

# Testing Distributions of Huge Objects

Oded Goldreich  

Department of Computer Science, Weizmann Institute of Science, Israel

Dana Ron  

School of Electrical Engineering, Tel Aviv University, Israel

---

## Abstract

---

We initiate a study of a new model of property testing that is a hybrid of testing properties of distributions and testing properties of strings. Specifically, the new model refers to testing properties of distributions, but these are distributions over huge objects (i.e., very long strings). Accordingly, the model accounts for the total number of local probes into these objects (resp., queries to the strings) as well as for the distance between objects (resp., strings). Specifically, the distance between distributions is defined as the earth mover's distance with respect to the relative Hamming distance between strings.

We study the query complexity of testing in this new model, focusing on three directions. First, we try to relate the query complexity of testing properties in the new model to the sample complexity of testing these properties in the standard distribution testing model. Second, we consider the complexity of testing properties that arise naturally in the new model (e.g., distributions that capture random variations of fixed strings). Third, we consider the complexity of testing properties that were extensively studied in the standard distribution testing model: Two such cases are uniform distributions and pairs of identical distributions, where we obtain the following results.

- Testing whether a distribution over  $n$ -bit long strings is uniform on some set of size  $m$  can be done with query complexity  $\tilde{O}(m/\epsilon^3)$ , where  $\epsilon > (\log_2 m)/n$  is the proximity parameter.
- Testing whether two distribution over  $n$ -bit long strings that have support size at most  $m$  are identical can be done with query complexity  $\tilde{O}(m^{2/3}/\epsilon^3)$ .

Both upper bounds are quite tight; that is, for  $\epsilon = \Omega(1)$ , the first task requires  $\Omega(m^c)$  queries for any  $c < 1$  and  $n = \omega(\log m)$ , whereas the second task requires  $\Omega(m^{2/3})$  queries. Note that the query complexity of the first task is higher than the sample complexity of the corresponding task in the standard distribution testing model, whereas in the case of the second task the bounds almost match.

**2012 ACM Subject Classification** Theory of computation  $\rightarrow$  Streaming, sublinear and near linear time algorithms

**Keywords and phrases** Property Testing, Distributions

**Digital Object Identifier** 10.4230/LIPIcs.ITCS.2022.78

**Related Version** *Full Version*: ECCC TR21-133

**Funding** *Oded Goldreich*: Partially supported by the Israel Science Foundation (grant No. 1041/18); received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819702).

*Dana Ron*: Partially supported by the Israel Science Foundation (grant No. 1041/18).

**Acknowledgements** We are grateful to Avi Wigderson for a discussion that started this research project.

## 1 Introduction

In the last couple of decades, the area of property testing has attracted much attention (see, e.g., a recent textbook [8]). Loosely speaking, property testing typically refers to sub-linear time probabilistic algorithms for deciding whether a given object has a predetermined



© Oded Goldreich and Dana Ron;  
licensed under Creative Commons License CC-BY 4.0  
13th Innovations in Theoretical Computer Science Conference (ITCS 2022).  
Editor: Mark Braverman; Article No. 78; pp. 78:1–78:19



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

property or is far from any object having this property. Such algorithms, called testers, obtain local views of the object by making adequate queries; that is, the object is modeled as a function and testers get oracle access to this function (and thus may be expected to work in time that is sub-linear in the size of the object).

The foregoing description fits much of the research in the area (see [8, Chap. 1-10]), but not the part that deals with testing properties of distributions (aka *distribution testing*, see [8, Chap. 11] and [6]). In this context, a tester gets samples from the tested distribution and sub-linearity means sub-linearity in the size of the distribution's domain.<sup>1</sup> Each element in the domain is considered to be small, and is assumed to be processed at unit time.

In this work we consider distributions over sets of huge (or very large) objects, and aim at complexities that are sublinear in the size of these objects. As an illustrative example, think of the distribution of DNA-sequences in a large population. We wish to sample this distribution and *query each sampled sequence at locations of our choice rather than read the entire sample*.

One key issue is the definition of the distance between such distributions (i.e., distributions of huge objects). A natural choice, which we use, is the *earth mover's distance* under the (relative) Hamming measure. Under this measure, the distance between distributions reflects the probability mass of the difference when weighted according to the Hamming distance between strings (see Definition 1.1).

## 1.1 The new model

We consider properties of distributions over sets of objects that are represented by  $n$ -bit long strings (or possibly  $n$ -symbol long sequences); that is, each object has size  $n$ . (In Section 5 of our report [10], this is extended to properties of tuples of distributions.) Each of these objects is considered huge, and so we do not read it in full but rather probe (or query) it at locations of our choice. Hence, the tester is an algorithm that may ask for few samples, and queries each sample at locations of its choice. This is modeled as getting oracle access to several oracles, where each of these oracles is selected independently according to the tested distribution (see Definition 1.2). We shall be mainly interested in the total number of queries (made into these samples), whereas the number of samples will be a secondary consideration.

The distance between such distributions,  $P$  and  $Q$  (over the same domain  $\Omega = \{0, 1\}^n$ ), is defined as the *earth mover's distance under the Hamming measure*; that is, the cost of transforming the distribution  $P$  to the distribution  $Q$ , where the cost of transforming a string  $x$  to a string  $y$  equals their relative Hamming distance.

► **Definition 1.1.** (distance between distributions over huge objects): *For two strings  $x, y \in \{0, 1\}^n$ , let  $\Delta_H(x, y)$  denote the relative Hamming distance between them; that is,*

$$\Delta_H(x, y) = \frac{1}{n} \cdot |\{i \in [n] : x_i \neq y_i\}|. \quad (1)$$

*For two distributions  $P, Q : \Omega \rightarrow [0, 1]$ , where  $\Omega = \{0, 1\}^n$ , the earth mover's distance under the Hamming measure between  $P$  and  $Q$ , is the optimal value of the following linear program:*

---

<sup>1</sup> This is the most standard and well studied model of testing properties of distributions. For a discussion of other models (e.g., providing the algorithm with the weight of any domain element of its choice) see [6, Part IV].

$$\begin{array}{l}
\forall x \in \Omega: \quad \min \\
\forall y \in \Omega: \quad \sum_{x \in \Omega} w_{x,y} = P(x) \\
\forall x, y \in \Omega: \quad w_{x,y} \geq 0
\end{array}
\left\{ \sum_{x, y \in \Omega} w_{x,y} \cdot \Delta_H(x, y) \right\} \quad (2)$$

We say that  $P$  is  $\epsilon$ -close to  $Q$  if the optimal value of the linear program is at most  $\epsilon$ ; otherwise, we say that  $P$  is  $\epsilon$ -far from  $Q$ .

As stated above, Definition 1.1 represents the earth mover's distance with respect to the relative Hamming distance between (binary) strings. Indeed, the earth mover's distance between distributions over a domain  $\Omega$  is always defined on top of a distance measure that is associated with  $\Omega$ . It is well known that the earth mover's distance with respect to the inequality function (i.e.,  $\text{InEq}(x, y) = 1$  if  $x \neq y$  and  $\text{InEq}(x, x) = 0$ ) coincides with the variation distance (between these distributions). That is, if we replace the distance  $\Delta_H(x, y)$  with  $\text{InEq}(x, y)$  in Definition 1.1, then we get the variation distance between  $P$  and  $Q$ . Furthermore,  $\Delta_H(x, y) \leq \text{InEq}(x, y)$  always holds. Hence, throughout this work, we shall be considering three distance measures:

1. The *distance between distributions* as defined above (i.e., in Definition 1.1). When we say that distributions are “close” or “far” we refer to this notion.
2. The *total variation distance between distributions*. In this case, we shall say that the distributions are “TV-close” or “TV-far” (or  $\epsilon$ -TV-close/far).
3. The *relative Hamming distance between strings*, which we denoted by  $\Delta_H(\cdot, \cdot)$ . In this case, we shall say that the strings are “H-close” or “H-far” (or  $\epsilon$ -H-close/far).

Referring to Definition 1.1 and to machines that have access to multiple oracles, we present the following definition of testing distributions on huge objects.

► **Definition 1.2.** (testing properties of distributions on huge objects (the DOHO model)): Let  $\mathcal{D} = \{\mathcal{D}_n\}_{n \in \mathbb{N}}$  be a property of distributions such that  $\mathcal{D}_n$  is a set of distributions over  $\{0, 1\}^n$ , and let  $s : \mathbb{N} \times (0, 1] \rightarrow \mathbb{N}$ . A tester, denoted  $T$ , of sample complexity  $s$  for the property  $\mathcal{D}$  is a probabilistic machine that, on input parameters  $n$  and  $\epsilon$ , and oracle access to a sequence of  $s = s(n, \epsilon)$  samples drawn from an unknown distribution  $P : \{0, 1\}^n \rightarrow [0, 1]$  outputs a verdict (“accept” or “reject”) that satisfies the following two conditions.

1. The tester accepts distributions that belong to  $\mathcal{D}$ : If  $P$  is in  $\mathcal{D}_n$ , then

$$\Pr_{x^{(1)}, \dots, x^{(s)} \sim P} [T^{x^{(1)}, \dots, x^{(s)}}(n, \epsilon) = 1] \geq 2/3,$$

where  $x^{(1)}, \dots, x^{(s)}$  are drawn independently from the distribution  $P$ .

2. The tester rejects distributions that are far from  $\mathcal{D}$ : If  $P$  is  $\epsilon$ -far from  $\mathcal{D}_n$  (i.e.,  $P$  is  $\epsilon$ -far from any distribution in  $\mathcal{D}_n$  (according to Definition 1.1)), then

$$\Pr_{x^{(1)}, \dots, x^{(s)} \sim P} [T^{x^{(1)}, \dots, x^{(s)}}(n, \epsilon) = 0] \geq 2/3,$$

where  $x^{(1)}, \dots, x^{(s)}$  are as in the previous item.

