)$, $\mathbf{F}_3(p) = \Theta(1/n^2)$, $\mathbf{F}_4(p) = O(m^{-4} + n^{-3})$, and $\mathbf{F}_5(p) = O(m^{-5} + n^{-4})$.

We henceforth condition on this event. The following lemma shows that our approximations to the second and third moments are appropriately accurate:

Lemma 2.5. *Let c be an appropriately small universal constant (selecting $c = 10^{-3}$ suffices). With probability at least 9/10 over the samples, the estimates for $\mathbf{F}_2(p)$ and $\mathbf{F}_3(p)$ obtained in Step 14 satisfy the following conditions:*

(i) $|\widehat{\mathbf{F}_2}(p) - \mathbf{F}_2(p)| \leq c \cdot \epsilon^2 \mathbf{F}_2(p)$.

(ii) $|\widehat{\mathbf{F}_3}(p) - \mathbf{F}_3(p)| \leq c \cdot \epsilon^2 \mathbf{F}_2^2(p)$.

Proof. The lemma follows using Fact 2.1 and an application of Chebyshev's inequality, crucially exploiting the improved variance bounds that hold when the above conditions are satisfied.

To prove part (i), note that $\mathbf{Var}[\widehat{\mathbf{F}}_2(p)] = O\left(m^{-2}\mathbf{F}_2(p) + m^{-1}\mathbf{F}_3(p)\right)$. We use that $\mathbf{F}_3(p) = \Theta(1/n^2) = \Theta(\mathbf{F}_2^2(p))$, where the second inequality uses the fact that $1/n = \Theta(\mathbf{F}_2(p))$ (as follows from Steps 2 and 3 of the algorithm). Now recall that the sample size m is defined to be $\Theta(n^{2/3}/\epsilon^{4/3})$, for a sufficiently large universal constant in the big-Θ. We can therefore bound the variance $\mathbf{Var}[\widehat{\mathbf{F}}_2(p)]$ from above by

$$O\left(m^{-2}n^{-1} + m^{-1}n^{-2}\right) = O\left(\epsilon^{8/3}n^{-7/3} + \epsilon^{4/3}n^{-8/3}\right) = O(\epsilon^4/n^2) ,$$

where we used the assumption that $\epsilon \geq n^{-1/4}$. By Chebyshev's inequality, we therefore get that

$$|\widehat{\mathbf{F}_2}(p) - \mathbf{F}_2(p)| \leq O(\epsilon^2/n) , \tag{2}$$

with probability at least 19/20. By selecting the constant factor in the definition of m appropriately, we can make the RHS in (2) at most $c \cdot \epsilon^2 \mathbf{F}_2(p)$, as desired.

Part (ii) is proved similarly. We have that $\mathbf{Var}[\widehat{\mathbf{F}}_3(p)] = O\left(m^{-3}\mathbf{F}_3(p) + m^{-2}\mathbf{F}_4(p) + m^{-1}\mathbf{F}_5(p)\right)$. We use that $\mathbf{F}_3(p) = \Theta(1/n^2)$, $\mathbf{F}_4(p) = O(m^{-4} + n^{-3})$, and $\mathbf{F}_5(p) = O(m^{-5} + n^{-4})$. Recalling that the sample size m is defined to be $\Theta(n^{2/3}/\epsilon^{4/3})$, we can bound the variance $\mathbf{Var}[\widehat{\mathbf{F}}_3(p)]$ from above by

$$O\left(m^{-3}n^{-2} + m^{-6} + m^{-2}n^{-3} + m^{-1}n^{-4}\right) = O\left(\epsilon^4/n^4\right) ,$$

where we used the assumption that $m = \Theta(n^{2/3}/\epsilon^{4/3})$ and $\epsilon \geq n^{-1/4}$. By Chebyshev's inequality, we therefore get that

$$|\widehat{\mathbf{F}_3}(p) - \mathbf{F}_3(p)| \leq O(\epsilon^2/n^2) , \tag{3}$$

with probability at least 19/20. By selecting the constant in the big-Θ defining m appropriately, it is clear that we can make the RHS in (3) at most $c \cdot \epsilon^2 \mathbf{F}_2^2(p)$, as desired. This completes the proof of Lemma 2.5. □

