Testing Distributions of Huge Objects*

Oded Goldreich† Dana Ron‡

May 23, 2023

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), and 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 $\widetilde{O}(m/\epsilon^3)$, 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 $\widetilde{O}(m^{2/3}/\epsilon^3)$.

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.

*An extended abstract of this work appeared in the proceedings of *ITCS22*. Partially supported by the Israel Science Foundation (grant No. 1041/18).

†Department of Computer Science, Weizmann Institute of Science, Rehovot, Israel. E-mail: oded.goldreich@weizmann.ac.il. Additional funding received from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819702).

‡School of Electrical Engineering, Tel Aviv University, Tel Aviv, Israel. E-mail: danaron@tau.ac.il

Contents

1 Introduction

In the last couple of decades, the area of property testing has attracted much attention (see, e.g., a recent textbook [15]). Loosely speaking, property testing typically refers to sub-linear time probabilistic algorithms for *approximate decision* making; that is, distinguishing between objects that have a predetermined property and ones that are 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 [15, Chap. 1-10]), but not the part that deals with testing properties of distributions (aka *distribution testing*, see [15, Chap. 11] and [8]). In this context, a tester get samples from the tested distribution and sub-linearity means sub-linearity in the size of the distribution's domain.[1] Each element in the domain is considered to be small, and is assumed to be processed at unit time, and the distance between distributions is defined as the total variation distance.

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 distance. Under this combined measure, the distance between distributions reflects the probability mass that "needs to be moved" when weighted according to the Hamming distance between strings (see Definition 1.1). Thus, a probability mass of p moved from a string x (in one distribution) to a string y (in the other distribution) contributes $p \cdot \delta$ units to the distance between the distributions, where δ is the relative Hamming distance between x and y.

Recall that the basic paradigm of distribution testing postulates that, in many settings, it suffices to distinguish distributions that satisfy the property (of interest) from ones that are far from satisfying it; that is, it postulates that a notion of approximate decision is meaningful. In fact, this basic paradigm underlies all of statistics. Here, we suggest that, in the case of distributions on huge objects, a distance measure that is more refined than total variation distance is more adequate. Specifically, *the suggested distance measure accounts for the distance between elements of the distributions' domain rather than viewing them as atoms.* Indeed, this suggestion goes hand in hand with a more refined notion of inspecting an element of the domain (or of the sample); it allows to partially inspect an element rather than forcing us to either look at the entire element or not look at it at all.

We mention that, while we follow the archetypical convention of property testing in using the Hamming distance as the distance between strings (i.e., objects), our framework allows to use any other distance between strings as a basis for defining distance between distributions on huge objects (i.e., long strings). For example, one may consider the edit distance between strings. Indeed, standard distribution testing is obtained as a special case by considering the equality function as a distance between strings.

To illustrate the flavor of our study, consider the problem of testing whether a distribution

[1] 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 [8, Part IV].

over n-bit strings has support size m (resp., is uniform over m strings). When n-bit strings are considered as atoms, the results of distribution testing apply and assert that, when ignoring the (polynomial) dependence on the proximity parameter, the sample complexity of this problem is $\Theta(m/\log m)$ [22, 23] (resp., $\Theta(m^{2/3})$ [3]). But the resulting tester requires entirely reading each of the samples (i.e., an n-bit long string), which means reading $\Theta(nm/\log m)$ bits. The question we address is whether the query complexity may be reduced (under the definitions outlined above and detailed next). (Jumping ahead, the answer is positive: Both properties can be tested using $\widetilde{O}(m)$ queries.)

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 this will be 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. The minimum cost of such a transformation (of P to Q) is captured by Eq. (2); an alternative formulation appears in Eq. (15) (in Appendix A.1).