We say that  $q : \mathbb{N} \times (0, 1] \rightarrow \mathbb{N}$  is the query complexity of  $T$  if  $q(n, \epsilon)$  is the maximum number of queries that  $T$  makes on input parameters  $n$  and  $\epsilon$ . If the tester accepts every distribution in  $\mathcal{D}$  with probability 1, then we say that it has one-sided error.

We may assume, without loss of generality, that the tester queries each of its samples, and that it never makes the same query twice. Hence,  $q(n, \epsilon) \in [s(n, \epsilon), s(n, \epsilon) \cdot n]$ .

The **sample** (resp., **query**) **complexity of testing** the property  $\mathcal{D}$  (in the DoHO model) is the minimal sample (resp., query) complexity of a tester for  $\mathcal{D}$  (in the DoHO model). Note that the tester achieving the minimal sample complexity is not necessarily the one achieving the minimal query complexity. As stated before, we shall focus on minimizing the query complexity, while using the sample complexity as a yardstick.

**Generalization.** The entire definitional treatment can be extended to  $n$ -long sequences over an alphabet  $\Sigma$ , where above (in Definitions 1.1 and 1.2) we used  $\Sigma = \{0, 1\}$ .

## 1.2 The standard notions of testing as special cases (and other observations)

We first observe that both the standard model of property testing (of strings) and the standard model of distribution testing are special cases of Definition 1.2.

**Standard property testing (of strings):** Specifically, we refer to testing properties of  $n$ -bit strings (equiv., Boolean functions over  $[n]$ ).

This special case corresponds to trivial distributions, where each distribution is concentrated on a single  $n$ -bit long string. Hence, a standard tester of query complexity  $q$  can be viewed as a tester in the sense of Definition 1.2 that has sample complexity 1 and query complexity  $q$ .

**Standard distribution testing:** Specifically, we refer to testing distributions over  $\Sigma$ .

This special case corresponds to the case of  $n = 1$ , where each distribution is over  $\Sigma$ .

Hence, a standard distribution tester of sample complexity  $s$  can be viewed as a tester in the sense of Definition 1.2 that has sample complexity  $s$  and query complexity  $q = s$ .

Indeed, here we used the generalization of the definitional treatment to sequences over  $\Sigma$ .

The basic version, which refers to bit sequences, can be used too (with a small overhead).<sup>2</sup>

Needless to say, the point of this paper is going beyond these standard notions. In particular, we seek testers (for the DoHO model) with query complexity  $q(n, \epsilon) = o(n) \cdot s(n, \epsilon)$ , where  $s(n, \epsilon) > 1$  is the sample complexity in the DoHO model. Furthermore, our focus is on cases in which  $s(n, \epsilon)$  is relatively small (e.g.,  $s(n, \epsilon) = \text{poly}(n/\epsilon)$  and even  $s(n, \epsilon) = o(n) \cdot \text{poly}(1/\epsilon)$ ), since in these cases a factor of  $n$  matters more.

We mention that the sample complexity in the DoHO model is upper-bounded by the sample complexity in the standard distribution testing model. This is the case because the distance between pairs of distributions according to Definition 1.1 is upper-bounded by the total variation distance between them (see the discussion following Definition 1.1).

► **Observation 1.3.** (on the sample complexity of testing distributions in two models): *The sample complexity of testing a property  $\mathcal{D}$  of distributions over  $\{0, 1\}^n$  in the DoHO model is upper-bounded by the sample complexity of testing  $\mathcal{D}$  in the standard distribution testing model.*

<sup>2</sup> Specifically, we consider a good error correcting code  $C : \Sigma \rightarrow \{0, 1\}^n$  such that  $n = O(\log |\Sigma|)$ ; that is,  $C$  has distance  $\Omega(n)$ . In this case, the total variation distance between distributions over codewords is proportional to their distance according to Definition 1.1, whereas the query complexity is at most  $n = O(\log |\Sigma|)$  times the sample complexity. The same effect can be obtained by using larger  $n$ 's, provided we use locally testable and correctable codes.

We mention that for some properties  $\mathcal{D}$  the sample complexity in the DoHO model may be much lower than in the standard distribution testing model, because in these cases the distance measure in the DoHO model is much smaller than the total variation distance.<sup>3</sup> Needless to say, this is not true in general, and we shall focus on cases in which the two sample complexities are closely related. In other words, we are not interested in the possible gap between the sample complexities (in the two models), but rather in the query complexity in the DoHO model. Furthermore, we are willing to increase the sample complexity of a tester towards reducing its query complexity in the DoHO model (e.g., see our tester for uniformity).

### 1.3 Our Results

We present three types of results. The first type consists of general results that relate the query complexity of testing in the DoHO model to the query and/or sample complexity of related properties in the standard (distribution and/or string) testing models. The second type consists of results for properties that have been studied (some extensively) in the standard distribution testing model. The third type consists of results for new properties that arise naturally in the DoHO model. A few of these results are presented in Section 2, and the rest can be found in our report [10].

#### 1.3.1 Some general bounds on the query complexity of testing in the DoHO model

A natural class of properties of distribution over huge objects is the class of all distributions that are supported by strings that have a specific property (of strings). That is, for a property of bit strings  $\Pi = \{\Pi_n\}_{n \in \mathbb{N}}$  such that  $\Pi_n \subseteq \{0, 1\}^n$ , let  $\mathcal{D}_\Pi = \{\mathcal{D}_n\}_{n \in \mathbb{N}}$  such that  $\mathcal{D}_n$  denotes the set of all distributions that have a support that is subset of  $\Pi_n$ . We observe that the query complexity of testing the set of distributions  $\mathcal{D}_\Pi$  (in the DoHO model) is related to the query complexity of testing the set of strings  $\Pi$  (in the standard model of testing properties of strings).

► **Theorem 1.4.** (from testing strings for membership in  $\Pi$  to testing distributions for membership in  $\mathcal{D}_\Pi$ ): *If the query complexity of testing  $\Pi$  is  $q$ , then the query complexity of testing  $\mathcal{D}_\Pi$  in the DoHO model is at most  $q'$  such that  $q'(n, \epsilon) = \tilde{O}(1/\epsilon) \cdot q(n, \epsilon/2)$ .*

While the proof of Theorem 1.4 is simple, we believe it is instructive towards getting familiar with the DoHO model. We thus include it here, while mentioning that some ramifications of it appear in Appendix A.2 of our report [10].

**Proof.** The main observation is that if the tested distribution  $P$  (whose domain is  $\{0, 1\}^n$ ) is  $\epsilon$ -far from  $\mathcal{D}_n$  (according to Definition 1.1), then, with probability at least  $\epsilon/2$ , an object  $x$  selected according to  $P$  is  $\epsilon/2$ -H-far from  $\Pi_n$ . Hence, with high constant probability, a sample of size  $O(1/\epsilon)$  will contain at least one string that is  $\epsilon/2$ -H-far from  $\Pi_n$ . If we have a one-sided error tester  $T$  for  $\Pi$ , then we can detect this event (and reject) by running  $T$  (with

<sup>3</sup> An obvious case in which testing distributions is trivial (in the DoHO model) is the case of the set of all distributions that are supported by a set of strings  $\Pi$  such that any string is H-close to  $\Pi$ . Specifically, if every  $n$ -bit long string is  $\epsilon$ -H-close to  $\Pi \subseteq \{0, 1\}^n$  and  $\mathcal{D}$  is set of distributions that contain every distribution that is supported by  $\Pi$ , then every distribution is  $\epsilon$ -close to  $\mathcal{D}$ . Additional examples are presented in Section 2.2.

proximity parameter  $\epsilon/2$ ) on each sampled string. If we only have a two-sided error tester for  $\Pi$ , then we invoke it  $O(\log(1/\epsilon))$  times on each sample, and reject if the majority rule regarding any of these samples is rejecting. Hence, in total we make  $O(\epsilon^{-1} \log(1/\epsilon)) \cdot q(n, \epsilon/2)$  queries.  $\blacktriangleleft$

**An opposite extreme.** Theorem 1.4 applies to any property  $\Pi$  of strings and concerns the set of *all* distributions that are supported by  $\Pi$  (i.e., all distributions  $P$  that satisfy  $\{x : P(x) > 0\} \subseteq \Pi$ ). Hence, Theorem 1.4 focuses on the support of the distributions and pays no attention to all other aspect of the distributions. The other extreme is to focus on properties of distributions that are invariant under relabeling of the strings (i.e., label-invariant properties of distributions).<sup>4</sup> We consider several such properties in Section 1.3.2, but in the current section we seek more general results. Our guiding question is the following.

► **Open Problem 1.5.** (a key challenge, relaxed formulation):<sup>5</sup> *For which label-invariant properties of distributions does it hold that testing them in the DoHO model has query complexity  $\text{poly}(1/\epsilon) \cdot \tilde{O}(s(n, \epsilon/2))$ , where  $s$  is the sample complexity of testing them in the DoHO model?*

Jumping ahead, we mention that Theorem 1.9 identifies one property that satisfies the foregoing requirement and another that does not satisfy it. More generally, we show that the requirement is satisfied for any property that is closed under mapping, where a property of distribution  $\mathcal{D}$  is closed under mapping if, for every distribution  $P : \{0, 1\}^n \rightarrow [0, 1]$  in  $\mathcal{D}$  and every  $f : \{0, 1\}^n \rightarrow \{0, 1\}^n$ , it holds that  $f(P)$  is in  $\mathcal{D}$ , where  $Q = f(P)$  is the distribution defined by  $Q(y) = P(f^{-1}(y))$ .

► **Theorem 1.6.** (testing distributions that are closed under mapping (see Theorem 2.2)): *Suppose that  $\mathcal{D} = \{\mathcal{D}_n\}$  is testable with sample complexity  $s(n, \epsilon)$  in the DoHO model, and that each  $\mathcal{D}_n$  is closed under mapping. Then,  $\mathcal{D}$  is testable in the DoHO model with query complexity  $\tilde{O}(\epsilon^{-1} \cdot s(n, \epsilon/2))$ .*

We stress that the tester in the hypothesis may have query complexity  $n \cdot s(n, \epsilon)$ , and recall that our focus is on the case that  $n \gg \text{poly}(\epsilon^{-1} \log s(n, \epsilon/2))$ .