We now have all the necessary ingredients to establish completeness and soundness in Case I. If $p \in \mathcal{C}_U$, it is easy to see that Steps (6), (10), and (13) succeed with high constant probability, as follows from the fact that the norms are minimal in this case and Lemma 2.2. Moreover, if the algorithm does not reject in any of these steps, the corresponding conditions on the magnitude of these norms are satisfied. If the conditions of Lemma 2.5 hold, then we have that

$$\left|\left(\mathbf{F}_3(p) - \mathbf{F}_2^2(p)\right) - \left(\widehat{\mathbf{F}}_3(p) - \widehat{\mathbf{F}}_2(p)^2\right)\right| \leq c \cdot \epsilon^2 \mathbf{F}_2^2(p) .$$

Therefore, the algorithm correctly distinguishes between the completeness and soundness cases, via Lemma 2.4. This completes the correctness analysis of Case I.

Case II: $[n^{-1/4}\log^{-1}(n) \leq \epsilon < n^{-1/4}]$. The correctness in the completeness case is straightforward. If $p \in \mathcal{C}_U$, it is easy to see that Conditions 20 and 21 will be satisfied with high constant probability. Moreover, the conditional distribution $(p|S)$ equals $\mathbf{u}_S$, and therefore the overall algorithm outputs "YES" with high constant probability.

The correctness of the soundness case is more involved. Suppose that $d_{\mathrm{TV}}(p, \mathcal{C}_U) \geq \epsilon$. If the algorithm does not output "NO" in Step 22, the following conditions hold with high probability: (a) $|S| \geq n/2$, (b) $p(S) \geq 1/2$, and (c) $p_i = O(\log n/n)$ for all $i \in \mathbf{\Omega}$. We will use these statements to prove the following lemma:

Lemma 2.6. *If $d_{\mathrm{TV}}(p, \mathcal{C}_U) \geq \epsilon$ and the conditions in Steps 20, 21 hold, then with high constant probability over the samples drawn in Step 19, we have that $d_{\mathrm{TV}}((p|S), \mathbf{u}_S) \geq \epsilon/10$.*

Proof. Suppose that $d_{\mathrm{TV}}(p, \mathcal{C}_U) \geq \epsilon$. We want to show that with high probability over the samples it holds $\sum_{i \in S} |p_i - p(S)/|S|| = \Omega(\epsilon)$. The main difficulty is that the value of $p(S)$ is unknown, hence we need a somewhat indirect argument. By Claim 2.3, for all $x \in [0, 1]$ we have that

$$\sum_{i \in \mathbf{\Omega}} \min\{p_i, |p_i - x|\} \geq \epsilon/2 \ . \tag{4}$$

To show that $\sum_{i \in S} |p_i - p(S)/|S|| = \Omega(\epsilon)$, it suffices to prove that the following holds:

Claim 2.7. *With probability at least 19/20, for all x in an additive grid with step size $O(\epsilon/n)$ such that $0 \leq x \leq \log n/n$, we have that $Z(x) \overset{\mathrm{def}}{=} \sum_{i \in S} |p_i - x| = \Omega(\epsilon)$.*

First note that for $x > 4/n$ or $x < 1/(4n)$, the above claim is satisfied automatically. Indeed, for $x > 4/n$, we have $\sum_{i \in S} |p_i - x| \geq |S| \cdot x - p(S) \geq (n/2)x - 1 \geq 1$. For $x < 1/(4n)$, we have $\sum_{i \in S} |p_i - x| \geq p(S) - |S| \cdot x \geq 1/2 - nx \geq 1/4$.

We henceforth focus on the setting where $1/(4n) \leq x \leq 4/n$. Here we show that $\mathbf{E}[Z(x)]$ is large and that Z is tightly concentrated around its expectation.

