Tugkan Batu, Clément L. Canonne

Generalized uniformity testing

Book section

Generalized Uniformity Testing

Tuğkan Batu[*] Clément L. Canonne[†]

August 16, 2017

Abstract

In this work, we revisit the problem of *uniformity testing* of discrete probability distributions. A fundamental problem in distribution testing, testing uniformity over a *known* domain has been addressed over a significant line of works, and is by now fully understood.

The complexity of deciding whether an unknown distribution is uniform over its unknown (and arbitrary) *support*, however, is much less clear. Yet, this task arises as soon as no prior knowledge on the domain is available, or whenever the samples originate from an unknown and unstructured universe. In this work, we introduce and study this *generalized* uniformity testing question, and establish nearly tight upper and lower bound showing that – quite surprisingly – its sample complexity significantly differs from the known-domain case. Moreover, our algorithm is intrinsically *adaptive*, in contrast to the overwhelming majority of known distribution testing algorithms.

[*] London School of Economics. Email: t.batu@lse.ac.uk.

[†] Columbia University. Email: ccanonne@cs.columbia.edu. Research supported by NSF grants CCF-1115703 and NSF CCF-1319788.

1 Introduction

Property testing, as introduced in the seminal works of [RS96, GGR98], is the analysis and study of ultra-efficient and randomized decision algorithms, which must answer a promise problem yet cannot afford to query their whole input. A very successful and prolific area of theoretical computer science, property testing also gave rise to several subfields, notably that of *distribution* testing, where the input consists of independent *samples* from a probability distribution, and one must now verify if the underlying unknown distribution satisfies a given property of interest (cf. [Ron08, Ron09, Rub12, Can15, Gol17] for surveys on property and distribution testing).

One of the earliest and most studied questions in distribution testing is that of *uniformity testing*, where, given independent samples from an arbitrary probability distribution $\mathbf{p}$ on a discrete domain Ω, one has to decide whether (i) $\mathbf{p}$ is uniform on Ω, or (ii) $\mathbf{p}$ is "far" (i.e., at total variation distance at least ε) from the uniform distribution on Ω. Arguably the most natural distribution testing problem, testing uniformity is also one of the most fundamental; algorithms for uniformity testing end up being crucial building blocks in many other distribution testing algorithms [BFF$^+$01, DK16, Gol16]. Fully understanding the sample complexity of the problem, as well as the possible trade-offs it entails, thus prompted a significant line of research.

Starting with the work of Goldreich and Ron [GR00] (which considered it in the context of testing expansion of graphs), uniformity testing was studied and analyzed in a series of work [BFF$^+$01, Pan08, VV14, DKN15, ADK15, DGPP16], which culminated with the tight sample complexity bound of $\Theta(\sqrt{n}/\varepsilon^2)$ for testing uniformity on a discrete domain of size n. (Moreover, the corresponding algorithms are also efficient, running in time linear in the number of samples they take.)

Given this state of affairs, testing uniformity of discrete distributions appears to be fully settled; however, as often is the case, the devil is in the detail. Specifically, all the aforementioned results address the case where the domain Ω is explicitly known, and the task is to find out whether $\mathbf{p}$ is the uniform distribution *on this domain*. Yet, in many cases, samples (or data points) are drawn from the underlying distribution without such prior knowledge, and the relevant question is whether $\mathbf{p}$ is uniform on its *support* – which is unknown, of arbitrary size, and can be completely unstructured.[1]

In this work, we focus on this latter question: in particular, we do not assume any *a priori* knowledge on the domain Ω, besides its being discrete. Our goal is then the following: given independent samples from an arbitrary probability distribution $\mathbf{p}$ on Ω, we must distinguish between the case (i) $\mathbf{p}$ is uniform on *some subset of* Ω, and (ii) $\mathbf{p}$ is far from *every* such uniform distribution. As we shall see, this is not merely a technicality: this new task is provably harder than the case where Ω is known. Indeed, this difference intuitively stems from the uncertainty on where the support of $\mathbf{p}$ lies, which prevents any reduction to the simple, known-domain case.

Furthermore, one crucial feature of the problem is that it intrinsically calls for *adaptive* algorithms. This is in sharp contrast to the overwhelming majority of distribution testing algorithms, which (essentially) draw a prespecified number of samples all at once, before processing them and outputting a verdict. This is because, in our case, an algorithm is provided only with the proximity parameter $\varepsilon \in (0,1]$, and has no upper bound on the domain size n nor on any other parameter of the problem. Therefore, it must keep on taking samples until it has "extracted" enough information – and is confident enough that it can stop and output an answer. (In this sense, our setting is closer in spirit to the line of work pioneered in Statistics by Ingster [Ing00, FL06] than to the "instance-optimal"

[1] In particular, one cannot without loss of generality assume that the support is the set of consecutive integers $\{1, \ldots, n\}$.

setting of Valiant and Valiant [VV14, BCG16], as in the latter the algorithm is still provided with a massive parameter in the form of the full description of a reference probability distribution.)

1.1 Our Results

Given a discrete, possibly unbounded domain Ω, we let $\mathcal{C}_U$ denote the set of all probability distributions that are supported and uniform on some subset of Ω, that is

$$\mathcal{C}_U \stackrel{\text{def}}{=} \{\, \mathbf{u}_S \;:\; S \subseteq \Omega \,\}$$

where, for a given set $S \subseteq \Omega$, $\mathbf{u}_S$ denote the uniform distribution on S. In what follows, we write $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})$ for the total variation distance between two distributions $\mathbf{p}, \mathbf{q}$ on Ω.

Theorem 1.1. *There exists an algorithm which, given sample access to an arbitrary distribution* $\mathbf{p}$ *over some unknown discrete domain* Ω*, as well as parameter* $\varepsilon \in (0, 1]$*, satisfies the following.*

1. *If* $\mathbf{p} \in \mathcal{C}_U$*, then the algorithm outputs* accept *with probability at least* $2/3$*; while*
2. *if* $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{C}_U) > \varepsilon$*, then the algorithm outputs* reject *with probability at least* $2/3$*.*

Moreover, the algorithm takes $O\Big(\frac{1}{\varepsilon^6 \|\mathbf{p}\|_3}\Big)$ *samples in expectation, and is efficient (in the number of samples taken).*

We note that if indeed $\mathbf{p}$ is uniform, i.e., $\mathbf{p} = \mathbf{u}_S$ for some $S \subseteq \Omega$, then, for constant ε, the above complexity becomes $O(\,|S|^{2/3}\,)$ – to be compared to the $\Theta\Big(\sqrt{|S|}\Big)$ sample complexity of testing whether $\mathbf{p} = \mathbf{u}_S$ for a fixed S. Our next result shows that this is not an artifact of our algorithm; namely, such a dependence is necessary, and testing the *class* of uniform distributions is strictly harder than testing any specific uniform distribution.

Theorem 1.2. *Fix any (non-uniform) distribution* $\mathbf{q}$ *over* Ω*, and let* $\varepsilon \stackrel{\text{def}}{=} \mathrm{d}_{\mathrm{TV}}(\mathbf{q}, \mathcal{C}_U)$ *be its distance to* $\mathcal{C}_U$*. Then, given sample access to a distribution* $\mathbf{p}$ *on* Ω*, distinguishing with high constant probability between (i)* $\mathbf{p}$ *is equal to* $\mathbf{q}$ *up to a permutation of the domain and (ii)* $\mathbf{p} \in \mathcal{C}_U$*, requires* $\Omega\Big(\frac{1}{\|\mathbf{q}\|_3}\Big)$ *samples. In particular, an algorithm that tests membership in* $\mathcal{C}_U$ *with high probability and for any proximity parameter* $\varepsilon' \leq \varepsilon$ *requires this many samples.*

It is worth discussing the above statement in detail, as its interpretation can be slightly confusing. Specifically, it does *not* state that testing identity to any fixed, known distribution $\mathbf{p}$ requires $\Omega(1/\|\mathbf{p}\|_3)$ (indeed, by the results of [VV14, BCG16], such a statement would be false). What is stated is essentially that, even given the full description of $\mathbf{p}$, it is hard to distinguish between $\mathbf{p}$ and a uniform distribution, *after relabeling of the elements of the domain.* Since the class of uniform distributions is invariant by such permutations, the last part of the theorem follows.