A middle ground between properties that contain all distributions that are supported by a specific set of strings and label-invariant properties of distributions is provided by properties of distributions that are label-invariant only on their support, where the **support of a property of distributions** is the union of the supports of all distributions in this property. That is, for a property  $\mathcal{D}_n$  of distributions over  $n$ -bit strings, we say that  $\mathcal{D}_n$  is **label-invariant over its support** if, for every bijection  $\pi : \{0, 1\}^n \rightarrow \{0, 1\}^n$  that *preserves the support of  $\mathcal{D}_n$*  (i.e.,  $x$  is in the support if and only if  $\pi(x)$  is in the support), it holds that the distribution  $P : \{0, 1\}^n \rightarrow [0, 1]$  is in  $\mathcal{D}_n$  if and only if  $\pi(P)$  is in  $\mathcal{D}_n$ . Indeed, generalizing Problem 1.5, one may ask

► **Open Problem 1.7.** (a more general challenge): *For which properties of distributions that are label-invariant over their support does it hold that testing them in the DoHO model has query complexity  $\text{poly}(1/\epsilon) \cdot \tilde{O}(s(n, \epsilon/2) \cdot q(n, \epsilon/2))$ , where  $s$  is the sample complexity of testing them in the DoHO model and  $q$  is the query complexity of testing their support?*

<sup>4</sup> Recall that a property of distributions over  $\{0, 1\}^n$  is called **label-invariant** if, for every bijection  $\pi : \{0, 1\}^n \rightarrow \{0, 1\}^n$  and every distribution  $P$ , it holds that  $P$  is in the property if and only if  $\pi(P)$  is in the property, where  $Q = \pi(P)$  is the distribution defined by  $Q(y) = P(\pi^{-1}(y))$ . We mention that label-invariant properties of distributions are often called symmetric properties.

<sup>5</sup> Less relaxed formulations may require query complexity  $\tilde{O}(s(n, \epsilon/2)/\epsilon)$  or even  $O(s(n, \epsilon))$ . On the other hand, one may ease the requirement by comparing the query complexity in the DoHO model to the sample complexity in the standard model.

The next theorem identifies a sufficient condition for a positive answer. Specifically, it requires that the support of the property, denoted  $S$ , has a (relaxed) self-correction procedure of query complexity  $q$ . We mention that such procedures may exist only in case the strings in  $S$  are pairwise far apart. Loosely speaking, on input  $i \in [n]$  and oracle access to an  $n$ -bit string  $x$ , the self-correction procedure is required to return  $x_i$  if  $x \in S$ , to reject if  $x$  is far from  $S$ , and otherwise it should either reject or return the  $i^{\text{th}}$  bit of the string in  $S$  that is closest to  $x$ .

► **Theorem 1.8.** (self-correction-based testers in the DoHO model, loosely stated (see Theorem 3.1 in our report [10])): *Let  $\mathcal{D}$  be a property of distributions over bit strings that is label-invariant over its support. Then, ignoring polylogarithmic factors, the query complexity of testing  $\mathcal{D}$  in the DoHO model is upper-bounded by the product of the sample complexity of testing  $\mathcal{D}$  in the standard model and the query complexity of testing and self-correcting the support of  $\mathcal{D}$ .*

One natural example to which Theorem 1.8 is applicable is a set of all distributions that are each have a support that contains few low-degree multi-variate polynomials; for size bound  $s(n)$  and the degree bound  $d(n)$ , we get query complexity  $\text{poly}(d(n)/\epsilon) \cdot \tilde{O}(s(n))$ .

### 1.3.2 Testing previously studied properties of distributions

Turning back to label-invariant properties of distributions, we consider several such properties that were studied previously in the context of the standard distribution testing model. Specifically, we consider the properties of having bounded support size (see, e.g., [11]), being uniform over a subset of specified size (see, e.g., [2]), and being  $m$ -grained (see, e.g., [9]).<sup>6</sup>

► **Theorem 1.9.** (testers for support size, uniformity, and  $m$ -grained in the DoHO model (see Corollary 2.3)): *For any  $m$ , the following properties of distributions over  $\{0, 1\}^n$  can be tested in the DoHO model using  $\text{poly}(1/\epsilon) \cdot \tilde{O}(m)$  queries:*

1. *All distributions having support size at most  $m$ .*
2. *All distributions that are uniform over some set of size  $m$ .*
3. *All distributions that are  $m$ -grained.*

Theorem 1.9 is proved by using Theorem 1.6. The foregoing upper bounds are quite tight. They also provide positive and negative cases regarding Problem 1.5 (see discussion following Theorem 1.10).

► **Theorem 1.10.** (lower bounds on testing support size, uniformity, and  $m$ -grained in the DoHO model (see Propositions 2.8, 2.10 and 2.9)):

1. *For every  $m \leq 2^{n-\Omega(n)}$ , testing whether a distribution over  $\{0, 1\}^n$  has support size at most  $m$  requires  $\Omega(m/\log m)$  samples.*
2. *For every constant  $c < 1$  and  $m \leq n$ , testing whether a distribution over  $\{0, 1\}^n$  is uniform over some subset of size  $m$  requires  $\Omega(m^c)$  queries.*
3. *For every constant  $c < 1$  and  $m \leq 2^{n-\Omega(n)}$ , testing whether a distribution over  $\{0, 1\}^n$  is  $m$ -grained requires  $\Omega(m^c)$  samples.*

Note that Parts 1 and 3 assert lower bounds on the *sample complexity* in the DoHO model, which imply the same lower bounds on the *query complexity* in this model. Combining the first part of Theorems 1.9 and 1.10 yields a property that satisfies the requirement of

<sup>6</sup> A distribution  $P : \{0, 1\}^n \rightarrow [0, 1]$  is called  $m$ -grained if any  $n$ -bit string appears in it with probability that is a multiple of  $1/m$ ; that is, for every  $x \in \{0, 1\}^n$  there exists an integer  $m_x$  such that  $P(x) = m_x/m$ .

Problem 1.5; that is, the query complexity in the DoHO model is closely related to the sample complexity (in this model). On the other hand, combining Part 2 of Theorem 1.10 with the tester of [2, 7] yields a property that does not satisfy the requirement Problem 1.5, since this tester uses  $O(m^{2/3}/\epsilon^2)$  samples (even in the standard distribution testing model).<sup>7</sup>

**Tuples of distributions.** In Section 5 of our report [10], we extend the DoHO model to testing tuples (e.g., pairs) of distributions, and consider the archetypal problem of testing equality of distributions (cf. [4, 5]). In this case, we obtain another natural property that satisfies the requirement of Problem 1.5.

► **Theorem 1.11.** (a tester for equality of distributions (see Theorem 5.2 in our report [10])):  
*For any  $m, n \in \mathbb{N}$  and  $\epsilon > 0$ , given a pair of distributions over  $\{0, 1\}^n$  that have support size at most  $m$ , we can distinguish between the case that the distributions are identical and the case that they are  $\epsilon$ -far from one another (according to Definition 1.1) using  $\tilde{O}(m^{2/3}/\epsilon^3)$  queries and  $O(m^{2/3}/\epsilon^2)$  samples.*

We note that  $m^{2/3}/\epsilon^2$  is a proxy for  $\max(m^{2/3}/\epsilon^{4/3}, m^{1/2}/\epsilon^2)$ , which is a lower bound on the sample complexity of testing this property in the standard distribution testing model [14]. This lower bound can be extended to the DoHO model. Hence, in this case, the query complexity in the DoHO model is quite close to the sample complexity in this model.

### 1.3.3 Distributions as variations of an ideal object

A natural type of distributions over huge objects arises by considering random variations of some ideal objects. Here we assume that we have no access to the ideal object, but do have access to a sample of random variations of this object, and we may be interested both in properties of the ideal object and in properties of the distribution of variations. In Section 4 of our report [10], we consider three types of such variations, and provide testers for the corresponding properties.

1. Noisy versions of a string, where we bound the noise level.  
 In this case it is easy to recover bits of the original string, and test that the noisy versions respect the predetermined noise level.
2. Random cyclic-shifts of a string.  
 In this case we use a tester of cyclic-shifts (i.e., given two strings the tester checks whether one is a cyclic shift of the other).
3. Random isomorphic copies of a graph represented by its adjacency matrix.  
 In this case we use an isomorphism tester.

We stress that the testers employed in the last two cases have sublinear complexity; specifically, pairs of  $n$ -bit long strings are tested using  $n^{0.5+o(1)}$  queries.

## 1.4 Orientation

As stated upfront, we seek testers that sample the distribution but do not read any of the samples entirely (and rather probe some of their bits).

---

<sup>7</sup> We mention that in [2, 7] the complexity bound is stated in terms of the second and third norms of the tested distribution, which can be roughly approximated by the number of samples required for seeing the first 2-way and 3-way collisions. To obtain complexity bounds in terms of  $m$ , we can take  $O(m^{2/3})$  samples and reject if no 3-way collision is seen (ditto for not seeing a 2-way collision among the first  $O(m^{1/2})$  samples).



In general, our proofs build on first principles, and are not technically complicated. Rather, each proof is based on one or few observations, which, once made, lead the way to obtaining the corresponding result. Hence, the essence of these proofs is finding the right point of view from which the observations arise.

**Upper bounds.** Some of our testers refer to label-invariant properties, and in this case it suffices to determine which samples are equal and which are different. Furthermore, viewing close samples as equal does not really create a problem, because we are working under Definition 1.1. Hence, testing equality between strings suffices, and it can be performed by probing few random locations in the strings. However, the analysis does not reduce to the foregoing comments, because we cannot afford to consider all strings in the (*a priori* unknown) support of the tested distribution. Instead, the analysis refers to the empirical distribution defined by the sequence of samples.