Let Z_i, $i \in \mathbf{\Omega}$, be the indicator of the event $i \in S$. Then, $Z(x) = \sum_{i \in \mathbf{\Omega}} |p_i - x| Z_i$. Note that Z_i is a Bernoulli random variable with $\mathbf{E}[Z_i] = 1 - e^{-p_i n}$ and that the Z_i's are mutually independent. Note that $\mathbf{E}[Z(x)] = \sum_{i \in \mathbf{\Omega}} (1 - e^{-p_i n}) |p_i - x|$. We recall the following concentration inequality for sums of non-negative random variables (see, e.g., Exercise 2.9 in [BLM13]):

Fact 2.8. *Let $X_1, \ldots, X_m$ be independent non-negative random variables, and $X = \sum_{j=1}^m X_j$. Then, for any $t > 0$, it holds that* $\Pr[X \leq \mathbf{E}[X] - t] \leq \exp\left(-t^2/(2\sum_{i=1}^m \mathbf{E}[X_i^2])\right)$.

Since $Z(x) = \sum_{i \in \mathbf{\Omega}} |p_i - x| Z_i$ where the Z_i's are independent Bernoulli random variables with $\mathbf{E}[Z_i^2] = 1 - e^{-p_i n}$, an application of Fact 2.8 yields that

$$\Pr\left[Z(x) \leq \mathbf{E}[Z(x)] - t\right] \leq \exp\left(\frac{-t^2}{2\sum_{i \in \mathbf{\Omega}} (1 - e^{-p_i n})(p_i - x)^2} \ .\right) \tag{5}$$

Let $S_l = \{i \in \mathbf{\Omega} : p_i \leq x/2\}$ and $S_h = \mathbf{\Omega} \setminus S_l$. By (4), we get that $\sum_{i \in S_l} p_i + \sum_{i \in S_h} |x - p_i| \geq \epsilon/2$. For $i \in S_l$, we have that $(1 - e^{-p_i n})|p_i - x| \geq n \cdot p_i \cdot |x/2| = \Omega(p_i)$. For $i \in S_h$, we have that $(1 - e^{-p_i n}) = \Omega(1)$ and therefore $(1 - e^{-p_i n})|p_i - x| = \Omega(1)|p_i - x|$. We therefore get that $\mathbf{E}[Z(x)] = \Omega(\epsilon)$. We now bound $\sum_{i \in \mathbf{\Omega}} (1 - e^{-p_i n})(p_i - x)^2$ from above using the fact that $p_i \leq \log n/n$, for all $i \in \Omega$. This assumption and the range of x imply that

$$\sum_{i \in \mathbf{\Omega}} (1 - e^{-p_i n})(p_i - x)^2 \leq O(\log n/n) \cdot \mathbf{E}[Z] \ .$$

So, by setting $t = \mathbf{E}[Z]/2$ in (5), we get that

$$\Pr[Z(x) \leq \mathbf{E}[Z(x)]/2] \leq \exp\left(-\Omega(\epsilon n/\log n)\right) = \exp\left(-n^{\Omega(1)}\right) \ ,$$

where the last inequality follows from the range of ϵ. Recalling that x lies in a grid of size $O(n/\epsilon)$, Claim 2.7 follows by a union bound. This completes the analysis of Case II.

Case III: $[\epsilon < n^{-1/4} \log^{-1}(n)]$. The correctness in this case is quite simple. In the completeness case, conditioning on Step 2 succeeding, we know that p is uniform over a domain of size $O(n)$.

Therefore, after $\Theta(n \log n)$ samples, we have seen all the elements of the domain with high probability, i.e., the set S has $p(S) = 1$. Therefore, the conditional distribution $p|S$ is identified with p, and the final tester outputs "YES". On the other hand, if p is ϵ-far from uniform. and the algorithm does not reject in Step 30, then it follows that $p(S) \geq 1 - O(\epsilon/n^{1/4}) > 1 - \epsilon/10$. Therefore, $p|S$ should be at least $\epsilon/2$-far from $\mathbf{u}_S$ and the tester will output "NO." This completes the proof of correctness. □

3 Sample Complexity Lower Bound

In this section, we prove a sample size lower bound matching our algorithm Gen-Uniformity-Test. One part of the lower bound is fairly easy. In particular, it is known [Pan08] that $\Omega(\sqrt{n}/\epsilon^2)$ samples are required to test uniformity of a distribution with a known support of size n. It is easy to see that the hard cases for this lower bound still work when $\|p\|_2 = \Theta(n^{-1/2})$.