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,*

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

For two distributions $P, Q : \Omega \to [0,1]$, 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:*

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

We say that P is ϵ-close to Q if the optimal value of the linear program is at most ϵ; otherwise, we say that P is ϵ-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 Ω is always defined on top of a distance measure that is associated

with Ω. It is well known that the earth mover's distance with respect to the inequality function (i.e., $\mathrm{InEq}(x,y) = 1$ if $x \neq y$ and $\mathrm{InEq}(x,x) = 0$) coincides with the variation distance (between these distributions). That is, if we replace the distance $\Delta_H(x,y)$ with $\mathrm{InEq}(x,y)$ in Definition 1.1, then we get the variation distance between P and Q (see Appendix A.1). Furthermore, $\Delta_H(x,y) \leq \mathrm{InEq}(x,y)$ always holds. Hence, the distance between pairs of distributions according to Definition 1.1 is upper-bounded by the total variation distance between them.

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 ϵ-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 ϵ-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 $s : \mathbb{N} \times (0,1] \to \mathbb{N}$. A* tester*, denoted T,* of sample complexity s for the property $\mathcal{D}$ *is a probabilistic machine that, on input parameters n and ϵ, and oracle access to a sequence of $s = s(n,\epsilon)$ samples drawn from an unknown distribution $P : \{0,1\}^n \to [0,1]$ 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*

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

   where $x^{(1)},\ldots,x^{(s)}$ are drawn independently from the distribution P, and the probability is taken ver these draws as well as over the internal random choices of T.
2. The tester rejects distributions that are far from $\mathcal{D}$: *If P is ϵ-far from $\mathcal{D}_n$* (i.e., P is ϵ-far from any distribution in $\mathcal{D}_n$ (according to Definition 1.1))*, then*

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

   where $x^{(1)},\ldots,x^{(s)}$ and the probability space are as in the previous item.

We say that $q : \mathbb{N} \times (0,1] \to \mathbb{N}$ is the query complexity of T *if $q(n,\epsilon)$ is the maximum number of queries that T makes on input parameters n and ϵ. 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 Σ, 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 Σ.

This special case corresponds to the case of $n = 1$, where each distribution is over Σ. 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 Σ. The basic version, which refers to bit sequences, can be used too (with a small overhead).[2]

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., $s(n,\epsilon) = \mathrm{poly}(n/\epsilon)$ and even $s(n,\epsilon) = o(n) \cdot \mathrm{poly}(1/\epsilon)$), 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). We state this observation for future reference.

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.*

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.[3] Needless to say, this is

[2] Specifically, we consider a good error correcting code $C : \Sigma \to \{0,1\}^n$ 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.

[3] An obvious case is one in which testing distributions is trivial (in the DoHO model). This is the case for the set of all distributions that are supported by a set of strings Π such that any string is H-close to Π. Specifically, if every n-bit long string is ϵ-H-close to $\Pi \subseteq \{0,1\}^n$ and $\mathcal{D}$ is set of distributions that contain every distribution that is supported by Π, then every distribution is ϵ-close to $\mathcal{D}$. On the other hand, testing $\mathcal{D}$ in the standard model may be non-trivial. Additional examples are presented in Section 2.2.

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), although this is a legitimate question too, 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.

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 Π_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 Π (in the standard model of testing properties of strings).

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

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.

Proof: The main observation is that if the tested distribution P (whose domain is $\{0,1\}^n$) is ϵ-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 Π_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 Π_n. If we have a one-sided error tester T for Π, then we can detect this event (and reject) by running T (with proximity parameter $\epsilon/2$) on each sampled string. If we only have a two-sided error tester for Π, 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. ■

An opposite extreme. Theorem 1.4 applies to any property Π of strings and concerns the set of *all* distributions that are supported by Π (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).[4] We