**Lower bounds.** Several of our lower bounds are obtained by transporting lower bounds from the standard distribution testing model. Typically, we transform distributions over an alphabet  $\Sigma$  to distributions over  $\{0, 1\}^n$  by using an error correcting code  $C : \Sigma \rightarrow \{0, 1\}^n$  that has constant relative distance (i.e.,  $\Delta_H(C(\sigma), C(\tau)) = \Omega(1)$  for every  $\sigma \neq \tau \in \Sigma$ ). For example, when proving a lower bound on testing the support size we transform a random variable  $Z$  that ranges over  $\Sigma$  to the random variable  $Z' = C(Z)$ . Note that in such a case it does not suffice to observe that if  $Z$  is TV-far from having a support of size at most  $m$ , then  $C(Z)$  is far (under Definition 1.1) from being supported on (at most)  $m$  codewords. We have to argue that  $C(Z)$  is far from being supported on *any* (subset of at most)  $m$  strings.

**Conventions.** As evident from the last paragraph, it is often convenient to treat distributions as random variables; that is, rather than referring to the distribution  $P : \Omega \rightarrow [0, 1]$  we refer to the random variable  $X$  such that  $\Pr[X = x] = P(x)$ . We stress that  $\epsilon$  always denotes the proximity parameter (for the testing task). Typically, the upper bounds specify the dependence on  $\epsilon$ , whereas the lower bound refer to some fixed  $\epsilon = \Omega(1)$ .

## 1.5 The current version and the full version

In the current version we present only the results that refer to a few natural properties of distributions that were studied previously in the context of the standard distribution testing model. These results also appear in Section 2 of our report [10].

The other sections of [10] are omitted; they include a proof of Theorem 1.8, a study (mentioned in Section 1.3.3) of the properties that capture random variations of some ideal objects, and an extension to testing tuples of distributions (including a proof of Theorem 1.11).

## 2 Support Size, Uniformity, and Being Grained

In this section we consider three natural types of label-invariant properties (of distributions). These properties refer to the support size, being uniform (over some subset), and being  $m$ -grained (i.e., each string appears with probability that is an integer multiple of  $1/m$ ). Recall that  $\mathcal{D}$  is a label-invariant property of distributions over  $\{0, 1\}^n$  if for every bijection  $\pi : \{0, 1\}^n \rightarrow \{0, 1\}^n$  and every distribution  $X$ , it holds that  $X$  is in  $\mathcal{D}$  if and only if  $\pi(X)$  is in  $\mathcal{D}$ . Label-invariant properties of distributions are of general interest and are also natural in the DoHO model, in which we wish to avoid reading samples in full. In this section we explore the possibility of obtaining testers for such properties.

We first present testers for these properties (in the DoHO model), and later discuss related “triviality results” and lower bounds. Our testers (for the DoHO model) are derived by emulating simple testers for the standard model (rather than emulating the best known such testers). The lower bounds justify this choice retroactively.

## 2.1 Testers

Our (DoHO-model) testers for support size, being uniform (over some subset), and being  $m$ -grained are obtained from a general result that refers to arbitrary properties (of distributions) that satisfy the following condition.

► **Definition 2.1.** (closure under mapping): *We say that a property  $\mathcal{D}$  of distributions over  $n$ -bit strings is closed under mapping if for every  $f : \{0, 1\}^n \rightarrow \{0, 1\}^n$  it holds that if  $X$  is in  $\mathcal{D}$  then  $f(X)$  is in  $\mathcal{D}$ .*

Note that closure under mapping implies being label-invariant (i.e., for every bijection  $\mu : \{0, 1\}^n \rightarrow \{0, 1\}^n$ , consider both the mapping  $\mu$  and  $\mu^{-1}$ ).

► **Theorem 2.2.** (testing distributions that are closed under mapping): *Suppose that  $\mathcal{D} = \{\mathcal{D}_n\}$  is testable with sample complexity  $s(n, \epsilon)$  in the DoHO model, and that each  $\mathcal{D}_n$  is closed under mapping. Then,  $\mathcal{D}$  is testable in the DoHO model with query complexity  $\tilde{O}(\epsilon^{-1} \cdot s(n, \epsilon/2))$ . Furthermore, the resulting tester uses  $3 \cdot s(n, \epsilon/2)$  samples, makes  $O(\epsilon^{-1} \log(s(n, \epsilon/2)/\epsilon))$  uniformly distributed queries to each sample, and preserves one-sided error of the original tester.*

The factor of 3 in the sample complexity is due to modest error reduction that is used to compensate for the small error that is introduced by our main strategy. Recall that a tester of sample complexity  $s$  in the standard distribution testing model constitutes a tester of sample complexity  $s$  in the DoHO model, alas this tester has query complexity  $n \cdot s$ .

**Proof.** The key observation is that, since  $\mathcal{D}$  is closed under mapping, for any  $\ell$ -subset  $J \subseteq [n]$ , it holds that if  $X$  is in  $\mathcal{D}$ , then  $X_J 0^{n-\ell}$  is in  $\mathcal{D}$ , whereas we can test  $X_J 0^{n-\ell}$  for membership in  $\mathcal{D}$  with  $\ell$  queries per sample. Even more importantly, for a typical  $\ell$ -subset  $J$ , we shall define a related random variable  $X'$  such that (i)  $X'_J \equiv X_J$ , (ii)  $X'$  is  $\epsilon/2$ -close to  $X$ , and (iii) the collision pattern of  $s = s(n, \epsilon/2)$  samples of  $X'_J$  is statistically close to the collision pattern of  $s$  samples of  $X'$ . Hence, applying the original tester to  $X_J 0^{n-\ell}$ , while setting the proximity parameter to  $\epsilon/2$ , yields the correct decision for  $X$  (in case  $X$  is either in  $\mathcal{D}$  or  $\epsilon$ -far from  $\mathcal{D}$ ).

**The actual tester.** Let  $T$  be the guaranteed tester of sample complexity  $s : \mathbb{N} \times [0, 1] \rightarrow \mathbb{N}$ . (Note that this tester operates in the DoHO model, but its query complexity is only bounded by  $n \cdot s$ .) Recall that we may assume, without loss of generality, that  $T$  is label-invariant, which means that it rules according to the collision pattern that it sees among its samples (i.e., the number of  $t$ -way collisions for each  $t \geq 2$ ). Using  $T$ , on input parameters  $n$  and  $\epsilon$ , for  $s = s(n, \epsilon/2)$ , given  $s$  samples, denoted  $x^{(1)}, \dots, x^{(s)}$ , that are drawn independently from a tested distribution  $X$ , we proceed as follows.

1. We select a set  $J \subseteq [n]$  of size  $\ell = O(\epsilon^{-1} \log(s/\epsilon))$  uniformly at random and query each of the samples at each location in  $J$ .
2. Invoking  $T$  with proximity parameter  $\epsilon/2$ , we output  $T'(x_J^{(1)}, \dots, x_J^{(s)})$ , where

$$T'(z^{(1)}, \dots, z^{(s)}) = T^{z^{(1)} 0^{n-\ell}, \dots, z^{(s)} 0^{n-\ell}}(n, \epsilon/2). \quad (3)$$

We emulate the execution of  $T^{x_J^{(1)}0^{n-\ell}, \dots, x_J^{(s)}0^{n-\ell}}(n, \epsilon/2)$  by answering  $T$ 's queries in a straightforward manner; that is, query  $j \in [n]$  to the  $i^{\text{th}}$  oracle is answered by the  $j^{\text{th}}$  bit of  $x_J^{(i)}$  if  $j \in [\ell]$  and by 0 otherwise. (Recall that  $x_J^{(i)}$  denotes the restriction of  $x^{(i)}$  to  $J$ .) As observed above, if  $X$  is in  $\mathcal{D}$ , then so is  $X_J0^{n-\ell}$ , for any choice of  $J$ . Hence, our tester accepts each distribution in  $\mathcal{D}$  with probability that is lower-bounded by the corresponding lower bound of  $T$ .

We now turn to the analysis of the case that  $X$  is  $\epsilon$ -far from  $\mathcal{D}$ . In this case, we proceed with a mental experiment in which we define, for each choice of  $J$ , a random variable  $X' = X'(J)$  such that (i)  $X'_J \equiv X_J$ , (ii)  $X'$  is  $\epsilon/2$ -close to  $X$ , and (iii) the collision pattern of  $s$  samples of  $X'_J$  is statistically close to the collision pattern of  $s$  samples of  $X'$ . Note that Condition (ii) implies that  $X'$  is  $\epsilon/2$ -far from  $\mathcal{D}$ , which means that  $T$  should reject  $s$  samples of  $X'$  (whp), Condition (iii) implies that  $T$  should also reject  $s$  samples of  $X'_J0^{n-\ell}$  (whp), whereas Condition (i) implies that the same holds for samples of  $X_J0^{n-\ell}$ , which in turn means that our tester rejects  $X$  (whp). In order to materialize the foregoing plan, we need a few terms.

**Terms and initial observations.** For integers  $\ell \leq n$  and  $s$ , and a generic random variable  $X$  that ranges over  $\{0, 1\}^n$ , we consider a sufficiently large  $s' = O(s^2 \cdot \ell)$ , and use the following terms.

- For an  $\ell$ -subset  $J$ , we say that  $\sigma \in \{0, 1\}^\ell$  is  *$J$ -heavy (w.r.t  $X$ )* if  $\Pr[X_J = \sigma] \geq \frac{0.01}{s^2}$ .
- For an  $\ell$ -subset  $J$ , we say that a sequence of  $s'$  strings  $(w^{(1)}, \dots, w^{(s')}) \in (\{0, 1\}^\ell)^{s'}$  is  *$J$ -good (for  $X$ )* if for every  $J$ -heavy string  $\sigma$  there exists  $i \in [s']$  such that  $w_J^{(i)} = \sigma$ . Note that, for every  $J$ ,

$$\Pr_{w^{(1)}, \dots, w^{(s')} \sim X}[(w^{(1)}, \dots, w^{(s')}) \text{ is } J\text{-good}] = 1 - o(1),$$

because  $2^\ell \cdot (1 - 0.01/s^2)^{s'} = 2^\ell \cdot \exp(-\Omega(s'/100s^2)) = o(1)$ .