The other half of the lower bound is somewhat more difficult and we rely on the lower bound techniques of [DK16]. In particular, for $n > 0$, and $1/10 > \epsilon > n^{-1/4}$ and for N sufficiently large, we produce a pair of distributions $\mathcal{D}$ and $\mathcal{D}'$ over positive measures on $[N]$, so that:

1. A random sample from $\mathcal{D}$ or $\mathcal{D}'$ has total mass $\Theta(1)$ with high probability.
2. A random sample from $\mathcal{D}$ or $\mathcal{D}'$ has support of size $\Theta(n)$ with high probability.
3. A sample from $\mu \in \mathcal{D}$ has $\mu/\|\mu\|_1$ be the uniform distribution over some subset of $[N]$ with probability 1.
4. A sample from $\mu \in \mathcal{D}'$ has $\mu/||\mu||_1$ be at least $\Omega(\epsilon)$-far from any uniform distribution with high probability.
5. Given a measure μ taking randomly from either $\mathcal{D}$ or $\mathcal{D}'$, no algorithm given the output of a Poisson process with intensity $k\mu$ for $k = o(\min(n^{2/3}/\epsilon^{4/3}, n))$ can reliably distinguish between a μ taken from $\mathcal{D}$ and μ taken from $\mathcal{D}'$.

Before we exhibit these families, we first discuss why the above is sufficient. This Poissonization technique has been used previously in various settings [VV14, DK16, WY16, DGPP17], so we only provide a sketch here. In particular, suppose that we have such families $\mathcal{D}$ and $\mathcal{D}'$, but that there is also an algorithm A that distinguishes between a distribution p being uniform and being ϵ-far from uniform in $m = o(\epsilon^{-4/3}/\|p\|_3)$ samples. We show that we can use algorithm A to violate property 5 above. In particular, letting $p = \mu/\|\mu\|_1$ for μ a random measure taken from either $\mathcal{D}$ or $\mathcal{D}'$, we note that with high probability p has support of size $\Theta(n)$, and thus $\|p\|_3 = O(n^{-2/3})$. Therefore, $m' = o(n^{2/3}/\epsilon^{4/3})$ samples are sufficient to distinguish between p being uniform and being $\Omega(\epsilon)$ far from uniform. However, by properties 3 and 4, this is equivalent to distinguish between μ being taken from $\mathcal{D}$ and being taken from $\mathcal{D}'$. On the other hand, given the output of a Poisson process with intensity $Cm'\mu$, for C a sufficiently large constant, a random m' of these samples (note that there are at least m' total samples with high probability) are distributed identically to m' samples from p. Thus, applying A to these samples distinguishes between μ taken from $\mathcal{D}$ and μ taken from $\mathcal{D}'$, thus contradicting property 5.

We now exhibit the families $\mathcal{D}$ and $\mathcal{D}'$. In both cases, we want to arrange $\mu_i := \mu(\{i\})$ to be i.i.d. for different i. We also want it to be the case that the first and second moments of μ_i are the same for $\mathcal{D}$ and $\mathcal{D}'$. Combining this with requirements on closeness to uniform, we are led to the following definitions:

For μ taken from $\mathcal{D}'$, we let

$$\mu_i = \begin{cases} \frac{1+\epsilon}{n} & \text{, with probability } \frac{n}{2N} \\ \frac{1-\epsilon}{n} & \text{, with probability } \frac{n}{2N} \\ 0 & \text{, otherwise .} \end{cases}$$

For μ taken from $\mathcal{D}$, we let

$$\mu_i = \begin{cases} \frac{1+\epsilon^2}{n} & \text{, with probability } \frac{n}{N(1+\epsilon^2)} \\ 0 & \text{, otherwise .} \end{cases}$$

Note that in both cases, the average total mass is 1, and it is easy to see by Chernoff bounds that the actual mass of μ is $\Theta(1)$ with high probability. Additionally, the support size is always $\Theta(n)$ times the total mass, and so is $\Theta(n)$ with high probability. For μ taken from $\mathcal{D}$, all of the μ_i are either 0 or $\frac{1+\epsilon^2}{n}$, and thus $\mu/\|\mu\|_1$ is uniform over its support. For μ taken from $\mathcal{D}'$, with high probability at least a third of the bins in its support have $\mu_i = \frac{1+\epsilon}{n}$, and at least a third have $\mu_i = \frac{1-\epsilon}{n}$. If this is the case, then at least a constant fraction of the mass of $\mu/\|\mu\|_1$ comes from bins with mass off from the average mass by at least a $(1 \pm \epsilon)$ factor, and this implies that $\mu/\|\mu\|_1$ is at least $\Omega(\epsilon)$-far from uniform.