1.2 Overview and Techniques

The key intuition and driving idea of both our upper and lower bounds is the observation that, by very definition of the problem, there is no structure nor ordering of the domain to leverage. That is, the class of uniform distributions over Ω is a "symmetric property" (broadly speaking, the actual labeling of the elements of the domain is irrelevant), and the domain itself can and should be thought of as a set of arbitrary points with no algebraic structure. Given this state of affairs, an algorithm should not be able to do much more than counting *collisions*, that is the number of pairs, or triples, or more generally k-tuples of samples which happen to "hit" the same domain element.

Equivalently, these collision counts correspond to the *moments* (that is, ℓ_p-norms) of the distribution; following a line of works on symmetric properties of distributions ([GR00, RRSS09, Val11, VV11], to cite a few), we thus need to, and can only, focus on estimating these moments. To relate this to our property $\mathcal{C}_U$, we first need a simple connection between ℓ_p norms and uniformity of a distribution. However, while getting an exact characterization is not difficult (Lemma 2.2), we are interested in a *robust* characterization, in order to derive a correspondence between approximate equality between ℓ_p norms and distance to uniformity. This is what we obtain in Lemma 3.4: roughly speaking, if $\|\mathbf{p}\|_2^4 \approx \|\mathbf{p}\|_3^3$ then $\mathbf{p}$ must be close to a uniform distribution on $1/\|\mathbf{p}\|_2^2$ elements.

This in turn allows us to design and analyze a simple and clean testing algorithm, which works in two stages: (i) estimate $\|\mathbf{p}\|_2^2$ to sufficient accuracy; (ii) using this estimate, take enough samples to estimate $\|\mathbf{p}\|_3^3$ as well; and $\textsf{accept}$ if and only if $\|\mathbf{p}\|_2^4 \approx \|\mathbf{p}\|_3^3$.

Turning to the lower bound, the idea is once again to only use the available information: namely, if all that *should* matter are the ℓ_p-norms of the distribution, then two distributions with similar low-order norms *should* be hard to distinguish; so it would suffice to come up with a pair of uniform and far-from-uniform distributions $\mathbf{p}^{\textsf{yes}}, \mathbf{p}^{\textsf{no}}$ with similar moments to establish our lower bound. Fortunately, this intuition – already present in [RRSS09] – was formalized and developed in an earlier work of Paul Valiant [Val11], which we thus can leverage for our purpose. Given this "Wishful Thinking Theorem" (see Theorem 2.1), what remains is to upper bound the discrepancy of the moments of our two candidate distributions $\mathbf{p}^{\textsf{yes}}, \mathbf{p}^{\textsf{no}}$ to show that some specific quantity is very small. Luckily, this last step also can be derived from the aforementioned robust characterization, Lemma 3.4.

1.3 Organization

After recalling some useful notation and results in Section 2, we establish our upper bound (Theorem 1.1) in Section 3. Section 4 is then dedicated to the proof of our lower bound, Theorem 1.2.

2 Preliminaries

2.1 Definitions and notation

All throughout this paper, we write $\Delta(\Omega)$ for the set of discrete probability distributions over domain Ω, i.e. the set of all real-valued functions $\mathbf{p}\colon \Omega \to [0,1]$ such that $\sum_{x\in\Omega} \mathbf{p}(x) = 1$. Considering a probability distribution as the vector of its probability mass function (pmf), we write $\|\mathbf{p}\|_r$ for its ℓ_r-norm, for any $r \in [1,\infty]$. A *property* of distributions over Ω is then a subset $\mathcal{P} \subseteq \Delta(\Omega)$, comprising all distributions that have the property.

As standard in distribution testing, we will measure the distance between two distributions $\mathbf{p}_1, \mathbf{p}_2$ on Ω by their *total variation distance*

$$\mathrm{d_{TV}}(\mathbf{p}_1, \mathbf{p}_2) \overset{\text{def}}{=} \frac{1}{2}\|\mathbf{p}_1 - \mathbf{p}_2\|_1 = \max_{S\subseteq\Omega}(\mathbf{p}_1(S) - \mathbf{p}_2(S))$$

which takes value in $[0,1]$. (This metric is sometimes referred to as *statistical distance*). Given a property $\mathcal{P}$ and a distribution $\mathbf{p} \subseteq \Delta(\Omega)$, we then write $\mathrm{d_{TV}}(\mathbf{p}, \mathcal{P}) \overset{\text{def}}{=} \inf_{\mathbf{q}\in\mathcal{P}} \mathrm{d_{TV}}(\mathbf{p}, \mathbf{q})$ for the distance of $\mathbf{p}$ to $\mathcal{P}$.

Finally, recall that a *testing algorithm* for a fixed property $\mathcal{P}$ is a randomized algorithm $\mathcal{T}$ which takes as input a proximity parameter $\varepsilon \in (0,1]$, and is granted access to independent samples from an unknown distribution $\mathbf{p}$:

1. if $\mathbf{p} \in \mathcal{P}$, the algorithm outputs $\textsf{accept}$ with probability at least $2/3$;
2. if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{p}') \geq \varepsilon$ for every $\mathbf{p}' \in \mathcal{P}$, it outputs $\textsf{reject}$ with probability at least $2/3$.

That is, $\mathcal{T}$ must accept if the unknown distribution has the property, and reject if it is ε-*far* from having it. The *sample complexity* of the algorithm is the number of samples it draws from the distribution in the worst case.

2.2 Useful results from previous work

We will heavily rely, for our lower bound, on the "Wishful Thinking Theorem" due to Paul Valiant [Val11], which applies to testing symmetric properties of distributions (that is, properties that are invariant under relabeling of the domain, as $\mathcal{C}_U$ happens to be). Intuitively, this theorem ensures that "if the low-degree moments (ℓ_p norms) of two distributions match, then these distributions (up to relabeling) are hard to distinguish."

Theorem 2.1 (Wishful Thinking Theorem [Val11, Theorem 4.10], restated)**.** *Given a positive integer k and two distributions $\mathbf{p}^{\textsf{yes}}, \mathbf{p}^{\textsf{no}}$, it is impossible to test in k samples any symmetric property that holds for $\mathbf{p}^{\textsf{yes}}$ and does not hold for $\mathbf{p}^{\textsf{no}}$, provided that following conditions hold:*

- $\|\mathbf{p}^{\textsf{yes}}\|_\infty, \|\mathbf{p}^{\textsf{no}}\|_\infty \leq \frac{1}{500k}$*;*
- *letting $m^{\textsf{yes}}$, $m^{\textsf{no}}$ be the k-based moments of $\mathbf{p}^{\textsf{yes}}, \mathbf{p}^{\textsf{no}}$ (defined below),*

$$\sum_{j=2}^{\infty} \frac{|m^{\textsf{yes}}(j) - m^{\textsf{no}}(j)|}{\sqrt{1+\max(m^{\textsf{yes}}(j), m^{\textsf{no}}(j))}} < \frac{1}{24},$$

where $m^{\textsf{yes}}(j) \stackrel{\text{def}}{=} k^j \|\mathbf{p}^{\textsf{yes}}\|_j^j$, $m^{\textsf{no}}(j) \stackrel{\text{def}}{=} k^j \|\mathbf{p}^{\textsf{no}}\|_j^j$ *for* $j \geq 0$.

(We observe that we only reproduced here one of the three sufficient conditions given in the original, more general theorem; as this will be the only one we need.)

2.3 Some structural results

We here state and establish some simple yet useful results. The first relates uniformity of a distribution to the ℓ_p-norms of its probability mass function, while the second provides inequalities between these norms.