- We say that  $(w^{(1)}, \dots, w^{(s')})$  is *good (for  $X$ )* if it is  $J$ -good for a  $1 - o(1)$  fraction of the  $\ell$ -subsets  $J$ 's. Note that

$$\Pr_{w^{(1)}, \dots, w^{(s')} \sim X}[(w^{(1)}, \dots, w^{(s')}) \text{ is good}] = 1 - o(1).$$

In fact, we shall only use the fact that there exists a good sequence of  $w^{(i)}$ 's.

We fix an arbitrary good (for  $X$ ) sequence  $(w^{(1)}, \dots, w^{(s')})$  for the rest of the proof.

Recall that, with probability  $1 - o(1)$  over the choice of  $J \in \binom{[n]}{\ell}$ , it holds that  $(w^{(1)}, \dots, w^{(s')})$  is  $J$ -good (for  $X$ ), which means that *all  $J$ -heavy strings (w.r.t  $X$ ) appear among the  $J$ -restrictions of the  $w^{(i)}$ 's*. Fixing such a set  $J$ , let  $I = I(J)$  be a maximal set of indices  $i \in [s']$  such that the  $w_J^{(i)}$ 's are distinct; that is,  $R \stackrel{\text{def}}{=} \{w_J^{(i)} : i \in I\}$  has size  $|I|$  and equals  $\{w_J^{(i)} : i \in [s']\}$ . We stress that  $R$  contains all  $J$ -heavy strings (w.r.t  $X$ ), which means that for every  $\sigma \notin R$  it holds that  $\Pr[X_J = \sigma] < 0.01/s^2$ . We now define  $X'$  by selecting  $x \sim X$ , and outputting  $w^{(i)}$  if  $x_J = w_J^{(i)}$  for some  $i \in I$ , and outputting  $x$  itself otherwise (i.e., if  $x_J \notin R$ ); that is,

$$\Pr[X' = x] = \begin{cases} \Pr[X_J = w_J^{(i)}] & \text{if } x = w^{(i)} \text{ for } i \in I \\ 0 & \text{if } x_J \in \{w_J^{(i)} : i \in I\} \text{ and } x \notin \{w^{(i)} : i \in I\} \\ \Pr[X = x] & \text{if } x_J \notin \{w_J^{(i)} : i \in I\} \end{cases} \quad (4)$$

Note that  $X'_J \equiv X_J$ . We claim that, for a typical  $J$ , it holds that  $X'$  is  $\epsilon/2$ -close to  $X$ .

The key observation is that  $X'$  differs from  $X$  only when  $X_J \in \{w_J^{(i)} : i \in I(J)\} = \{w_J^{(i)} : i \in [s']\}$ . In this case, strings that are  $\epsilon/4$ -H-close to  $\{w^{(i)} : i \in I\}$  contribute at most  $\epsilon/4$  units (to the distance between  $X$  and  $X'$  (as in Definition 1.1)), and so we upper-bound

## 78:12 Testing Distributions of Huge Objects

the probability mass of strings  $x \sim X$  that are  $\epsilon/4$ -H-far from  $\{w^{(i)} : i \in I(J)\}$  but satisfy  $x_J \in \{w_J^{(i)} : i \in [s']\} = R$ . Denoting this bad event by  $\text{Bad}_x(J)$ , we have

$$\begin{aligned} \Pr_{J,X}[\text{Bad}_X(J)] &= \mathbb{E}_{x \sim X}[\Pr_{J \in \binom{[n]}{\ell}}[\text{Bad}_x(J)]] \\ &\leq \sum_{i \in [s']} \mathbb{E}_{x \sim X} \left[ \Pr_J \left[ x \text{ is } \epsilon/4\text{-far from } w^{(i)} \text{ and } x_J = w_J^{(i)} \text{ and } i \in I(J) \right] \right] \\ &\leq \sum_{i \in [s']} \mathbb{E}_{x \sim X} \left[ \Pr_J \left[ x \text{ is } \epsilon/4\text{-far from } w^{(i)} \text{ and } x_J = w_J^{(i)} \right] \right] \\ &\leq s' \cdot (1 - (\epsilon/4))^\ell, \end{aligned}$$

which is  $o(\epsilon)$  by the definition of  $\ell = O(\epsilon^{-1} \log(s/\epsilon))$  (and  $s' = \tilde{O}(s^2/\epsilon)$ ). Hence, with probability  $1 - o(1)$  over the choice of  $J$ , it holds that the probability that  $x \sim X$  is  $\epsilon/4$ -H-far from  $\{w^{(i)} : i \in [s']\}$  but satisfies  $x_J \in R$  is at most  $\epsilon/4$ . It follows that, with probability  $1 - o(1)$  over the choice of  $J$ , it holds that  $X'$  is  $\epsilon/2$ -close to  $X$ . Recalling that  $X$  is  $\epsilon$ -far from  $\mathcal{D}$ , this implies that  $X'$  is  $\epsilon/2$ -far from  $\mathcal{D}$ , which implies that (with probability at least  $2/3$ ), the tester  $T$  rejects  $X'$  (i.e., rejects when fed with  $s$  samples selected according to  $X'$ ). However, we are interested in the probability that our tester (rather than  $T$ ) rejects  $X$  (rather than  $X'$ ).

We say that  $J$  is useful if  $(w^{(1)}, \dots, w^{(s)})$  is  $J$ -good (for  $X$ ) and  $\Pr_{x \sim X}[\text{Bad}_x(J)] \leq \epsilon/4$ , and recall that each of the two events holds with probability  $1 - o(1)$  (over the choice of  $J \in \binom{[n]}{\ell}$ ). Recalling that  $X'_J = X_J$ , while relying on the hypothesis that  $(w^{(1)}, \dots, w^{(s)})$  is  $J$ -good (for  $X$ ), we observe that the probability that our tester rejects  $X$  equals

$$\begin{aligned} &\Pr_{x^{(1)}, \dots, x^{(s)} \sim X} [T'(n, \epsilon/2; x_J^{(1)}, \dots, x_J^{(s)}) = 0] \\ &= \Pr_{x^{(1)}, \dots, x^{(s)} \sim X'} [T'(n, \epsilon/2; x_J^{(1)}, \dots, x_J^{(s)}) = 0] && \text{[using } X'_J = X_J \text{]} \\ &= \Pr_{x^{(1)}, \dots, x^{(s)} \sim X'} [T^{x_J^{(1)} 0^{n-\ell}, \dots, x_J^{(s)} 0^{n-\ell}}(n, \epsilon/2) = 0] && \text{[definition of } T' \text{]} \\ &= \Pr_{x^{(1)}, \dots, x^{(s)} \sim X'} [T^{x^{(1)}, \dots, x^{(s)}}(n, \epsilon/2) = 0] \pm \frac{\binom{s}{2}}{100 \cdot s^2} && \text{[from } x_J^{(\cdot)} \text{'s to } x^{(\cdot)} \text{'s]} \end{aligned}$$

where the approximate equality is justified as follows (based on the definition of  $X'$ ).

- On the one hand, the equality-relations between samples of  $X'$  with a  $J$ -restriction in  $R$  are identical to those of their  $J$ -restrictions, because for each  $\sigma \in R$  there is a unique  $x$  in the support of  $X'_J$  such that  $x_J = \sigma$  (i.e.,  $x = w^{(i)}$  such that  $w_J^{(i)} = \sigma$ ).
- On the other hand, the probability of collision among the  $J$ -restrictions of the other samples (i.e., those with a  $J$ -restriction in  $\{0, 1\}^\ell \setminus R$ ) is upper-bounded by  $\frac{\binom{s}{2}}{100 \cdot s^2} < 0.005$ , since these  $J$ -restrictions are all non-heavy. Needless to say, the collision probability between these (other) samples themselves can only be smaller.

It follows that our tester rejects with probability at least  $(1 - o(1)) \cdot (\frac{2}{3} - 0.005) > 0.66$ , where the first factor represents the probability that  $J$  is useful and the second factor lower-bounds the probability that  $T$  rejects when presented with  $s$  samples of  $X'$ . Using mild error reduction (via three experiments), the theorem follows. ◀

► **Corollary 2.3.** (testers for support size, uniformity, and  $m$ -grained in the DoHO model):  
*For any  $m$ , the following properties of distributions over  $\{0, 1\}^n$  can be tested in the DoHO model using  $\text{poly}(1/\epsilon) \cdot \tilde{O}(m)$  queries:*

1. All distributions having support size at most  $m$ .

*Furthermore, the tester uses  $O(m/\epsilon)$  samples, makes  $O(\epsilon^{-1} \log(m/\epsilon))$  queries to each sample, and has one-sided error.*

2. All distributions that are uniform over some set of size  $m$ .

Furthermore, the tester uses  $O(\epsilon^{-2}m \log m)$  samples, and makes  $q = O(\epsilon^{-1} \log(m/\epsilon))$  queries to each sample if  $\epsilon \geq \frac{2^{\lceil \log_2 m \rceil}}{n}$ , and  $O(q^2)$  queries otherwise.

3. All distributions that are  $m$ -grained.

Furthermore, the tester uses  $O(\epsilon^{-2}m \log m)$  samples, and makes  $O(\epsilon^{-1} \log(m/\epsilon))$  queries to each sample.

Moreover, all testers make the same uniformly distributed queries to each of their samples.

**Proof.** For Parts 1 and 3 we present testers for the standard model and apply Theorem 2.2, whereas for Part 2 we observe that the tester for  $m$ -grained distributions will do. (Recall that a tester of sample complexity  $s$  in the standard distribution testing model constitutes a tester of sample complexity  $s$  in the DoHO model.)

