Testing Monotone Continuous Distributions on High-dimensional Real Cubes *†

Artur Czumaj ‡
Joint work with Michał Adamaszek§ and Christian Sohler¶

We study the task of testing properties of probability distributions and our focus is on understanding the role of continuous distributions in this setting. We consider a scenario in which we have access to independent samples of an unknown distribution $\mathfrak{D}$ with infinite (perhaps even uncountable) support (e.g., on the interval $[0,1]$ or on the n-dimensional real cube $[0,1]^n$). Our goal is to test whether $\mathfrak{D}$ has a given property or it is ε-far from it (in the statistical distance, with the L_1-distance measure).

We study the task of testing properties of probability distributions. We consider a scenario in which we have access to independent samples of an unknown distribution $\mathfrak{D}$ with infinite (perhaps even uncountable) support. Our goal is to test whether $\mathfrak{D}$ has a given property or it is ε-far from it (in the statistical distance, with the L_1-distance measure).

The topic of testing basic properties of the underlying probability distributions has been extensively studied for many decades. While the standard approach in statistics (and also more modern approaches, e.g., in data mining) have led to the development of many high quality techniques and algorithms, until very recently little attention has been paid to the computational complexity of testing in the situations when the underlying distributions are over very large domains. Motivated by these considerations, a number of new studies have emerged that aim at developing efficient testers for various properties of distributions with the focus on the small number of samples used for testing. In particular, it has been shown that for a number of fundamental properties, such as independence, entropy estimation, and the closeness between distributions, it is possible to test the underlying distribution with the number of samples sublinear in the domain size.

While these studies lead to very efficient testers for various properties for distributions on finite support, they seem to be useless when the underlying distribution is on a continuous, or infinite, or even uncountable domain. In this paper, our goal is to study the phenomenon of testability of continuous distributions.

Setting. We assume that there is an underlying probability distribution $\mathfrak{D}$ from which we can draw *independent identically distributed samples* (see, e.g., [4]). We assume that each sample is of infinite precision and we will not consider the issue of representation of the real numbers. The *complexity of the tester* is measured in terms of the *number of samples* required in order to obtain a desired information about the distribution.

*Research supported in part by EPSRC award EP/G064679/1, DFG grant So 514/3-1, and by the Centre for Discrete Mathematics and its Applications (DIMAP), EPSRC award EP/D063191/1.

†A preliminary, conference version of this work appeared in *Proceedings of the 21st ACM-SIAM Symposium on Discrete Algorithms (SODA'10)*, pages 56–65, 2010. SIAM, Philadelphia, PA, 2010.

‡Department of Computer Science and Centre for Discrete Mathematics and its Applications (DIMAP), University of Warwick. Email: A.Czumaj@warwick.ac.uk.

§Warwick Mathematics Institute and Centre for Discrete Mathematics and its Applications (DIMAP), University of Warwick. Email: M.J.Adamaszek@warwick.ac.uk.

¶Department of Computer Science, Technische Universität Dortmund. Email: sohler@informatik.uni-bonn.de.

We study probability distributions over a domain Ω which will be either finite or infinite; our main focus is on the domain $\Omega = [0,1]^n$, $n \in \mathbb{N}$, that is, (continuous) n-dimensional unit cube.

Recall, by the Radon-Nikodym theorem, that every distribution on Ω has a Lebesgue decomposition into a sum of two parts: (i) continuous (with respect to the standard Lebesgue measure), that is, given by a measurable density function $\mathfrak{f}$, (ii) singular (concentrated on a set of Lebesgue measure 0).

We study the similarity and dissimilarity between various distributions. Following the mainstream research of testing properties of distributions in theoretical computer science, we use the *total variation distance to measure the similarity between distributions* (L_1-distance). For any two *discrete* distributions $\mathcal{X}$ and $\mathcal{Y}$ over Ω, defined by the probability functions $\mathbf{Pr}_{\mathcal{X}}$ and $\mathbf{Pr}_{\mathcal{Y}}$, respectively, we say $\mathcal{Y}$ *is ε-far from* $\mathcal{X}$ if

$$\frac{1}{2} \cdot \sum_{\omega \in \Omega} |\mathbf{Pr}_{\mathcal{X}}[\omega] - \mathbf{Pr}_{\mathcal{Y}}[\omega]| \quad \geq \quad \varepsilon \ .$$