Lemma 2.2. *Let $\mathbf{p} \in \Delta(\Omega)$. Then, $\|\mathbf{p}\|_2^4 = \|\mathbf{p}\|_3^3$ if and only if $\mathbf{p} \in \mathcal{C}_U$.*

Proof. If $\mathbf{p} \in \mathcal{C}_U$, it is immediate to see that $\|\mathbf{p}\|_2^4 = \|\mathbf{p}\|_3^3$. We thus consider the converse implication. By the Cauchy–Schwarz inequality,

$$\|\mathbf{p}\|_2^2 = \sum_{i\in\Omega} \mathbf{p}_i^2 \leq \left(\sum_{i\in\Omega}\left(\mathbf{p}_i^{3/2}\right)^2\right)^{1/2} \left(\sum_{i\in\Omega}\left(\mathbf{p}_i^{1/2}\right)^2\right)^{1/2} = \left(\sum_{i\in\Omega}\mathbf{p}_i^3\right)^{1/2}\left(\sum_{i\in\Omega}\mathbf{p}_i\right)^{1/2} = \|\mathbf{p}\|_3^{3/2}\cdot 1$$

with equality if, and only if, $(\mathbf{p}_i^{3/2})_{i\in\Omega}$ and $(\mathbf{p}_i^{1/2})_{i\in\Omega}$ are linearly dependent. Thus, $\|\mathbf{p}\|_2^4 = \|\mathbf{p}\|_3^3$ implies that there exist non-zero $\alpha, \beta \in \mathbb{R}$ such that $\alpha\mathbf{p}_i^{3/2} = \beta\mathbf{p}_i^{1/2}$ for all $i \in \Omega$, or equivalently that $\mathbf{p}_i \in \{0, \frac{\beta}{\alpha}\}$ for all $i \in \Omega$. This, in turn, implies that $\mathbf{p}$ is uniform on a subset of $\frac{\alpha}{\beta}$ elements. □

Fact 2.3. *For any vector $x \in \mathbb{R}^{\mathbb{N}}$ such that $\|x\|_1 < \infty$, we have*

$$\|x\|_2^{2(j-1)} \leq \|x\|_1^{j-2} \|x\|_j^j,$$

for all $j \geq 2$. In particular, for any distribution $\mathbf{p} \in \Delta(\Omega)$, we have $\|\mathbf{p}\|_2^{2(j-1)} \leq \|\mathbf{p}\|_j^j$ for all $j \geq 2$ (and, thus, for instance, $\|\mathbf{p}\|_2^4 \leq \|\mathbf{p}\|_3^3$).

Proof. The inequality is trivially true for $j = 2$, and, so, we henceforth assume $j \geq 3$. Let $x \in \mathbb{R}^{\mathbb{N}}$ be such a vector: we wish to show that $(\sum_{i=0}^{\infty} x_i^2)^{j-1} \leq (\sum_{i=0}^{\infty} |x_i|)^{j-2} \left(\sum_{i=0}^{\infty} |x_i|^j\right)$, or equivalently $\sum_{i=0}^{\infty} x_i^2 \leq (\sum_{i=0}^{\infty} |x_i|)^{\frac{j-2}{j-1}} \left(\sum_{i=0}^{\infty} |x_i|^j\right)^{\frac{1}{j-1}}$. Set $p' \overset{\text{def}}{=} \frac{j-1}{j-2}$, and $q' \overset{\text{def}}{=} j-1$ so that $p', q' \geq 1$ with $\frac{1}{p'} + \frac{1}{q'} = 1$. Observing that $|x_i|^2 = |x_i|^{\frac{j-2}{j-1}} |x_i|^{\frac{j}{j-1}}$, we then apply Hölder's inequality:

$$\begin{aligned}
\sum_{i=0}^{\infty} |x_i|^2 &= \sum_{i=0}^{\infty} |x_i|^{\frac{1}{p'}} |x_i|^{\frac{j}{q'}} \\
&\leq \left(\sum_{i=0}^{\infty} |x_i|^{\frac{p'}{p'}}\right)^{\frac{1}{p'}} \left(\sum_{i=0}^{\infty} |x_i|^{\frac{jq'}{q'}}\right)^{\frac{1}{q'}} \\
&= \left(\sum_{i=0}^{\infty} |x_i|\right)^{\frac{j-2}{j-1}} \left(\sum_{i=0}^{\infty} |x_i|^j\right)^{\frac{1}{j-1}}
\end{aligned}$$

concluding the proof. □

3 The Upper Bound

Our algorithm for testing uniformity first estimates the ℓ_2 norm of the input distribution and uses this estimate to obtain a surrogate value for the size of the support set for the distribution. In the case the input distribution is a uniform distribution, the ℓ_2 norm estimate indeed provides a good approximation to the size of the support set. Our algorithm for the ℓ_2 norm estimation is presented in the following section, followed by our algorithm for testing uniformity.

3.1 Estimating the ℓ_2 norm of a distribution

In this section, we present an algorithm that, given independent samples from a distribution $\mathbf{p}$ over $\mathbb{N}$, estimates $\|\mathbf{p}\|_2^2$. Note that a similar result was presented in Batu et al. [BFR+13] in the case when the size of the domain is bounded and known to the algorithm. Furthermore, an algorithm based on the same ideas have been presented by Batu et al. [BDKR05] to estimate the entropy of a distribution that is uniform on a subset of its domain. The algorithm is presented below in Algorithm 1.

Algorithm 1 Estimating the ℓ_2 norm of a distribution from samples

```
procedure Estimate-$\ell_2$-norm($\mathbf{p}, \varepsilon$)
    $k \leftarrow \lceil \frac{C}{\varepsilon^4} \rceil$                    ▷ $C = 6500$
    Keep taking samples from $\mathbf{p}$ until $k$ 2-collisions are observed.
    Let $m$ be the number of samples taken.
    return $\frac{k}{\binom{m}{2}}$
end procedure
```

Lemma 3.1. *Algorithm* Estimate-ℓ_2-norm*, given independent samples from a distribution* $\mathbf{p}$ *over* $\mathbb{N}$ *and* $0 < \varepsilon < \frac{1}{2}$*, outputs a value* γ *such that*

$$(1-\varepsilon) \cdot \|\mathbf{p}\|_2^2 \leq \gamma \leq (1+\varepsilon) \cdot \|\mathbf{p}\|_2^2, \tag{1}$$

with probability at least $3/4$*. Whenever the algorithm produces an estimate satisfying* (1) *above, the number of samples taken by the algorithm is* $\Theta(\frac{1}{\varepsilon^2 \|\mathbf{p}\|_2})$*. Moreover, the algorithm takes* $O(\frac{1}{\varepsilon^2 \|\mathbf{p}\|_2})$ *samples in expectation.*

Proof. Let M be the random variable that denotes the number of samples that were taken by the algorithm until k pairwise collisions are observed. We will show that, with constant probability, M is close to its expected value nearly $\sqrt{k}/\|\mathbf{p}\|_2$.

Consider a set of m samples from $\mathbf{p}$. For $1 \leq i < j \leq m$, let X_{ij} be an indicator random variable denoting a collision between ith and jth samples. Let $S_m = \sum_{1 \leq i < j \leq m} X_{ij}$ be the total number of collisions among the samples.

For any $i < j$, $\mathbb{E}[X_{ij}] = \|\mathbf{p}\|_2^2$. Therefore, $\mathbb{E}[S_m] = \binom{m}{2} \cdot \|\mathbf{p}\|_2^2$. We will also need an upper bound on the variance $\mathrm{Var}[S_m]$ to show that the k collisions are not observed too early or too late.

$$\mathbb{E}\left[S_m^2\right] = \mathbb{E}\left[\left(\sum_{i<j} X_{ij}\right)\left(\sum_{i'<j'} X_{i'j'}\right)\right] = \sum_{i<j, i'<j'} \mathbb{E}[X_{ij} X_{i'j'}].$$

The terms of the last summation above can be grouped according to the cardinality of the set $\{i, j, i', j'\}$.

- If $|\{i, j, i', j'\}| = 2$, then $\mathbb{E}[X_{ij} X_{i'j'}] = \mathbb{E}[X_{ij}] = \|\mathbf{p}\|_2^2$. There are $\binom{m}{2}$ such terms.
- If $|\{i, j, i', j'\}| = 3$, then $\mathbb{E}[X_{ij} X_{i'j'}] = \mathbb{E}[X_{ij} X_{ij'}] = \|\mathbf{p}\|_3^3$. There are $6\binom{m}{3}$ such terms.
- If $|\{i, j, i', j'\}| = 4$, then $\mathbb{E}[X_{ij} X_{i'j'}] = \mathbb{E}[X_{ij}]\mathbb{E}[X_{i'j'}] = \|\mathbf{p}\|_2^4$. There are $6\binom{m}{4}$ such terms.