Let us start with Part 2. The key observation is that any distribution that is uniform over some  $m$ -subset is  $m$ -grained, whereas any distribution that is  $m$ -grained is  $\frac{\lceil \log_2 m \rceil}{n}$ -close (under Definition 1.1) to being uniform over some set of  $m$  elements (e.g., by modifying the first  $\lceil \log_2 m \rceil$  bits in each string in the support).<sup>8</sup> Hence, for  $\epsilon > 2 \cdot \frac{\lceil \log_2 m \rceil}{n}$ , we test uniformity over  $m$ -subsets by testing for being  $m$ -grained (using proximity parameter  $\epsilon/2$ ). If  $\epsilon \leq \frac{2^{\lceil \log_2 m \rceil}}{n}$ , then we can afford reading entirely each sample, since  $n = O(\epsilon^{-1} \log m)$ . In the latter case we make  $O(\epsilon^{-2} \log^2 n)$  (rather than  $O(\epsilon^{-1} \log n)$ ) queries to each sample.

Turning to Parts 1 and 3, it is tempting to use known (standard model) testers of complexity  $O(\epsilon^{-2}m/\log m)$  for these properties (cf. [13]), while relying on the fact that these properties are label-invariant. However, these bounds hold only when the tested distribution ranges over a domain of size  $O(m)$ , and so some additional argument is required. Furthermore, this may not allow us to argue that the tester for support-size has one-sided error. Instead, we present direct (standard model) testers of sample complexity  $O(m/\epsilon)$  and  $\tilde{O}(m/\epsilon^2)$ , respectively.

**Testing support size.** On input parameters  $n$  and  $\epsilon$ , given  $s = O(m/\epsilon)$  samples, denoted  $x^{(1)}, \dots, x^{(s)}$ , that are drawn independently from a tested distribution  $X$ , we accept if and only if  $|\{x^{(i)} : i \in [s]\}| \leq m$ . Suppose that  $X$  is  $\epsilon$ -TV-far from having support size at most  $m$ , and note that for any set  $S$  of at most  $m$  strings it holds that  $\Pr[X \notin S] > \epsilon$ . Then, for each  $t \in [s-1]$ , either  $W_t = \{x^{(i)} : i \in [t]\}$  has size exceeding  $m$  or  $\Pr[x^{(t+1)} \notin W_t] > \epsilon$ . It follows that  $\Pr[|W_s| \leq m] = \exp(-\Omega(m))$ .

**Testing the set of  $m$ -grained distributions.** On input parameters  $n$  and  $\epsilon$ , we set  $s = O(m \log m)$  and  $s' = O(\epsilon^{-2}m \log m)$ . Given  $s + s'$  samples, denoted  $x^{(1)}, \dots, x^{(s+s')}$ , that are drawn independently from a tested distribution  $X$ , we proceed in two steps.

1. We construct  $W = \{w^{(i)} : i \in [s]\}$ , the set of strings seen in the first  $s$  samples.

(We may reject if  $|W| > m$ , but this is inessential.)

2. For each  $w \in W$ , we approximate  $\Pr[X = w]$  by  $p_w \stackrel{\text{def}}{=} |\{i \in [s'] : x^{(s+i)} = w\}|/s'$ . We reject if we either encountered a sample not in  $W$  or one of the  $p_w$ 's is not within a  $1 \pm 0.1\epsilon$  factor of a positive integer multiple of  $1/m$ .

<sup>8</sup> Saying that  $X$  is  $m$ -grained means that it is uniform on a multiset  $\{x^{(1)}, \dots, x^{(m)}\}$  of  $n$ -bit strings. We modify  $X$  by replacing each  $x^{(i)}$  by  $y^{(i)}$  such that  $y^{(i)}$  encodes the binary expansion of  $i-1$  in the first  $\ell = \lceil \log_2 m \rceil$  locations and equals  $x^{(i)}$  otherwise. That is, we set  $y_j^{(i)}$  to equal the  $j^{\text{th}}$  bit in the binary expansion of  $i-1$  if  $j \in [\ell]$ , and  $y_j^{(i)} = x_j^{(i)}$  otherwise (i.e., if  $j \in \{\ell+1, \dots, n\}$ ).

Note that if  $X$  is  $m$ -grained, then, with high probability,  $W$  equals the support of  $X$ , and (whp) each of the  $p_w$ 's is within a  $1 \pm 0.1\epsilon$  factor of a positive integer multiple of  $1/m$ . On the other hand, suppose that  $X$  is accepted with high probability. Then, for any choice of  $W$  (as determined in Step 1), for each  $w \in W$ , it holds that  $\Pr[X = w] = (1 \pm 0.1\epsilon) \cdot p_w$ , since  $p_w$  is within a  $(1 \pm 0.1\epsilon)$  factor of a positive integer multiple of  $1/m$ . Furthermore,  $\Pr[X \notin W] < 0.1\epsilon$ . It follows that  $X$  is  $\epsilon$ -TV-close to being  $m$ -grained. ◀

## 2.2 Triviality results

An obvious case in which testing is trivial is the property of all distributions (on  $n$ -bit strings) that have support size  $2^n$ . In this case, each distribution is infinitesimally close (under Definition 1.1) to being supported on all  $2^n$  strings. A less obvious result is stated next.

► **Observation 2.4.** (triviality of testing  $2^n$ -grained distributions in the DoHO model): *Under Definition 1.1, every distribution over  $\{0, 1\}^n$  is  $O(\frac{\log n}{n})$ -close to being  $2^n$ -grained.*

**Proof.** We first show that, for every  $\ell \in \mathbb{N}$ , it holds that every distribution over  $\{0, 1\}^n$  is  $\frac{\ell}{n}$ -close to a distribution that is supported by  $\{0, 1\}^{n-\ell}0^\ell$ . Next we show that each distribution of the latter type is  $2^{-\ell}$ -close to being  $2^n$ -grained. Letting  $\ell = \lfloor \log_2 n \rfloor$ , the claim follows.

In the first step, given an arbitrary distribution  $X$ , we consider the distribution  $X'$  obtained by setting the last  $\ell$  bits of  $X$  to zero; that is, let  $\Pr[X' = x'0^\ell] = \sum_{x'' \in \{0, 1\}^\ell} \Pr[X = x'x'']$ . Then,  $X'$  is  $(\ell/n)$ -close to  $X$  (according to Definition 1.1).

In the second step, we consider  $X''$  obtained by letting  $\Pr[X'' = x'0^\ell]$  equal  $2^{-n} \cdot [2^n \cdot \Pr[X' = x'0^\ell]]$ , and assigning the residual probability to (say)  $1^n$ . Then,  $X''$  is  $2^n$ -grained and is at total variation distance at most  $2^{n-\ell} \cdot 2^{-n} = 2^{-\ell}$  from  $X'$ , since the support size of  $X'$  is at most  $2^{n-\ell}$ . Hence,  $X''$  is  $(\frac{\ell}{n} + 2^{-\ell})$ -close to  $X$ . ◀

**Non-triviality results.** It is easy to see that any property of distributions that includes only distributions having a support of size  $2^{n-\Omega(n)}$  is non-trivial in the sense that *not* all distributions are close to it under Definition 1.1. This is the case because any such distribution is far from the uniform distribution over  $\{0, 1\}^n$  (since, w.h.p., a uniformly distributed  $n$ -bit string is at Hamming distance  $\Omega(n)$  from a set that contains  $2^{n-\Omega(n)}$  strings). Additional non-triviality results follow from the lower bounds presented in Section 2.3.

## 2.3 Lower bounds

We first consider three notions of uniformity: Uniformity over the entire potential support (i.e., all  $n$ -bit strings), uniformity over the support of the distribution (where the size of the support is not specified), and uniformity over a support of a specified size. In all three cases (as well as in the results regarding testing support size and the set of grained distributions), we prove lower bounds on the sample (and query) complexity of testing the corresponding property in the DoHO model. As usual, the lower bounds refer to testing with  $\epsilon = \Omega(1)$ ; that is, to the case that the proximity parameter is set to some positive constant. Our proofs rely on two standard simplifying assumptions:

1. When considering the task of testing a label-invariant property, one may assume, without loss of generality, that the tester is label-invariant [1] (see also [8, Thm. 11.12]); that is, for every bijection  $\pi$  on the potential support, the tester's verdict on the samples  $x^{(1)}, \dots, x^{(s)}$  is identical to its verdict on the samples  $\pi(x^{(1)}), \dots, \pi(x^{(s)})$ .

2. To prove a lower bound of  $L$  on the complexity of testing, it suffices to show two distributions  $X$  and  $Y$  that an algorithm of complexity  $L - 1$  cannot distinguish (with constant positive gap)<sup>9</sup> such that  $X$  has the property and  $Y$  is  $\Omega(1)$ -far from having the property (cf. [8, Thm. 7.2]).

Combining these two observations, we focus on presenting distributions that cannot be distinguished by label-invariant algorithms of low complexity such that one distribution has the property while the other is  $\Omega(1)$ -far from having the property.

► **Observation 2.5.** (lower bound on testing uniformity over  $\{0, 1\}^n$ ): *For every  $c \in (0, 0.5)$  there exists  $\epsilon > 0$  such that testing with proximity parameter  $\epsilon$  whether a distribution is uniform over  $\{0, 1\}^n$  requires  $2^{c \cdot n}$  samples in the DoHO model.*

**Proof.** Let  $S$  be an arbitrary  $2^{2c \cdot n}$ -subset of  $\{0, 1\}^n$ , and  $X$  be uniform over  $S$ . Then, a sample of  $s = o(2^{cn})$  strings does not allow for (label-invariant) distinguishing between  $X$  and the uniform distribution over  $\{0, 1\}^n$ ; that is, for every label-invariant decision procedure  $D : (\{0, 1\}^n)^s \rightarrow \{0, 1\}$ , it holds that

$$\Pr_{x^{(1)}, \dots, x^{(s)} \in S} [D(x^{(1)}, \dots, x^{(s)}) = 1] = \Pr_{x^{(1)}, \dots, x^{(s)} \in \{0, 1\}^n} [D(x^{(1)}, \dots, x^{(s)}) = 1] \pm o(1).$$

On the other hand, for every  $S$  as above, it holds that  $X$  is  $\Omega(1)$ -far from the uniform distribution over  $\{0, 1\}^n$  (according to Definition 1.1). This is the case because the probability mass of each  $x$  in the support of  $X$  must be distributed among  $2^n / 2^{2cn}$  strings, whereas most of these strings are at relative Hamming distance at least  $\epsilon = \Omega(1)$  from the support of  $X$  (provided that  $\epsilon$  is chosen such that  $H_2(\epsilon) < 1 - 2c$ ). ◀

► **Observation 2.6.** (lower bound on testing uniformity over an unspecified support size): *For every  $c \in (0, 0.5)$  there exists  $\epsilon > 0$  such that testing with proximity parameter  $\epsilon$  whether a distribution is uniform over some set requires  $2^{c \cdot n}$  samples in the DoHO model.*

**Proof.** We consider the following two families of distributions, where each of the distributions is parameterized by an  $2^{2c \cdot n}$ -subset of  $n$ -bit strings, denoted  $S$ .