For general distributions, the definition is analogous: for any two distributions $\mathcal{X}$ and $\mathcal{Y}$ over Ω, with density functions $\mathfrak{f}_{\mathcal{X}}$ and $\mathfrak{f}_{\mathcal{Y}}$, respectively, we say $\mathcal{Y}$ *is ε-far from* $\mathcal{X}$ if

$$\frac{1}{2} \cdot \int_{\boldsymbol{x} \in \Omega} |\mathfrak{f}_{\mathcal{X}}(\boldsymbol{x}) - \mathfrak{f}_{\mathcal{Y}}(\boldsymbol{x})| \, d\boldsymbol{x} \quad \geq \quad \varepsilon \ . \tag{1}$$

We say $\mathcal{Y}$ *is ε-close to* $\mathcal{X}$ if $\mathcal{Y}$ is not ε-far from $\mathcal{X}$.

Note that inequality (1) is equivalent to

$$\int_{\boldsymbol{x} \in \Omega : \mathfrak{f}_{\mathcal{X}}(\boldsymbol{x}) \geq \mathfrak{f}_{\mathcal{Y}}(\boldsymbol{x})} (\mathfrak{f}_{\mathcal{X}}(\boldsymbol{x}) - \mathfrak{f}_{\mathcal{Y}}(\boldsymbol{x})) \, d\boldsymbol{x} \quad \geq \quad \varepsilon.$$

Let us remind that a distribution $\mathfrak{D}$ over $[0,1]^n$ with density function $\mathfrak{f}$ is *uniform* if $\mathfrak{f}$ is identically 1. Therefore, for the uniform distribution, (1) can be rephrased as follows: A distribution $\mathfrak{D}$ over $[0,1]^n$ with density function $\mathfrak{f}$ is *ε-far from uniform* if

$$\frac{1}{2} \cdot \int_{\boldsymbol{x} \in [0,1]^n} |\mathfrak{f}(\boldsymbol{x}) - 1| \, d\boldsymbol{x} \quad \geq \quad \varepsilon \ .$$

Our goal is to design an algorithm that for a given positive ε and a given underlying probability distribution $\mathfrak{Q}$ and a probability distribution $\mathfrak{D}$ available through random sampling, is able to distinguish between the case when $\mathfrak{Q} = \mathfrak{D}$ and when $\mathfrak{D}$ is ε-far from it $\mathfrak{Q}$. The algorithm is allowed to be randomized and can err with probability at most $\frac{1}{4}$.

Continuous distributions are typically not testable. In general, when using the total variation distance to measure the similarity between distributions, it is infeasible to investigate interesting properties of distributions on infinite domains without any assumptions on the density function. For example, one can show that for every integer t there is no tester A that distinguishes with at most t samples between uniform distribution $\mathfrak{D}_U$ on $[0,1]$ and any distribution that is ε-far from uniform (for example, take a uniform distribution on t^3 randomly chosen points from the interval $[0,1]$; such distribution is discrete and hence it is $\frac{1}{2}$-far from uniform). This observation can be easily generalized to testing a number of natural properties for distributions on infinite domains.

One can also derive similar impossibility results from the existing lower bounds for testing properties of discrete distributions. For example, Batu et al. [4] (see also [6]) show that testing if a given distribution on the support of size n is uniform requires $\Omega(\sqrt{n}/\varepsilon^2)$ samples. With that, by taking $n \to \infty$, the lower

bound in [4] immediately implies that no algorithm can test if a given distribution on $[0, 1]$ is uniform. This approach implies also similar impossibility results for testing if a given distribution is monotone, unimodal, or if two distributions are identical, are independent, and so on (see [1, 2, 3, 4, 5, 7, 8, 9] for more examples).

Once we see these negative result, the natural question is: what properties of distributions on infinite domains can be tested?

Testing if a distribution is discrete on N points. In order to understand the problem of testing distributions on infinite domains, the very first question should be to test if a given distribution has infinite support. We first briefly consider a dual question: to verify if a given distribution has support of up to a given size. Recall, that a point $\boldsymbol{x}$ is called an *atom* of $\mathfrak{D}$ if $\mathbf{Pr}_{\mathfrak{D}}[\boldsymbol{x}] > 0$. Detection of a single atom is not possible, since its probability may be arbitrarily small, beyond the resolution of any given algorithm. Instead we may try to determine whether a large part of the probability mass is concentrated on the atoms: for a given parameter N, distinguish between distributions that have the entire support on at most N points (*discrete on N points*) and those that are ε-far from discrete on N points.