Hence, we can bound the variance of S_m as follows.

$$\begin{aligned}
\mathrm{Var}[S_m] &= \mathbb{E}[S_m^2] - \mathbb{E}[S_m]^2 \\
&= \binom{m}{2} \cdot \|\mathbf{p}\|_2^2 + 6\binom{m}{3} \cdot \|\mathbf{p}\|_3^3 + 6\binom{m}{4} \cdot \|\mathbf{p}\|_2^4 - \left(\binom{m}{2} \cdot \|\mathbf{p}\|_2^2\right)^2 \\
&= \binom{m}{2} \cdot \|\mathbf{p}\|_2^2 + 2m \cdot \|\mathbf{p}\|_3^3 + (m^3 - 3m^2) \cdot (\|\mathbf{p}\|_3^3 - \|\mathbf{p}\|_2^4) \\
&\leq \binom{m}{2} \cdot \|\mathbf{p}\|_2^2 + m^3 \cdot \|\mathbf{p}\|_2^3,
\end{aligned}$$

where the inequality arises from $\|\mathbf{p}\|_3 \leq \|\mathbf{p}\|_2$.

The probability that the output of the algorithm is less than $(1-\varepsilon)\cdot\|\mathbf{p}\|_2^2$ (that is, an underestimation) is bounded from above by the probability of the random variable M taking a value m such that $(1-\varepsilon)\binom{m}{2}\cdot\|\mathbf{p}\|_2^2 > k$. Analogously, the probability of an overestimation is bounded above by the probability of the random variable M taking a value m such that $(1+\varepsilon)\binom{m}{2}\cdot\|\mathbf{p}\|_2^2 < k$.

Let m be the smallest integer such that $(1+\varepsilon)\binom{m}{2}\cdot\|\mathbf{p}\|_2^2 \geq k$, so that $(1+\varepsilon)\binom{m-1}{2}\cdot\|\mathbf{p}\|_2^2 < k$. Then,

$$\begin{aligned}
\Pr\left[\text{overestimation}\right] &= \Pr[M < m] = \Pr[\exists \ell \leq m-1,\ S_\ell \geq k] = \Pr\left[S_{m-1} \geq k\right] \\
&= \Pr\left[S_{m-1} - \mathbb{E}[S_{m-1}] \geq k - \binom{m-1}{2}\cdot\|\mathbf{p}\|_2^2\right] \\
&\leq \Pr\left[|S_{m-1} - \mathbb{E}[S_{m-1}]| > \varepsilon\cdot\binom{m-1}{2}\cdot\|\mathbf{p}\|_2^2\right] \\
&\leq \frac{\mathrm{Var}[S_{m-1}]}{\left(\varepsilon\cdot\binom{m-1}{2}\cdot\|\mathbf{p}\|_2^2\right)^2} \qquad \text{(Chebyshev's inequality)} \\
&\leq \frac{\binom{m-1}{2}\cdot\|\mathbf{p}\|_2^2 + (m-1)^3\cdot\|\mathbf{p}\|_2^3}{\varepsilon^2\cdot\binom{m-1}{2}^2\cdot\|\mathbf{p}\|_2^4} \\
&\leq \frac{1}{\varepsilon^2}\left(\frac{1}{\binom{m-1}{2}\cdot\|\mathbf{p}\|_2^2} + \frac{9}{(m-1)\cdot\|\mathbf{p}\|_2}\right) \\
&= \frac{1}{\varepsilon^2}\left(\frac{m}{m-2}\frac{1}{\binom{m}{2}\cdot\|\mathbf{p}\|_2^2} + \frac{m}{m-1}\frac{9}{m\cdot\|\mathbf{p}\|_2}\right) \\
&\leq \frac{1}{\varepsilon^2}\left(\frac{10}{8}\frac{1}{\binom{m}{2}\cdot\|\mathbf{p}\|_2^2} + \frac{10}{9}\frac{9}{m\cdot\|\mathbf{p}\|_2}\right) \\
&\qquad (m \geq \sqrt{2k}+1 \geq 10, \text{ or } \Pr[S_{m-1} \geq k] = 0.) \\
&\underset{(*)}{\leq} \frac{1}{\varepsilon^2}\left(\frac{10}{8}\cdot\frac{1+\varepsilon}{k} + \frac{10\sqrt{1+\varepsilon}}{\sqrt{2k}}\right) \leq \frac{10}{\varepsilon^2}\left(\frac{1}{4k} + \frac{1}{\sqrt{k}}\right) \\
&\leq \frac{5\varepsilon^2}{2C} + \frac{10}{\sqrt{C}} \leq \frac{5}{2C} + \frac{10}{\sqrt{C}} \\
&< \frac{1}{8}
\end{aligned}$$

for $C \geq 6500$, where $(*)$ follows from the choice of m.

To upper bound the probability of underestimation, take m to be largest integer such that

[2] In particular, this implies $\binom{m+1}{2} > k$, from which $m > \sqrt{2k}+1 \gg \frac{1}{\varepsilon^2}$.

$(1-\varepsilon)\binom{m}{2}\cdot\|\mathbf{p}\|_2^2 \le k$ (so that $(1-\varepsilon)\binom{m+1}{2}\cdot\|\mathbf{p}\|_2^2 > k$, i.e. $(1-\varepsilon)\frac{m+1}{m-1}\binom{m}{2}\cdot\|\mathbf{p}\|_2^2 > k$).[2] Then,

$$\begin{aligned}
\Pr\left[\text{underestimation}\right] &= \Pr[M > m] = \Pr[\forall \ell \le m,\ S_\ell < k] = \Pr\left[S_m < k\right] \\
&= \Pr\left[\mathbb{E}[S_m] - S_m > \mathbb{E}[S_m] - k\right] \\
&\le \Pr\left[\mathbb{E}[S_m] - S_m > \left(1-(1-\varepsilon)\frac{m+1}{m-1}\right)\cdot\binom{m}{2}\cdot\|\mathbf{p}\|_2^2\right] \\
&= \Pr\left[\mathbb{E}[S_m] - S_m > \frac{\varepsilon m-1}{m-1}\cdot\binom{m}{2}\cdot\|\mathbf{p}\|_2^2\right] && \text{(Note that } \varepsilon m > 1\text{)} \\
&\le \left(\frac{m-1}{\varepsilon m-1}\right)^2 \frac{\operatorname{Var}[S_m]}{\left(\binom{m}{2}\cdot\|\mathbf{p}\|_2^2\right)^2} \le \left(\frac{2}{\varepsilon}\right)^2 \frac{\operatorname{Var}[S_m]}{\left(\binom{m}{2}\cdot\|\mathbf{p}\|_2^2\right)^2} \\
&\le 4\frac{\binom{m}{2}\cdot\|\mathbf{p}\|_2^2 + m^3\cdot\|\mathbf{p}\|_2^3}{\varepsilon^2\cdot\binom{m}{2}^2\cdot\|\mathbf{p}\|_2^4} \le \frac{4}{\varepsilon^2}\left(\frac{1}{\binom{m}{2}\cdot\|\mathbf{p}\|_2^2} + \frac{9}{m\cdot\|\mathbf{p}\|_2}\right) \\
&\le \frac{4}{\varepsilon^2}\left(\frac{12}{10}\frac{1}{\binom{m}{2}\cdot\|\mathbf{p}\|_2^2} + \frac{11}{10}\frac{9}{m\cdot\|\mathbf{p}\|_2}\right) && (m \ge \sqrt{2k}+1 \ge 10.) \\
&\underset{(*)}{\le} \frac{6}{\varepsilon^2}\left(\frac{1-\varepsilon}{k} + \frac{\sqrt{1-\varepsilon}}{\sqrt{2k}}\right) \le \frac{6}{\varepsilon^2}\left(\frac{1}{k}+\frac{1}{\sqrt{2k}}\right) \\
&\le \frac{6\varepsilon^2}{C} + \frac{6}{\sqrt{2C}} \le \frac{6}{C} + \frac{6}{\sqrt{2C}} \\
&< \frac{1}{8}
\end{aligned}$$

for $C \ge 1250$, where $(*)$ follows from the choice of m.