We have thus verified 1-4. Property 5 will be somewhat more difficult to prove. For this, let X be a random $\{0, 1\}$ random variable with equal probabilities. Let μ be chosen randomly from $\mathcal{D}$ if $X = 0$, and randomly from $\mathcal{D}'$ if $X = 1$. Let our Poisson process with intensity $k\mu$ return A_i samples from bin i. We note that, by the same arguments as in [DK16], it suffices to show that the shared information $I(X; A_1, \ldots, A_N) = o(1)$. In order to prove this, we note that the A_i are conditionally independent on X, and thus we have that $I(X; A_1, \ldots, A_N) \leq \sum_{i=1}^{N} I(X; A_i) = NI(X; A_1)$. Thus, we need to show that $I(X; A_1) = o(1/N)$. For notational simplicity, we drop the subscript in A_1.

This boils down to an elementary but tedious calculation. We begin by noting that we can bound

$$I(X; A) = \sum_{t=0}^{\infty} O\left(\frac{(\Pr(A = t|X = 0) - \Pr(A = t|X = 1))^2}{\Pr(A = t)}\right) .$$

(This calculation is standard. See Fact 81 in [CDKS17] for a proof.) We seek to bound each of these terms. The distribution of A conditioned on μ_1 is Poisson with parameter $k\mu_1$. Thus, the distribution of A conditioned on X is a mixture of two or three Poisson distributions, one of which is the trivial constant 0. We start by giving explicit expressions for these probabilities.

Firstly, for the $t = 0$ term, note that

$$\Pr(A = t|X = 1) = 1 - \frac{n}{N}\left(1 - \frac{e^{-k(1+\epsilon)/n} + e^{-k(1-\epsilon)/n}}{2}\right) ,$$

$$\Pr(A = t|X = 0) = 1 - \frac{n}{N(1+\epsilon^2)}(1 - e^{-k(1+\epsilon^2)/n}) .$$

Note that $\Pr(A = 0)$ is at least $1 - n/N \geq 1/2$ and $\Pr(A = t|X = 1) - \Pr(A = t|X = 0) \leq n/N$. Thus, the contribution from this term, $\frac{(\Pr(A=0|X=0)-\Pr(A=0|X=1))^2}{\Pr(A=0)}$, is $O(n/N)^2 = o(1/N)$.

For $t \geq 1$, there is no contribution from $\mu_1 = 0$. We can compute the probabilities involved exactly as

$$\Pr(A = t|X = 1) = \frac{n}{N}\frac{(k(1+\epsilon)/n)^t e^{-k(1+\epsilon)/n} + (k(1-\epsilon)/n)^t e^{-k(1-\epsilon)/n}}{2t!} ,$$

$$\Pr(A=t|X=0) = \frac{n}{N(1+\epsilon^2)} \frac{(k(1+\epsilon^2)/n)^t e^{-k(1+\epsilon^2)/n}}{t!} ,$$

and obtain that $\frac{(\Pr(A=t|X=0)-\Pr(A=t|X=1))^2}{\Pr(A=t)}$ is

$$O\left(\left(\frac{n^{1-t}k^t}{2Nt!}\right)\frac{\left((1+\epsilon)^t e^{-k(1+\epsilon)/n}+(1-\epsilon)^t e^{-k(1-\epsilon)/n}-2(1+\epsilon^2)^{t-1}e^{-k(1+\epsilon^2)/n}\right)^2}{(1+\epsilon)^t e^{-k(1+\epsilon)/n}+(1-\epsilon)^t e^{-k(1-\epsilon)/n}+2(1+\epsilon^2)^{t-1}e^{-k(1+\epsilon^2)/n}}\right) .$$