A related question has been studied recently by Raskhodnikova et al. [7], who were interested in estimating the size of the support of a given distribution under the assumption that every element in the support is an atom (distribution is singular) with the probability at least $\frac{1}{M}$. For such problem, Raskhodnikova et al. [7] (see also [10]) show that one needs at least $\Omega(M^{1-o(1)})$ samples to estimate the size of the support. On the other hand, it is easy to compute (exactly) the size of the support with $\mathcal{O}(M \log M)$ samples (e.g., by using the approach from the coupon collector problem). Our goal is different than that in [7], because on one hand, we do not have any lower bound for the probability of the points in the support (which makes the task of even estimating the size of the support impossible), and on the other hand, we want to test if a given distribution has at most N points in the support (rather than estimate the size of the support). Still, one can rather easily that the lower bound result from Raskhodnikova et al. [7] carries over for our problem and gives a lower bound for the sample size of $\Omega(N^{1-o(1)})$. One can also show that the following algorithm sampling $\mathcal{O}(N/\varepsilon)$ elements is a testing algorithm that distinguishes between a discrete distribution on N points and any distribution that is ε-far from discrete on N points with $\mathcal{O}(N/\varepsilon)$ samples.

Testing discreteness (N):

- Draw a sample (according to the distribution $\mathfrak{D}$) $S = \langle s_1, \ldots, s_\ell \rangle$ from Ω with $\ell = \lceil 2N/\varepsilon \rceil$
- If S has more than N distinct elements then **Reject**
- else **Accept**

Observe that this result immediately implies that we can estimate the smallest number $\mathcal{N}$ of points in the domain of $\mathfrak{D}$ such that $\mathfrak{D}$ has $\mathcal{N}$ points that have the total probability at least $1 - \varepsilon$ using $\mathcal{O}(\mathcal{N}/\varepsilon)$ samples.

The obtained lower and upper bound it is an interesting **open question** of whether the upper bound is tight, that is, if every algorithm testing if a distribution is discrete on N points requires $\Omega(N/\varepsilon)$ samples.

Testing if a monotone high-dimensional distribution on a real hypercube is uniform. The main goal of this work is to investigate if there are any interesting distributions on infinite domains that are testable. One of a very few properties of discrete distributions considered in the Computer Science literature that has only a light dependency on the size of the support (the condition that by our discussion above seems to be necessary to hope for a fast tester) is that of *testing if a monotone distribution*[1] *on the Boolean cube is*

[1]Distribution $\mathfrak{D}$ *is monotone* if for any $\boldsymbol{x} = (x_1, \ldots, x_n)$, $\boldsymbol{y} = (y_1, \ldots, y_n)$, if $x_i \leq y_i$ for every i then $\mathbf{Pr}_{\mathfrak{D}}[\boldsymbol{x}] \leq \mathbf{Pr}_{\mathfrak{D}}[\boldsymbol{y}]$.

uniform. Rubinfeld and Servedio [9] consider the following problem:

given a monotone distribution $\mathfrak{D}$ on a Boolean n-dimensional cube $\{0,1\}^n$, test if $\mathfrak{D}$ is uniform.

Rubinfeld and Servedio [9] show that without any assumption about the monotonicity of $\mathfrak{D}$, every testing algorithm requires $2^{\Omega(n)}$ samples (because the domain's size is 2^n), however, if $\mathfrak{D}$ is monotone, then one distinguishes between the case when $\mathfrak{D}$ is uniform and when $\mathfrak{D}$ is ε-far from uniform using $\mathcal{O}(n\log(1/\varepsilon)/\varepsilon^2)$ samples. Furthermore, this result is almost optimal in the sense that $\Omega(n/\log^2 n)$ samples are necessary [9].

Our main contribution is the analysis of this problem in the setting when $\mathfrak{D}$ is a monotone distribution[2] on an n-dimensional (*real*) cube $[0,1]^n$. On high-level our approach is similar to that used by Rubinfeld and Servedio [9] in the case of Boolean n-cubes. However, the fact that we have to deal with continuous domain makes our proof of the main result, Lemma 1, more complicated.