1.  $X_S$  is uniform on  $S$ .
2. With probability half,  $Y_S$  is uniform on  $S$ , and otherwise it is uniform on  $\bar{S} \stackrel{\text{def}}{=} \{0, 1\}^n \setminus S$ . Now, on the one hand, a label-invariant algorithm cannot distinguish  $X_S$  from  $Y_S$  by using  $o(2^{cn})$  samples. On the other hand, we prove that  $Y_S$  is far from being uniform on any set. Suppose that  $Y = Y_S$  is  $\delta$ -close to a distribution that is uniform on the set  $S' \subseteq \{0, 1\}^n$ . We shall show that  $\delta = \Omega(1)$ , by considering two cases regarding  $S'$ :

**Case 1:**  $|S'| \leq 2^{(0.5+c) \cdot n}$  (recall that  $c < 0.5$ ). In this case, the probability mass assigned by  $Y$  to  $\bar{S} \setminus S'$  should be moved to  $S'$ , whereas the average relative Hamming distance between a random element of  $\bar{S} \setminus S'$  and the set  $S'$  is  $\Omega(1)$ . Specifically, letting  $U_n$  denote the uniform distribution on  $\{0, 1\}^n$ , we upper-bound the probability that  $U_n \in \bar{S} \setminus S'$  is H-close to  $S'$  by noting that  $|\bar{S} \setminus S'| > 2^{n-1}$ , since  $|S| + |S'| = o(2^n)$ , whereas  $|S'| \leq 2^{(0.5+c) \cdot n} = 2^{n - \Omega(n)}$ .

**Case 2:**  $|S'| > 2^{(0.5+c) \cdot n}$ . In this case, almost all the probability assigned by  $Y$  to  $S$  should be distributed among more than  $2^{(0.5+c) \cdot n}$  strings such that each of these strings is assigned equal weight. This implies that almost all the weight assigned by  $Y$  to  $S$  must be moved to strings that are at Hamming distance  $\Omega(n)$  from  $S$ , since  $|S| = 2^{2cn} = 2^{(0.5+c) \cdot n - \Omega(n)} < 2^{-\Omega(n)} \cdot |S'|$ .

<sup>9</sup> We say that  $A$  distinguishes  $s$  samples of  $X$  from  $s$  samples of  $Y$  with gap  $\gamma$  if

$$|\Pr_{z_1, \dots, z_s \sim X} [A(z_1, \dots, z_s) = 1] - \Pr_{z_1, \dots, z_s \sim Y} [A(z_1, \dots, z_s) = 1]| \geq \gamma.$$

Hence, in both cases, a significant probability weight of  $Y$  must be moved to strings that are  $\Omega(1)$ -H-far from their origin. The claim follows.  $\blacktriangleleft$

► **Observation 2.7.** (lower bound on testing parameterized uniformity, grained, and support size): *For every  $m \leq 2^{n-\Omega(n)}$ , the following testing tasks regarding properties of distributions over  $\{0,1\}^n$  require  $\Omega(\sqrt{m})$  samples in the DoHO model:*

- *The set of distributions that are uniform over some  $m$ -subset;*
- *The set of  $m$ -grained distributions;*
- *The set of distributions with support size at most  $m$ .*

Stronger results are presented in Propositions 2.8 and 2.9.

**Proof.** A label-invariant algorithm cannot distinguish the uniform distribution over  $\{0,1\}^n$  from a distribution that is uniform over an  $m$ -subset unless it sees  $\Omega(\sqrt{m})$  samples. However, the uniform distribution over  $\{0,1\}^n$  is far from any of the foregoing properties (also under Definition 1.1), since  $m \leq 2^{n-\Omega(n)}$ .  $\blacktriangleleft$

► **Proposition 2.8.** (lower bound on testing parameterized support size): *For every  $m \leq 2^{n-\Omega(n)}$ , testing that a distribution over  $\{0,1\}^n$  has support size at most  $m$  requires  $\Omega(m/\log m)$  samples in the DoHO model.*

**Proof.** We use the  $\Omega(m/\log m)$  (sample complexity) lower bound of [12] that refers to testing distributions over  $[O(m)]$  for support size at most  $m$ , in the standard testing model (that is, under the total variation distance). This lower bound is proved in [12] by presenting two distributions,  $X$  and  $Y$ , that cannot be distinguished by a label-invariant algorithm that gets  $s = o(m/\log m)$  samples, where  $X$  has support size at most  $m$  and  $Y$  is far (in total variation distance) from having support size at most  $m$ . We use an error correcting code  $C : [O(m)] \rightarrow \{0,1\}^n$  of constant relative distance, and consider the distributions  $X' = C(X)$  and  $Y' = C(Y)$ .

Evidently, a label-invariant algorithm that obtains  $m$  samples cannot distinguish  $X'$  and  $Y'$ . On the other hand,  $X'$  has support size at most  $m$  whereas we claim that  $Y'$  is far from having support size at most  $m$ , under Definition 1.1. Intuitively, this is the case because reducing the support size of  $Y'$  requires moving a constant amount of probability weight from elements in the support of  $Y'$ , which resides on strings that are far away in Hamming distance, to fewer strings. Each such movement can be charged in proportion to the relative distance of the code  $C$ . The actual argument follows.

Let  $Z$  be a distribution that is closest to  $Y'$ , under Definition 1.1, among all distributions that are supported on at most  $m$  strings, and let  $\gamma$  denote the distance between  $Y'$  and  $Z$ . By Definition 1.1, this means that there exists a “weight relocation” function  $W : \{0,1\}^{2n} \rightarrow [0,1]$  that satisfies  $\sum_z W(y',z) = \Pr[Y'=y']$  for every  $y'$ , and  $\sum_{y'} W(y',z) = \Pr[Z=z]$  for every  $z$ . Furthermore,  $\sum_{y'} \sum_z W(y',z) \cdot \Delta_H(y',z) = \gamma$ , where we refer to this sum as the *cost* associated with  $W$ . Note that  $\sum_{y'} \sum_z W(y',z) \cdot \text{InEq}(y',z)$  is lower-bounded by the total variation distance between  $Y'$  and  $Z$ , where  $\text{InEq}(y',z) = 1$  if  $y' \neq z$  and  $\text{InEq}(y',y') = 0$ .

Let  $S$  denote the support of  $Z$  (so that  $W(y',z) = 0$  for every  $z \notin S$ ), and let  $S'$  be the subset of  $S$  that contains those strings that are  $(0.4 \cdot \delta)$ -H-close to the code  $C$ . Recall that the support of  $Y'$  is a subset of  $C$  (so that  $W(y',z) = 0$  for every  $y' \notin C$ ). The cost associated with  $W$  is the sum of three terms. The first is  $\sum_{y'} \sum_{z \in S \setminus S'} W(y',z) \cdot \Delta_H(y',z)$ , the second is  $\sum_{y'} \sum_{z \in S' \setminus C} W(y',z) \cdot \Delta_H(y',z)$  and the third is  $\sum_{y'} \sum_{z \in S' \cap C} W(y',z) \cdot \Delta_H(y',z)$ . We analyze each separately, while letting  $R$  denote the support of  $Y'$ .



- By the definition of  $S'$  (and since the support of  $Y'$  is a subset of  $C$ ), for each  $y'$  in the support of  $Y'$  and each  $z \in S \setminus S'$ , we have that  $\Delta_H(y', z) > 0.4 \cdot \delta$ . Therefore, the first term is lower-bounded by  $\sum_{y'} \sum_{z \in S \setminus S'} W(y', z) \cdot 0.4 \cdot \delta$ .
- Turning to the second term, for each  $z \in S' \setminus C$ , let  $\text{cc}(z) \in C$  be the codeword in  $C$  that is closest to  $z$ . By the definition of  $S'$  we have that  $\delta'(z) \stackrel{\text{def}}{=} \Delta_H(\text{cc}(z), z) \leq 0.4 \cdot \delta$ , and for every  $y' \in R \setminus \{\text{cc}(z)\}$ , we have that  $\Delta_H(y', z) \geq \delta - \delta'(z) \geq 0.6 \cdot \delta$ .

We claim that (for every  $z \in S' \setminus C$ ), at least half the probability mass that is relocated by  $W$  to  $z$  (from  $Y'$ ) must come from codewords  $y'$  (in the support of  $Y'$ ) that are different from  $\text{cc}(z)$ ; that is,  $\sum_{y' \in R \setminus \{\text{cc}(z)\}} W(y', z) \geq \frac{1}{2} \cdot \sum_{y'} W(y', z)$ . We prove that  $\sum_{y' \in R \setminus \{\text{cc}(z)\}} W(y', z) \geq W(\text{cc}(z), z)$ , by showing that otherwise we could modify  $Z$  (and  $W$ ) to obtain a distribution  $Z'$  with support size at most  $m$  (and a corresponding weight relocation function  $W'$ ) such that  $Z'$  is closer to  $Y'$  than  $Z$  (i.e.,  $W'$  has lower cost than  $W$ ).