By the union bound, overestimation or underestimation happens with probability at most 1/4. Finally, in the event that we have a good estimation, we have that the number m of samples satisfy

$$\frac{k}{(1+\varepsilon)\cdot\|\mathbf{p}\|_2^2} \le \binom{m}{2} \le \frac{k}{(1-\varepsilon)\cdot\|\mathbf{p}\|_2^2}.$$

Therefore, we have that $m = \Theta(\sqrt{k}/\|\mathbf{p}\|_2) = \Theta(1/(\varepsilon^2\cdot\|\mathbf{p}\|_2))$.

To bound the expected number of samples, we consider two cases (recall that the asymptotics here are taken, unless specified otherwise, while viewing $\mathbf{p}$ as a sequence of distributions $(\mathbf{p}^{(n)})_{n\ge 0}$ and letting $n \to \infty$):

- if $\|\mathbf{p}\|_\infty = \Omega(\|\mathbf{p}\|_2)$ (i.e., $\|\mathbf{p}\|_\infty = \Theta(\|\mathbf{p}\|_2)$), then we denote by i_∞ the element such that $\mathbf{p}_{i_\infty} = \|\mathbf{p}\|_\infty$. It follows from properties of the negative binomial distribution that the expected number M_∞ of draws necessary to see $\ell = \Theta(\sqrt{k})$ different draws of i_∞ (and thus $k = \binom{\ell}{2}$ collisions) is $\Theta(\sqrt{k}/\|\mathbf{p}\|_\infty)$, so that $\mathbb{E}[M] \le \mathbb{E}[M_\infty] = O(\frac{1}{\varepsilon^2\|\mathbf{p}\|_\infty})$.
- on the other hand, if $\|\mathbf{p}\|_\infty = o(\|\mathbf{p}\|_2)$, then we can apply Theorem 4 of [CP00] (see also [eh17]) to get that $\mathbb{E}[M] \sim_{n\to\infty} \frac{C_k}{\|\mathbf{p}\|_2}$, where $C_k = \binom{k-\frac{1}{2}}{k-1}\sqrt{\frac{\pi}{2}} \sim_{k\to\infty} \sqrt{2k}$. Recalling that $k = \Theta(1/\varepsilon^4)$, we obtain $\mathbb{E}[M] = \Theta(\frac{1}{\varepsilon^2\|\mathbf{p}\|_2})$, as claimed.

□

Note that the sample complexity of Algorithm Estimate-ℓ_2-norm is tight for near-uniform distributions (at least, in terms of dependency on $\|\mathbf{p}\|_2$). Consider a distribution $\mathbf{p}$ on n elements with probability values in $\{(1-\delta)/n, (1+\delta)/n\}$ for some small δ. Even though $\|\mathbf{p}\|_2$ can have sufficiently high $\|\mathbf{p}\|_2$ and should be distinguished from the uniform distribution on n elements, there will be no repetition in the sample until $\Omega(\sqrt{n}) = \Omega(1/\|\mathbf{p}\|_2)$ samples are taken. The following lemma generalizes this argument.

Lemma 3.2. *For any distribution $\mathbf{p}$ and $\varepsilon \in (0, 1/3)$, estimation of $\|\mathbf{p}\|_2^2$ within a multiplicative factor of $(1+\varepsilon)$ requires $\Omega(1/(\sqrt{\varepsilon}\|\mathbf{p}\|_2))$ samples from $\mathbf{p}$.*

Proof. Take any distribution $\mathbf{p}$. We first consider the case $\varepsilon \geq \|\mathbf{p}\|_2^2$. Fix any element $c \in \mathbb{N}$ such that $\mathbf{p}(c) = 0$ (we can assume for simplicity one exists; otherwise, since we can find, for any $\eta > 0$, $c \in \mathbb{N}$ such that $\mathbf{p}(c) < \eta$, we can repeat the argument below for an arbitrarily small η), and let $\gamma \overset{\text{def}}{=} \frac{\|\mathbf{p}\|_2+\sqrt{3\varepsilon+(1+3\varepsilon)\|\mathbf{p}\|_2^2}}{1+\|\mathbf{p}\|_2^2}$. Then, we define the distribution $\mathbf{q}$ on $\mathbb{N}$ as the mixture

$$\mathbf{q} \overset{\text{def}}{=} (1-\gamma\|\mathbf{p}\|_2)\mathbf{p} + \gamma\|\mathbf{p}\|_2 \mathbf{1}_{\{c\}}$$

which satisfies $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = \gamma\|\mathbf{p}\|_2$, and

$$\|\mathbf{q}\|_2^2 = (1-\gamma\|\mathbf{p}\|_2)^2\|\mathbf{p}\|_2^2 + \gamma^2\|\mathbf{p}\|_2^2 = ((1-\gamma\|\mathbf{p}\|_2)^2 + \gamma^2)\|\mathbf{p}\|_2^2 = (1+3\varepsilon)\|\mathbf{p}\|_2^2$$

the last equality from our choice of γ. Since $\varepsilon < 1$, any algorithm that estimates the squared ℓ_2 norm of an unknown distribution can be used to distinguish between $\mathbf{p}$ and $\mathbf{q}$. However, from the very definition of total variation distance, distinguishing between $\mathbf{p}$ and $\mathbf{q}$ requires $\Omega(1/\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}))$ samples. Since

$$\gamma \leq \|\mathbf{p}\|_2 + \sqrt{3\varepsilon + 2\|\mathbf{p}\|_2^2} \leq (1+\sqrt{5})\sqrt{\varepsilon}$$

(as $\|\mathbf{p}\|_2^2 \leq \varepsilon$) we get a lower bound of $\Omega\Big(\frac{1}{\sqrt{\varepsilon}\|\mathbf{p}\|_2}\Big)$.

We now turn to the case $\varepsilon < \|\mathbf{p}\|_2^2$. The construction will be similar, but setting $\gamma \overset{\text{def}}{=} 3\varepsilon/\|\mathbf{p}\|_2$, and spreading the $\gamma\|\mathbf{p}\|_2 = 3\varepsilon$ probability uniformly on $m \overset{\text{def}}{=} \frac{3\varepsilon}{(1-3\varepsilon)\|\mathbf{p}\|_2^2}$ elements $c_1, \dots, c_m$ outside the support of $\mathbf{p}$, instead of just one. It is straightforward to check that in this case, the distribution $\mathbf{q}$ we defined is such that

$$\|\mathbf{q}\|_2^2 = (1-3\varepsilon)^2\|\mathbf{p}\|_2^2 + \frac{9\varepsilon^2}{m} = (1-3\varepsilon)\|\mathbf{p}\|_2^2$$

so again, by the same argument, any algorithm which can approximate $\|\mathbf{p}\|_2^2$ to $1+\varepsilon$ can be used to distinguish between $\mathbf{p}$ and $\mathbf{q}$, and thus requires $\Omega\Big(\frac{1}{\gamma\|\mathbf{p}\|_2}\Big) = \Omega\Big(\frac{1}{\sqrt{\varepsilon}\|\mathbf{p}\|_2}\Big)$ samples. □

Remark 3.3. We emphasize that the above theorem is on an instance-by-instance basis, and applies to *every* probability distribution $\mathbf{p}$. In contrast, it is not hard to see that for *some* distributions $\mathbf{p}$, a lower bound of $\Omega(1/(\|\mathbf{p}\|_2\varepsilon^2))$ holds: this follows from instance from [AOST17, Theorem 15]. This latter bound, however, cannot hold for every probability distribution, as one can see e.g. from a (trivial) distribution $\mathbf{p}$ supported on a single element, for which ℓ_2-norm estimation can be done with $O(1/\varepsilon) = O(1/(\|\mathbf{p}\|_2\varepsilon)$ samples.