Factoring out the $e^{-k/n}$ terms and noting that, since $k\epsilon/n = o(1)$, the denominator is $\Omega(e^{-k/n})$ yields that

$$O\left(\left(\frac{n^{1-t}k^t e^{-k/n}}{2Nt!}\right)\left((1+\epsilon)^t e^{-k(1+\epsilon)/n}+(1-\epsilon)^t e^{-k(1-\epsilon)/n}-2(1+\epsilon^2)^{t-1}e^{-k(1+\epsilon^2)/n}\right)^2\right) .$$

Noting that $k/n = o(1)$, we can ignore this e^{-kn} term and Taylor expanding the exponentials, we have that

$$\begin{aligned}&\frac{(\Pr(A=t|X=0)-\Pr(A=t|X=1))^2}{\Pr(A=t)} = \\ &O\left(\left(\frac{n^{1-t}k^t}{2Nt!}\right)\left((1+\epsilon)^t(1-k(1+\epsilon)/n)+(1-\epsilon)^t(1+k(1-\epsilon)/n)\right.\right. \\ &\left.\left.-\ 2(1+\epsilon^2)^{t-1}(1-k(1+\epsilon^2)/n)+O((k\epsilon/n)^2(1+\epsilon)^t)\right)^2\right) .\end{aligned}$$

We deal separately with the cases $t=1, t=2$ and $t>2$. For the $t=1$ term, we have

$$\begin{aligned}&O\left(\left(\frac{k}{N}\right)\left((1+\epsilon)(1-k\epsilon/n)+(1-\epsilon)(1+k\epsilon/n)-2(1-k\epsilon^2/n)+O((k\epsilon/n)^2)\right)^2\right) \\ =&O\left(\left(\frac{k}{N}\right)O((k\epsilon/n)^2)^2\right) .\end{aligned}$$

Since $k = o(n^{2/3}/\epsilon^{4/3})$ and $\epsilon > n^{-1/4}$, $\epsilon k/n = o(n^{-1/3}/\epsilon^{1/3}) = o(n^{-1/4})$, and we find that this is

$$O\left(\left(\frac{k}{N}\right)o(1/n)\right) = o(1/N) .$$

This appropriately bounds the contribution from this term.

When $t=2$, we have

$$\begin{aligned}&O\left(\left(\frac{k^2}{nN}\right)\left((1+\epsilon)^2(1-k(1+\epsilon)/n)+(1-\epsilon)^2(1-k(1-\epsilon)/n)\right.\right. \\ &\left.\left.-2(1+\epsilon^2)(1-k(1+\epsilon^2)/n)+O((k\epsilon/n)^2)\right)^2\right) .\end{aligned}$$

Note that the terms without k/n factors cancel out, $(1+\epsilon)^2+(1-\epsilon)^2-2(1+\epsilon^2)=0$, yielding

$$O(k^2/nN)(k\epsilon^2/n+o(n^{-1/2}))^2 = O(k^4\epsilon^4/n^3N)+o(k^2/n^2N) = o(k^3\epsilon^4/n^2N)+o(1/N) = o(1/N) ,$$

using both $k = o(n^{2/3}/\epsilon^{4/3})$ and $k = o(n)$.

For $t > 2$, we let $f_t(x) = (1+x)^t(1-kx/n)$. In terms of f_t, we have that $\frac{(\Pr(A=t|X=0)-\Pr(A=t|X=1))^2}{\Pr(A=t)}$ is:

$$O\left(\left(\frac{n^{1-t}k^t}{2Nt!}\right)(f_t(\epsilon)+f_t(-\epsilon))/2 - f_t(0) - (f_{t-1}(\epsilon^2)-f_{t-1}(0)) + o(n^{-1/2})^2\right) .$$

Using the Taylor expansion of f_t in terms of its first two derivatives and f_{t-1} in terms of its first, we see that

$$(f_t(\epsilon)+f_t(-\epsilon))/2 - f_t(0) = \epsilon^2 f_t''(\xi)$$

and

$$f_{t-1}(\epsilon^2) - f_{t-1}(0) = \epsilon^2 f_{t-1}'(\xi') ,$$

for some $\xi \in [-\epsilon, \epsilon]$ and $\xi' \in [0, \epsilon^2]$. However, the derivatives are