First, we provide a characterization of monotone distributions that are ε-far from uniform:

Lemma 1 *Let $\mathfrak{D}$ be a monotone distribution on $[0,1]^n$ with density function $\mathfrak{f}$. If $\mathfrak{D}$ is ε-far from uniform then*

$$\mathbb{E}_{\mathfrak{f}}[\|\boldsymbol{x}\|_1] \;=\; \int_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \cdot \mathfrak{f}(\boldsymbol{x})\, d\boldsymbol{x} \geq \frac{n}{2} + \frac{\varepsilon}{2}\ .$$

The proof of this lemma can be deduced from the following result (which is the main technical contribution of the paper) by substituting $\mathfrak{g}(\boldsymbol{x}) = \mathfrak{f}(\boldsymbol{x}) - 1$.

Lemma 2 *Let $\mathfrak{g} : [0,1]^n \to \mathbb{R}$ be a monotone function with $\int_{\boldsymbol{x}} \mathfrak{g}(\boldsymbol{x})\, d\boldsymbol{x} = 0$. Then*

$$\int_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \cdot \mathfrak{g}(\boldsymbol{x})\, d\boldsymbol{x} \;\;\geq\;\; \frac{1}{4}\int_{\boldsymbol{x}} |\mathfrak{g}(\boldsymbol{x})|\, d\boldsymbol{x}\ .$$

By combining Lemma 1 with the fact that for uniform distribution $\mathfrak{U}$ on $[0,1]^n$ we have $\mathbb{E}_{\mathfrak{U}}[\|\boldsymbol{x}\|_1] = \frac{n}{2}$, we can show that the following simple algorithm tests if a distribution is uniform or it is ε-far from uniform:

Testing uniformity:

- **Repeat** $r = 20$ times:

  Draw a sample (according to the distribution $\mathfrak{D}$) $S = \langle \boldsymbol{x}_1, \ldots, \boldsymbol{x}_s\rangle$ from $[0,1]^n$ with $s = \lceil \frac{40n}{\varepsilon^2} \rceil$

  If $\sum_{i=1}^{s} \|\boldsymbol{x}_i\|_1 \geq s(\frac{n}{2} + \frac{\varepsilon}{4})$ then **Reject** and exit
- **Accept**

Theorem 3 Testing uniformity *distinguishes between uniform distribution on $[0,1]^n$ and any monotone distribution over $[0,1]^n$ that is ε-far from uniform. Its sample complexity is $\mathcal{O}(n/\varepsilon^2)$ and it errs with the probability at most $\frac{1}{4}$.*

Our analysis not only extends the result from [9] to real cubes, but also leads to an algorithm slightly faster than that from [9] (we shave off an $\mathcal{O}(\log(n/\varepsilon))$ factor) for both the Boolean and real hypercube. We observe that since the lower bound from [9] can be directly carried over to the case of real $[0,1]^n$ cubes, our upper bound is almost optimal.

Let us also notice that our tester will work with the same complexity if the input is a monotone distribution on a discrete cube $\{0,1,\ldots,k\}^n$. The obtained sample size is independent of k.

[2] A distribution $\mathfrak{D}$ on $[0,1]^n$ with density function $\mathfrak{f}$ is *monotone* if for any $\boldsymbol{x} = (x_1,\ldots,x_n)$, $\boldsymbol{y} = (y_1,\ldots,y_n)$, if $x_i \leq y_i$ for every i then $\mathfrak{f}(\boldsymbol{x}) \leq \mathfrak{f}(\boldsymbol{y})$.

Theorem 4 *Let k be any positive integer and consider any n-dimensional finite grid $\{0, 1, 2, \ldots, k\}^n$. One can test if a given monotone distribution $\mathfrak{D}$ over $\{0, 1, 2, \ldots, k\}^n$ is uniform with $\mathcal{O}(n/\varepsilon^2)$ samples.*

References

[1] N. Alon, A. Andoni, T. Kaufman, K. Matulef, R. Rubinfeld, and N. Xie. Testing k-wise and almost k-wise Independence. *Proceedings of the 39th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 496–505, 2007.

[2] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 678–687, 2002.

[3] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 442–415, 2001.

[4] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. *Proceedings of the 41st IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 259–269, 2000.

[5] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 381–390, 2004.

[6] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, Report No. 7, 2000.

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

[8] R. Rubinfeld. Sublinear time algorithms. *Proceedings of the International Congress of Mathematicians*, Madrid, Spain, August 22–30, 2006.

[9] R. Rubinfeld and R. A. Servedio. Testing monotone high-dimensional distributions. *Proceedings of the 37th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 147–156, 2005.

[10] P. Valiant. Testing symmetric properties of distributions. *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 383–392, 2008.