3.2 Testing Uniformity

In this section, we present our algorithm for testing uniformity of a distribution. We first give a brief overview of the algorithm. The algorithm first estimates the ℓ_2 norm of the input distribution and uses this value to obtain an estimate on the support size of the distribution. Then, the algorithm tries to distinguish a uniform distribution from a distribution that is far from any uniform distribution by using the number of 3-way collisions in a freshly taken sample set. For two distributions with the same ℓ_2 norm, where one is a uniform distribution and the other is far from being uniform, the latter is expected to produce more 3-way collisions in a large enough sample set. The algorithm keeps taking samples up to a number based on the support-size estimate and keeps track of the 3-way collisions in the sample set to decide whether to accept or reject the input distribution.

The following lemma formalizes the intuition that if the ℓ_2 and the ℓ_3 norm of a distribution is close to those of the uniform distribution on N elements, then the distribution is close to being uniform.

Lemma 3.4. *Let* $\mathbf{p}$ *be a distribution over* $\mathbb{N}$ *and* $N \in \mathbb{N}$ *such that*

$$\frac{1-\varepsilon}{N} \leq \|\mathbf{p}\|_2^2 \leq \frac{1+\varepsilon}{N}$$

and

$$\|\mathbf{p}\|_3^3 \leq \frac{1+\delta}{N^2},$$

for some $0 < \varepsilon, \delta < 0.04$. *Then, the distance of* $\mathbf{p}$ *to* $\mathcal{C}_U$ *can be upper bounded as*

$$\mathrm{d_{TV}}(\mathbf{p}, \mathcal{C}_U) \leq 9\sqrt[3]{\delta + 3\varepsilon}.$$

Proof. Note that the condition on the $\|\mathbf{p}\|_2^2$ implies that $\mathbf{p}$ "ought to be" distributed roughly uniformly over N elements, or otherwise would deviate significantly enough from uniformity to impact its ℓ_3 norm. The condition on $\|\mathbf{p}\|_3^3$ further strengthens how evenly $\mathbf{p}$ is distributed, ensuring that this latter case cannot happen. Below we formalize this intuition and, in particular, use the conditions on the norms to upper bound the total mass on the items that have probability significantly larger than $1/N$.

Let R be a random variable such that R takes value p_i with probability p_i, for each element i in the support set of $\mathbf{p}$. Then, $\mathbb{E}[R] = \sum_{i \in \mathbb{N}} p_i^2 = \|\mathbf{p}\|_2^2$, which implies

$$\frac{1-\varepsilon}{N} \leq \mathbb{E}[R] \leq \frac{1+\varepsilon}{N}$$

and

$$\begin{aligned}
\mathrm{Var}[R] &= \mathbb{E}[R^2] - \mathbb{E}[R]^2 \\
&= \sum_{i \in \mathbb{N}} p_i^3 - \|\mathbf{p}\|_2^4 \\
&\leq \frac{1+\delta}{N^2} - \frac{(1-\varepsilon)^2}{N^2} \\
&\leq \frac{\delta + 2\varepsilon}{N^2}.
\end{aligned}$$

We now derive an upper bound on the ℓ_1 distance $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{C}_U)$. We first obtain an upper bound on the total weight of elements with probability significantly above or below $\frac{1}{N}$. Then, we can proceed to compare the distribution $\mathbf{p}$ to a uniform distribution with support size close to N.

First, we can bound the total probability mass of items i such that $p_i > \frac{1+\sqrt[3]{\delta+3\varepsilon}}{N}$ or $p_i < \frac{1-\sqrt[3]{\delta+3\varepsilon}}{N}$ by looking at the probability of a large deviation of R from its expectation. In particular,

$$\begin{aligned}\Pr\left[\left(R > \frac{1+\sqrt[3]{\delta+3\varepsilon}}{N}\right) \vee \left(R < \frac{1-\sqrt[3]{\delta+3\varepsilon}}{N}\right)\right] &\leq \Pr\left[|R - \mathbb{E}[R]| > \frac{\sqrt[3]{\delta+3\varepsilon}-\epsilon}{N}\right] \\ &\leq \Pr\left[|R - \mathbb{E}[R]| > \frac{\sqrt[3]{\delta+2\varepsilon}}{N}\right] \\ &\leq \frac{\mathrm{Var}[R]\cdot N^2}{\sqrt[3]{(\delta+2\varepsilon)^2}} \\ &\leq \sqrt[3]{\delta+2\varepsilon}\end{aligned}$$

Note that the second inequality above follows from that $\sqrt[3]{\delta+3\varepsilon} - \epsilon \geq \sqrt[3]{\delta+2\varepsilon}$ when $\delta + 2\epsilon \leq 3^{-3/2} \leq 0.18$, by the concavity of the function $f(x) = \sqrt[3]{x}$ and $f'(x) \geq 1$ for $x \leq 3^{-3/2}$.

We now have established that a probability mass of at least $1 - \sqrt[3]{\delta+2\varepsilon}$ of $\mathbf{p}$ is placed on elements with individual probabilities in the interval $[\frac{1-\sqrt[3]{\delta+3\varepsilon}}{N}, \frac{1+\sqrt[3]{\delta+3\varepsilon}}{N}]$. Call this set F. Thus, we have that

$$\frac{(1-\sqrt[3]{\delta+2\varepsilon})N}{1+\sqrt[3]{\delta+3\varepsilon}} \leq |F| \leq \frac{N}{1-\sqrt[3]{\delta+3\varepsilon}}.$$

Now consider the uniform distribution $\mathbf{u}_F$ on the set F. Since $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathcal{C}_U) \leq \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_F)$, it suffices to upper bound the latter. Given that

$$1 - \sqrt[3]{\delta+3\varepsilon} < 1 - \sqrt[3]{\delta+2\varepsilon} < 1 + \sqrt[3]{\delta+2\varepsilon} < \frac{1-\sqrt[3]{\delta+3\varepsilon}}{1-\sqrt[3]{\delta+2\varepsilon}},$$

for any $i \in F$, we have that

$$|p_i - \frac{1}{|F|}| \leq \frac{4\sqrt[3]{\delta+3\varepsilon}}{N}.$$

Finally, we can conclude that

$$\begin{aligned}\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{u}_F) &= \mathbf{p}(\mathbb{N} \setminus F) + \sum_{i \in F} |p_i - \frac{1}{|F|}| \\ &\leq \sqrt[3]{\delta+2\varepsilon} + \sum_{i \in F} \frac{4\sqrt[3]{\delta+3\varepsilon}}{N} \\ &\leq \sqrt[3]{\delta+2\varepsilon} + \frac{4\sqrt[3]{\delta+3\varepsilon}}{1-\sqrt[3]{\delta+3\varepsilon}} \\ &\leq 9\sqrt[3]{\delta+3\varepsilon}\end{aligned}$$

establishing the lemma. □

The algorithm for testing uniformity is presented below in Algorithm 2.

Algorithm 2 Testing Uniformity

```
procedure Test-Uniformity($\mathbf{p}, \varepsilon$)
    $\delta \leftarrow \varepsilon^3/5832$
    $N \leftarrow 1/$Estimate-$\ell_2$-norm$(\mathbf{p}, \delta)$
    $k \leftarrow \lceil \varepsilon^{-18} \rceil$
    Keep taking samples from $\mathbf{p}$ until you see $k$ 3-way collisions or reach
    $M = \sqrt[3]{3(1-4\delta)k}N^{2/3}$ samples, whichever happens first.
    if more than $k$ 3-way collisions are observed in the sample set then
        return reject
    else
        return accept
    end if
end procedure
```

Note that, for a uniform distribution, ℓ_2 norm estimation will give a reliable estimate N for the support size. Then, we will show that $M = O(\varepsilon^{-6}N^{2/3})$ samples will be unlikely to produce more than k 3-way collision. On the other hand, for a distribution that is far from a uniform distribution, the support size estimation in the algorithm will be an underestimation. In additions, the ℓ_3 norm of such a distribution will be higher than that of the uniform distribution with that estimated support size. As a result, the algorithm will observe more than k 3-way collisions in the subsequent samples with high probability as an evidence that the input distribution is not uniform.