$$f_t'(x) = (1+x)^{t-1}(t-(1+x+tx)k/n)$$

and

$$f_t''(x) = (1+x)^{t-2}(t(t-1)-t(t+1)xk/n) ,$$

and so $|f_t''(\xi)| \leq O(t^2(1+\epsilon)^{t-1})$ and $f_{t-1}'(\xi') \leq O(t(1+\epsilon^2)^{t-2})$. Hence, the term

$$\frac{(\Pr(A=t|X=0)-\Pr(A=t|X=1))^2}{\Pr(A=t)}$$

is at most

$$\begin{aligned}
&O(n^{1-t}k^t/Nt!)(\epsilon^4 t^4(1+\epsilon)^{2t-2}) + o(1/n))\\
&= O\left((k^3\epsilon^4/n^2)(t^4(1+\epsilon)^2/N)(k(1+\epsilon)^2/n)^{t-3}/t!\right) + o\left((k/n)^t/(Nt!)\right)\\
&= o(1/N)t^4/t! ,
\end{aligned}$$

using both $k = o(n^{2/3}/\epsilon^{4/3})$ and $k = o(n)$. Since $(t+1)^4/(t+1)! \leq t^4/2t!$ for all $t \geq 4$, even summing the above over all $t \geq 3$ still leaves $o(1/N)$.

Thus, we have that $I(X;A) = o(1/N)$, and therefore that $I(X : A_1, \ldots, A_N) = o(1)$. This proves that $X = 0$ and $X = 1$ cannot be reliably distinguished given $A_1, \ldots, A_N$, and thus proves property 5, completing the proof of our lower bound.

4 Conclusions

In this paper, we gave tight upper and lower bounds on the sample complexity of generalized uniformity testing – a natural non-trivial generalization of uniformity testing, recently introduced in [BC17]. The obvious research question is to understand the sample complexity of testing more general symmetric properties (e.g., closeness, independence, etc.) for the regime where the domain of the underlying distributions is discrete but unknown (of unknown size). Is it possible to obtain sub-learning sample complexities for these problems? And what is the optimal sample complexity for each of these tasks? It turns out that the answer to the first question is affirmative. These extensions require more sophisticated techniques and will appear in a forthcoming work.

References

[ADK15] J. Acharya, C. Daskalakis, and G. Kamath. Optimal testing for properties of distributions. In *NIPS*, pages 3591–3599, 2015.

[AOST17] J. Acharya, A. Orlitsky, A. T. Suresh, and H. Tyagi. Estimating renyi entropy of discrete distributions. *IEEE Trans. Information Theory*, 63(1):38–56, 2017.

[BC17] T. Batu and C. Canonne. Generalized uniformity testing. *CoRR*, abs/1708.04696, 2017. To appear in FOCS'17.

[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.

[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.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[Can15] 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.

[CDGR16] C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. Testing shape restrictions of discrete distributions. In *33rd Symposium on Theoretical Aspects of Computer Science, STACS 2016*, pages 25:1–25:14, 2016.

[CDKS17] C. L. Canonne, I. Diakonikolas, D. M. Kane, and A. Stewart. Testing bayesian networks. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 370–448, 2017.

[CDS17] C. L. Canonne, I. Diakonikolas, and A. Stewart. Fourier-based testing for families of distributions. *CoRR*, abs/1706.05738, 2017.

[CDVV14] S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014.

[DDK16] C. Daskalakis, N. Dikkala, and G. Kamath. Testing ising models. *CoRR*, abs/1612.03147, 2016.

[DDS$^+$13] C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *SODA*, pages 1833–1852, 2013.

[DGPP16] 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.

[DGPP17] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. Sample-optimal identity testing with high probability. *CoRR*, abs/1708.02728, 2017.

[DK16] 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.

[DKN15a] 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.

[DKN15b] 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.

[DKN17] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Near-optimal closeness testing of discrete histogram distributions. In *44th International Colloquium on Automata, Languages, and Programming, ICALP 2017*, pages 8:1–8:15, 2017.

[DP17] C. Daskalakis and Q. Pan. Square hellinger subadditivity for bayesian networks and its applications to identity testing. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 697–703, 2017.

[Gol16] O. Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. *ECCC*, 23, 2016.

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

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

[RRSS09] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM J. Comput.*, 39(3):813–842, 2009.

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

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

[VV14] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *FOCS*, 2014.

[WY16] 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.