Specifically,  $Z'$  is obtained by moving the probability mass that  $Z$  assigns  $z$  to the codeword  $\text{cc}(z)$ ; that is,  $\Pr[Z' = z] = 0$  and  $\Pr[Z' = \text{cc}(z)] = \Pr[Z = \text{cc}(z)] + \Pr[Z = z]$  (and  $\Pr[Z' = z'] = \Pr[Z = z']$  for every  $z' \notin \{z, \text{cc}(z)\}$ ), while noting that  $Z'$  has support size at most  $m$ . The weight relocation function  $W'$  is define accordingly (i.e., for each  $y'$ , we set  $W'(y', z) = 0$  and  $W'(y', \text{cc}(z)) = W(y', \text{cc}(z)) + W(y', z)$  (leaving  $W'(y', z') = W(y', z')$  for every  $z' \notin \{z, \text{cc}(z)\}$ )). Then, the cost of  $W'$  (which upper-bounds the distance between  $Y'$  and  $Z'$ ) equals the cost of  $W$  minus  $\sum_{y'} W(y', z) \cdot \Delta_H(y', z)$  plus  $\sum_{y'} W(y', z) \cdot \Delta_H(y', \text{cc}(z))$ . Now,

$$\sum_{y' \in R} W(y', z) \cdot \Delta_H(y', z) \geq W(\text{cc}(z), z) \cdot \delta'(z) + \sum_{y' \in R \setminus \{\text{cc}(z)\}} W(y', z) \cdot (\delta - \delta'(z)), \quad (5)$$

since for  $y' \in R \setminus \{\text{cc}(z)\}$  it holds that  $\Delta_H(y', z) \geq \Delta_H(y', \text{cc}(z)) - \Delta_H(z, \text{cc}(z)) \geq \delta - \delta'(z)$ , whereas

$$\sum_{y' \in R} W(y', z) \cdot \Delta_H(y', \text{cc}(z)) \leq \sum_{y' \in R \setminus \{\text{cc}(z)\}} W(y', z) \cdot \delta. \quad (6)$$

Using the counter hypothesis (i.e.,  $W(\text{cc}(z), z) > \sum_{y' \in R \setminus \{\text{cc}(z)\}} W(y', z)$ ), we lower-bound Eq. (5) by  $\sum_{y' \in R \setminus \{\text{cc}(z)\}} W(y', z) \cdot \delta$ , and reach a contradiction to the optimality of  $W$  (since the cost of  $W'$  is smaller than the cost of  $W$ ).

Hence, for each  $z \in S \setminus C$  we have that  $\sum_{y'} W(y', z) \cdot \Delta_H(y', z) \geq \frac{1}{2} \sum_{y'} W(y', z) \cdot 0.6 \cdot \delta$ , implying that second term in the cost of  $W$  is lower-bounded by  $\sum_{y'} \sum_{z \in S' \setminus C} W(y', z) \cdot 0.3 \cdot \delta$ .

- Lastly, for each  $y'$  in the support of  $Y'$  and each  $z \in S' \cap C$  such that  $z \neq y'$ , we have that  $\Delta_H(y', z) \geq \delta$ . Therefore, the third term is lower-bounded by  $\sum_{y'} \sum_{z \in (S' \cap C) \setminus \{y'\}} W(y', z) \cdot \delta$ , which we rewrite as  $\sum_{y'} \sum_{z \in (S' \cap C)} W(y', z) \cdot \text{InEq}(y', z) \cdot \delta$ . To summarize, the distance  $\gamma$  between  $Y'$  and  $Z$ , under Definition 1.1, is at least a  $0.3\delta$  factor of the total variation distance between these two distributions. ◀

► **Proposition 2.9.** (lower bound on testing  $m$ -grained distributions): *For every constant  $c < 1$  and  $m \leq 2^{n - \Omega(n)}$ , testing that a distribution over  $\{0, 1\}^n$  is  $m$ -grained requires  $\Omega(m^c)$  samples in the DOHO model.*

We comment that the foregoing lower bound (for DOHO model) matches the best known lower bound for the standard distribution testing model [9]. See Section 2.4 of our report [10] for further discussion.

**Proof.** We use the  $\Omega(m^c)$  lower bound of [9] that refers to testing whether a distribution over  $[O(m)]$  is  $m$ -grained, under the total variation distance. This lower bound is proved in [9] by presenting two ( $2m$ -grained) distributions,  $X$  and  $Y$ , that cannot be distinguished by a label-invariant algorithm that gets  $s = o(m^c)$  samples, where  $X$  is  $m$ -grained and  $Y$  is far (in total variation distance) from being  $m$ -grained.

As in the proof of Proposition 2.8, applying an error correcting code  $C : [O(m)] \rightarrow \{0, 1\}^n$  to  $X$  and  $Y$ , we observe that  $X' = C(X)$  is  $m$ -grained whereas  $Y' = C(Y)$  is far from being  $m$ -grained (also under Definition 1.1). To see that  $Y'$  is far from any distribution  $Z$  that is  $m$ -grained and is supported by a set  $S$ , we (define  $S'$  and) employ the same case-analysis as in the proof of Proposition 2.8. (This shows that the distance (under Definition 1.1) between  $Y'$  and  $Z$  is lower-bounded by a constant fraction of their total variation distance.)<sup>10</sup> ◀

► **Proposition 2.10.** (lower bound on testing parameterized uniformity): *For every constant  $c < 1$  and  $m \leq n$ , testing that a distribution over  $\{0, 1\}^n$  is uniform over some  $m$ -subset requires  $\Omega(m^c)$  queries in the DOHO model.*

We stress that, unlike Proposition 2.9, which lower-bounds the sample complexity of testers, in Proposition 2.10 we only lower-bound their query complexity.<sup>11</sup>

**Proof.** Let  $X'$  and  $Y'$  denote the distributions derived in the proof of Proposition 2.9. Recall that  $X'$  is  $m$ -grained, whereas  $Y'$  is far from being  $m$ -grained (under Definition 1.1). Note that  $Y'$  is  $\Omega(1)$ -far from being uniform over any set of size  $m$ , and observe that  $X'$  is  $\frac{\log_2 m}{n}$ -close to a distribution  $X''$  that is uniform over a set of size  $m$ . Specifically, we can transform  $X'$  to  $X''$  by modifying only the bits that reside in  $\log_2 m$  locations, where the choice of these locations is arbitrary.<sup>12</sup> Hence, a potential tester that make  $o(n/\log m)$  queries is unlikely to hit these locations, if we select these locations uniformly at random. Using  $m \leq n$ , we conclude that a potential tester that makes  $\min(o(m^c), o(n/\log m)) = o(m^c)$  queries cannot distinguish between the distribution  $X''$  and distribution  $Y'$ , which implies that it fails to test uniformity in the DOHO model. ◀

**Conditional lower bounds.** The lower bounds (for the DOHO model) presented in Proposition 2.9 and 2.10 build on the best known lower bound for testing the set of grained distributions in the standard distribution testing model. In Section 2.4 of our report [10], we present stronger (i.e., almost-linear) lower bounds on the complexity of testing in the DOHO model that rely on an analogous conjecture regarding the sample complexity of testing grained distributions in the standard model.

<sup>10</sup>Note that in the second case (i.e., probability mass relocated from  $Y'$  to  $z \in S' \setminus C$ ), the potential replacement (of  $z$  by the codeword closest to it) preserves  $m$ -grained-ness.

<sup>11</sup>We actually use  $m \log m = o(n^{1/c})$ , which follows from  $m \leq n$ .

<sup>12</sup>Saying that  $X'$  is  $m$ -grained means that it is uniform on a multiset  $\{x^{(1)}, \dots, x^{(m)}\}$  of  $n$ -bit strings. We modify  $X'$  by replacing each  $x^{(i)}$  by  $y^{(i)}$  such that  $y^{(i)}$  encodes the binary expansion of  $i - 1$  in the chosen locations and equals  $x^{(i)}$  otherwise. That is, letting  $\ell_1 < \ell_2 < \dots < \ell_{\log_2 m}$  denote the chosen locations, we set  $y_{\ell_j}^{(i)}$  to equal the  $j^{\text{th}}$  bit in the binary expansion of  $i - 1$  and set  $y_{\ell}^{(i)} = x_{\ell}^{(i)}$  if  $\ell \in [n] \setminus \{\ell_1, \ell_2, \dots, \ell_{\log_2 m}\}$ .

---

**References**

---

- 1 Tugkan Batu. *Testing properties of distributions*. PhD thesis, Computer Science department, Cornell University, 2001.
- 2 Tugkan Batu and Clement L. Canonne. Generalized uniformity testing. In *Proceedings of the Fiftieth-Eighth Annual Symposium on Foundations of Computer Science (FOCS)*, pages 880–889, 2017.
- 3 Tugkan Batu, Lance Fortnow, Eldar Fischer, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of the Forty-Second Annual Symposium on Foundations of Computer Science (FOCS)*, pages 442–451, 2001.
- 4 Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *Proceedings of the Forty-First Annual Symposium on Foundations of Computer Science (FOCS)*, pages 259–269, 2000.
- 5 Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *Journal of the ACM*, 60(1):4:1–4:25, 2013. This is a long version of [4].
- 6 Clément L. Canonne. *A Survey on Distribution Testing: Your Data is Big. But is it Blue?* Number 9 in Graduate Surveys. Theory of Computing Library, 2020. doi:10.4086/toc.gs.2020.009.
- 7 Ilias Diakonikolas, Daniel Kan, and Alistair Stewart. Sharp bounds for generalized uniformity testing. Technical Report TR17-132, Electronic Colloquium on Computational Complexity (ECCC), 2017.
- 8 Oded Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017.
- 9 Oded Goldreich and Dana Ron. Lower bounds on the complexity of testing grained distributions. Technical Report TR21-129, Electronic Colloquium on Computational Complexity (ECCC), 2021.
- 10 Oded Goldreich and Dana Ron. Testing distributions of huge objects. Technical Report TR21-133, Electronic Colloquium on Computational Complexity (ECCC), 2021.
- 11 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.
- 12 Gregory Valiant and Paul Valiant. Estimating the unseen: an  $n/\log(n)$ -sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the Fourty-Third Annual ACM Symposium on the Theory of Computing (STOC)*, pages 685–694, 2011.
- 13 Gregory Valiant and Paul Valiant. Estimating the unseen: Improved estimators for entropy and other properties. *Journal of the ACM*, 64(6), 2017.
- 14 Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011.