Theorem 3.5. *Algorithm* Test-Uniformity*, given independent samples from a distribution* $\mathbf{p}$ *over* $\mathbb{N}$ *and* $0 < \varepsilon < \frac{1}{2}$*, accepts if* $\mathbf{p} \in \mathcal{C}_U$ *and rejects* $\mathbf{p}$ *such that* $\Delta(\mathbf{p}, \mathcal{C}_U) \geq \varepsilon$*, with probability at least 3/4. The sample complexity of the algorithm is* $\Theta(1/\varepsilon^6 \|\mathbf{p}\|_3)$*.*

Proof. In the proof, we will need simple distributional properties of the number of 3-way collisions, analogous to the arguments in the proof of Lemma 3.1. Let T_m be the total number of 3-way collisions in m samples from a distribution $\mathbf{p}$. Then, we have that

$$\mathbb{E}[T_m] = \binom{m}{3} \cdot \|\mathbf{p}\|_3^3$$

and

$$\mathrm{Var}[T_m] \leq O\left(m^3\|\mathbf{p}\|_3^3 + m^4\|\mathbf{p}\|_3^4 + m^5\|\mathbf{p}\|_3^5\right).$$

For the completeness argument, take $\mathbf{p} = U_S$ for some subset S of $\mathbb{N}$. Then, by Lemma 3.1, variable N from the algorithm will be within $(1 \mp \delta)$ of $|S|$, with probability 3/4. Then, the probability that the number of 3-way collisions in $m = M$ samples from $\mathbf{p}$ is more than k is

$$\begin{aligned}
\Pr\left[T_m > k\right] &\leq \Pr\left[T_m - \mathbb{E}[T_m] > k - (1-4\delta)kN^2 \cdot \frac{1}{|S|^2}\right] \\
&\leq \Pr\left[T_m - \mathbb{E}[T_m] > k - (1-4\delta)kN^2 \cdot \frac{(1+\delta)^2}{N^2}\right] \\
&\leq \Pr\left[T_m - \mathbb{E}[T_m] > \delta k\right] \\
&\leq \delta^{-2}k^{-2} \cdot \mathrm{Var}[T_m] \\
&\leq O(\varepsilon^{-6}k^{-2}) \cdot O(k^{5/3}) \\
&\leq \frac{1}{O(\varepsilon^6 k^{1/3})} \\
&\leq \frac{1}{8}.
\end{aligned}$$

Hence, with constant probability, there will be at most k 3-way collisions in the samples from $\mathbf{p}$ and it will be accepted. The sample and running time complexity is then

$$\Theta\left(\frac{1}{\varepsilon^6\|\mathbf{p}\|_2} + \varepsilon^{-6}N^{2/3}\right) = \Theta\left(\frac{1}{\varepsilon^6\|\mathbf{p}\|_2} + \frac{1}{\varepsilon^6\|\mathbf{p}\|_3}\right) = \Theta\left(\frac{1}{\varepsilon^6\|\mathbf{p}\|_3}\right).$$

Now, for the soundness argument, suppose that after $m = M$ samples, at most k 3-way collisions are observed. We can then argue that, with some constant probability, $\|\mathbf{p}\|_3^3$ is less than $\frac{1+5\delta}{N^2}$. If $\|\mathbf{p}\|_3^3 > \frac{1+5\delta}{N^2}$, then

$$\mathbb{E}[T_m] = \binom{m}{3} \cdot \|\mathbf{p}\|_3^3 > (1-4\delta)kN^2 \cdot \frac{1+5\delta}{N^2} \geq (1+\delta/2)k.$$

Then,

$$\begin{aligned}
\Pr[T_m \leq k] &= \Pr[|T_m - \mathbb{E}[T_m]| \geq \delta k/2] \\
&\leq \frac{4\,\mathrm{Var}[T_m]}{\delta^2 k^2} \\
&\leq O\left(\frac{4k^{5/3}N^{10/3}(1+5\delta)^{5/3}}{\varepsilon^6 k^2 N^{10/3}}\right) \\
&\leq O\left(\frac{1}{\varepsilon^{16}k^2}\right) \\
&\leq \frac{1}{4}
\end{aligned}$$

Hence, we have that

$$\frac{1-\delta}{N} \leq \|\mathbf{p}\|_2^2 \leq \frac{1+\delta}{N}$$

and

$$\|\mathbf{p}\|_3^3 \leq \frac{1+5\delta}{N^2}.$$

By Lemma 3.4, we have that $\mathbf{p}$ is within $9\sqrt[3]{8\delta} = \varepsilon$ of $\mathcal{C}_U$.

For a distribution $\mathbf{p}$ that is ε-far from uniform, the algorithm will stop after observing k 3-way collisions with constant probability. Similar to the arguments above, this will happen when the number m of samples satisfies

$$\binom{m}{3} \cdot \|\mathbf{p}\|_3^3 \approx k.$$

Hence, the sample complexity of the algorithm in this case is

$$\Theta\left(\frac{1}{\varepsilon^6\|\mathbf{p}\|_2} + \varepsilon^{-6}N^{2/3}\right) = \Theta\left(\frac{1}{\varepsilon^6\|\mathbf{p}\|_2} + \frac{1}{\varepsilon^6\|\mathbf{p}\|_3}\right) = \Theta\left(\frac{1}{\varepsilon^6\|\mathbf{p}\|_3}\right).$$

□

4 The Lower Bound

In this section, we prove our main lower bound, restated below.

Theorem 1.2. *Fix any (non-uniform) distribution $\mathbf{q}$ over Ω, and let $\varepsilon \overset{\text{def}}{=} \mathrm{d}_{\mathrm{TV}}(\mathbf{q}, \mathcal{C}_U)$ be its distance to $\mathcal{C}_U$. Then, given sample access to a distribution $\mathbf{p}$ on Ω, distinguishing with high constant probability between (i) $\mathbf{p}$ is equal to $\mathbf{q}$ up to a permutation of the domain and (ii) $\mathbf{p} \in \mathcal{C}_U$, requires $\Omega\Big(\frac{1}{\|\mathbf{q}\|_3}\Big)$ samples. In particular, an algorithm that tests membership in $\mathcal{C}_U$ with high probability and for any proximity parameter $\varepsilon' \leq \varepsilon$ requires this many samples.*

Proof. Let $\mathbf{q} \in \Delta(\Omega)$ and $\varepsilon \in (0, 1]$ be as in the statement of the theorem. To argue that (a permutation of) $\mathbf{q}$ is hard to distinguish from some $\mathbf{u} \in \mathcal{C}_U$ with few samples (where "few" is a function of $\mathbf{q}$ and ε only), we will rely on the Wishful Thinking Theorem of Valiant [Val11]. Indeed, this theorem, broadly speaking, ensures that two distributions with moments (nearly) matching are hard to distinguish given only their fingerprints (equivalently, that distinguishing between relabelings of $\mathbf{q}$ and relabelings of $\mathbf{u}$ is hard). This will be enough to conclude, as $\mathcal{C}_U$ is a symmetric property.

Specifically, we define the two distributions $\mathbf{p}^{\mathsf{yes}}, \mathbf{p}^{\mathsf{no}}$ (respectively in $\mathcal{C}_U$ and ε-far from it) as follows:

- $\mathbf{p}^{\mathsf{no}}$ is the "no-distribution" imposed to us – that is, $\mathbf{p}^{\mathsf{no}} = \mathbf{q}$;
- $\mathbf{p}^{\mathsf{yes}}$ is a uniform distribution on a set $S \subseteq \Omega$ of $1/\|\mathbf{q}\|_2^2$ elements.

(To see why this is a natural choice: the natural "yes-distribution" to consider in order to fool an algorithm is, by the Wishful Thinking Theorem, a distribution that matches as many moments of $\mathbf{p}^{\mathsf{no}} = \mathbf{q}$ as possible; which, in our case, will mean matching the $\|\cdot\|_1$, and $\|\cdot\|_2$ moments. Note that we could try to *approximately* match the third moment, $\|\cdot\|_3$, as well, but that there is no hope to match it perfectly: if we could do so with a uniform distribution, this by Lemma 2.2 would imply that $\mathbf{q}$ was in $\mathcal{C}_U$ to begin with.)

In what follows, in view of deriving our lower bound we suppose that $k\|\mathbf{q}\|_3 \ll 1$. Let $\mathbf{p}^{\mathsf{yes}}$ be a uniform distribution on a subset of $m \overset{\text{def}}{=} \frac{1}{\|\mathbf{q}\|_2^2}$ elements. Computing the k-based moments of $\mathbf{p}^{\mathsf{yes}}$ is straightforward: for any $j \geq 2$, we have

$$m^{\mathsf{yes}}(j) = \frac{k^j}{m^{j-1}} = k^j\|\mathbf{q}\|_2^{2(j-1)} = \frac{\left(k\|\mathbf{q}\|_2^2\right)^j}{\|\mathbf{q}\|_2^2}$$

while, of course, $m^{\mathsf{no}}(j) = k^j\|\mathbf{q}\|_j^j$. It follows that

$$\begin{aligned}\sum_{j=2}^{\infty} \frac{|m^{\mathsf{yes}}(j) - m^{\mathsf{no}}(j)|}{\sqrt{1+\max(m^{\mathsf{yes}}(j), m^{\mathsf{no}}(j))}} &= \sum_{j=3}^{\infty} \frac{|m^{\mathsf{yes}}(j) - m^{\mathsf{no}}(j)|}{\sqrt{1+\max(m^{\mathsf{yes}}(j), m^{\mathsf{no}}(j))}} \\ &= \sum_{j=3}^{\infty} k^j \frac{\left|\|\mathbf{q}\|_j^j - \|\mathbf{q}\|_2^{2(j-1)}\right|}{\sqrt{1+k^j\max(\|\mathbf{q}\|_2^{2(j-1)}, \|\mathbf{q}\|_j^j)}}\ .\end{aligned}$$

Now, we will use Fact 2.3 to get rid of the absolute value; as it enables us to rewrite our sum as

$$\sum_{j=2}^{\infty} \frac{|m^{\mathsf{yes}}(j) - m^{\mathsf{no}}(j)|}{\sqrt{1+\max(m^{\mathsf{yes}}(j), m^{\mathsf{no}}(j))}} = \sum_{j=3}^{\infty} k^j \frac{\|\mathbf{q}\|_j^j - \|\mathbf{q}\|_2^{2(j-1)}}{\sqrt{1+k^j\|\mathbf{q}\|_j^j}}\ .$$

In order to handle this last expression, we can drop the denominator, to get

$$\begin{aligned}\sum_{j=2}^{\infty} \frac{|m^{\mathsf{yes}}(j) - m^{\mathsf{no}}(j)|}{\sqrt{1+\max(m^{\mathsf{yes}}(j), m^{\mathsf{no}}(j))}} &\le \sum_{j=3}^{\infty} k^j \left(\|\mathbf{q}\|_j^j - \|\mathbf{q}\|_2^{2(j-1)}\right) \le \sum_{j=3}^{\infty} k^j\|\mathbf{q}\|_j^j \\ &\le \sum_{j=3}^{\infty} k^j\|\mathbf{q}\|_3^j \qquad \text{(Monotonicity of } \ell_p \text{ norms)} \\ &= \frac{k^3\|\mathbf{q}\|_3^3}{1-k\|\mathbf{q}\|_3} < \frac{1}{24}\end{aligned}$$

using our assumption that $k\|\mathbf{q}\|_3 \ll 1$.

This last bound will allow us to apply Theorem 2.1 and obtain the lower bound, provided that $\|\mathbf{p}^{\mathsf{yes}}\|_\infty, \|\mathbf{p}^{\mathsf{no}}\|_\infty \le \frac{1}{500k}$. But this last condition follows from observing that $k\max(\|\mathbf{p}^{\mathsf{yes}}\|_\infty, \|\mathbf{p}^{\mathsf{no}}\|_\infty) \le k\max(\|\mathbf{p}^{\mathsf{yes}}\|_3, \|\mathbf{p}^{\mathsf{no}}\|_3) = k\max(\|\mathbf{q}\|_3, \|\mathbf{q}\|_2^{4/3}) \le k\|\mathbf{q}\|_3 \ll 1$.

□

Remark 4.1. Although our lower bound does not directly feature a dependence on the distance parameter ε (besides applying to any $\varepsilon' \le \varepsilon$), we conjecture that the right dependence should be linear in $1/\varepsilon$, i.e., $\Omega(1/(\varepsilon\|\mathbf{q}\|_3))$. (Indeed, while a *square* dependence on ε appears natural, it cannot hold on an instance-by-instance basis for *all* distributions, analogously to that of Lemma 3.2: as one could see by considering a degenerate distribution $\mathbf{q}$ with $1-\varepsilon$ probability weight on a single element, for which uniformity testing can be done with $O(1/\varepsilon) = \Omega(1/(\varepsilon\|\mathbf{q}\|_3))$ samples.) Establishing this linear dependence with our techniques, however, would require at the very least a significant strengthening of the above chain of inequalities, especially at step (†).

References

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Proceedings of NIPS*, pages 3577–3598, 2015. 1

[AOST17] Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. Estimating renyi entropy of discrete distributions. *IEEE Trans. Information Theory*, 63(1):38–56, 2017. 3.3

[BCG16] Eric Blais, Clément Louis Canonne, and Tom Gur. Alice and bob show distribution testing lower bounds (they don't talk to each other anymore.). *Electronic Colloquium on Computational Complexity (ECCC)*, 23:168, 2016. 1, 1.1

[BDKR05] Tuğkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005. 3.1

[BFF$^+$01] Tuğkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of FOCS*, pages 442–451, 2001. 1

[BFR$^+$13] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *J. ACM*, 60(1):4:1–4:25, 2013. 3.1

[Can15] Clément L. Canonne. A Survey on Distribution Testing: your data is Big. But is it Blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, April 2015. 1

[CP00] Michael Camarri and Jim Pitman. Limit distributions and random trees derived from the birthday problem with unequal probabilities. *Electron. J. Probab.*, 5:18 pp., 2000. 3.1

[DGPP16] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Collision-based testers are optimal for uniformity and closeness. *ArXiV*, abs/1611.03579, 2016. 1

[DK16] Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *Proceedings of FOCS*. IEEE Computer Society, 2016. 1

[DKN15] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing Identity of Structured Distributions. In *Proceedings of SODA*, pages 1841–1854. Society for Industrial and Applied Mathematics (SIAM), January 2015. 1

[eh17] esg (http://mathoverflow.net/users/48831/esg). Birthday problem with unequal probability: expected number of draws before the m-th collision? MathOverflow, March 2017. http://mathoverflow.net/q/263749 (version: 2017-03-05). 3.1

[FL06] Magalie Fromont and Béatrice Laurent. Adaptive goodness-of-fit tests in a density model. *Ann. Statist.*, 34(2):680–720, 04 2006. 1

[GGR98] Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, 45(4):653–750, July 1998. 1

[Gol16] Oded Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:15, 2016. 1

[Gol17] Oded Goldreich. *Introduction to Property Testing.* Forthcoming, 2017. 1

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 7:20, 2000. 1, 1.2

[Ing00] Yu. I. Ingster. Adaptive chi-square tests. *Journal of Mathematical Sciences*, 99(2):1110–1119, apr 2000. 1

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008. 1

[Ron08] Dana Ron. Property testing: A learning theory perspective. *Foundations and Trends in Machine Learning*, 1(3):307–402, 2008. 1

[Ron09] Dana Ron. Algorithmic and analysis techniques in property testing. *Foundations and Trends in Theoretical Computer Science*, 5(2):73–205, 2009. 1

[RRSS09] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distributions support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009. 1.2

[RS96] Ronitt Rubinfeld and Madhu Sudan. Robust characterization of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, 1996. 1

[Rub12] Ronitt Rubinfeld. Taming big probability distributions. *XRDS: Crossroads, The ACM Magazine for Students*, 19(1):24, sep 2012. 1

[Val11] Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011. 1.2, 2.2, 2.1, 4

[VV10a] Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:179, 2010. 4

[VV10b] Gregory Valiant and Paul Valiant. Estimating the unseen: A sublinear-sample canonical estimator of distributions. *Electronic Colloquium on Computational Complexity (ECCC)*, 17:180, 2010. 4

[VV11] Gregory Valiant and Paul Valiant. The power of linear estimators. In *Proceedings of FOCS*, pages 403–412, October 2011. See also [VV10a] and [VV10b]. 1.2

[VV14] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *Proceedings of FOCS*, pages 51–60, 2014. 1, 1.1