[4]Recall that a property of distributions over $\{0,1\}^n$ is called label-invariant if, for every bijection $\pi : \{0,1\}^n \to \{0,1\}^n$ and every distribution P, it holds that P is in the property if and only if $\pi(P)$ is in the property, where

consider several such specific 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):[5] *For which label-invariant properties of distributions does it hold that testing them in the* DoHO *model has query complexity* $\mathrm{poly}(1/\epsilon) \cdot \widetilde{O}(s(n,\epsilon/2))$*, where s is the sample complexity of testing them in the* DoHO *model?*

Jumping ahead, we mention that in Section 1.3.2 we identify two label-invariant properties for which the relation between the query complexity and the sample complexity is as sought in Problem 1.5, and one for which this relation does not hold. More generally, we show that a relaxed form of such a relation (in which s is the sample complexity in the standard model) 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 $P : \{0,1\}^n \to [0,1]$ in $\mathcal{D}$ and every $f : \{0,1\}^n \to \{0,1\}^n$, 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 standard model, and that each $\mathcal{D}_n$ is closed under mapping. Then, $\mathcal{D}$ is testable in the* DoHO *model with query complexity $\widetilde{O}(\epsilon^{-1} \cdot s(n,\epsilon/2))$.*

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 $n \cdot s$ (whereas our focus is on the case that $n \gg \mathrm{poly}(\epsilon^{-1} \log s(n,\epsilon/2))$). We wonder whether a result similar to Theorem 1.6 holds when s is the sample complexity in the DoHO model.[6]

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 $\pi : \{0,1\}^n \to \{0,1\}^n$ 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 $P : \{0,1\}^n \to [0,1]$ 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* $\mathrm{poly}(1/\epsilon) \cdot \widetilde{O}(s(n,\epsilon/2) \cdot (q(n,\epsilon/2)+1))$*, where s is the sample complexity of testing them in the* DoHO *model and q is the query complexity of testing their support?*

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

$Q = \pi(P)$ is the distribution defined by $Q(y) = P(\pi^{-1}(y))$. We mention that label-invariant properties of distributions are often called symmetric properties.

[5] Less relaxed formulations may require query complexity $\widetilde{O}(s(n,\epsilon/2)/\epsilon)$ 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.

[6] Such a result was wrongly claimed in Revision 1 of our ECCC TR21-133 (and in the extended abstract that appeared in the proceedings of *ITCS22*). Partial progress towards such a result is presented in Appendix A.3.

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^{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)): *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; specifically, for support-size bound $s(n)$ and the degree bound $d(n)$, we get query complexity $\text{poly}(d(n)/\epsilon) \cdot \widetilde{O}(s(n))$. Note that the dependency on the support size is quite tight in light of Theorem 1.10 (see the proof of Proposition 2.8).

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., [21]), being uniform over a subset of specified size (see, e.g., [3]), and being m-grained (see, e.g., [17]).[7]

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 $\text{poly}(1/\epsilon) \cdot \widetilde{O}(m)$ 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 (see Theorem 1.10). 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.*

[7] A distribution $P : \{0,1\}^n \to [0,1]$ 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 $P(x) = m_x/m$.

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 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 [3, 11] yields a property that does not satisfy the requirement in Problem 1.5, since this tester uses $O(m^{2/3}/\epsilon^2)$ samples (even in the standard distribution-testing model).[8]

Tuples of distributions. In Section 5 we extend the DoHO model to testing tuples (e.g., pairs) of distributions, and consider the archetypical problem of testing equality of distributions (cf. [5, 6]). 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)): *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 ϵ-far from one another* (according to Definition 1.1) *using $\widetilde{O}(m^{2/3}/\epsilon^3)$ 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 [24]. 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. We consider three types of such variations, and provide testers for the corresponding properties.

Noisy versions of a string, where we bound the noise level. Specifically, we consider a noise model in which each bit may be flipped with some bounded probability $\eta < 1/2$ such that the resulting string is at relative Hamming distance at most δ from the original string. In this case it is easy to recover bits of the original string, and test that the noisy versions respect the predetermined noise model. Furthermore, we can test properties of the original string with rather small overhead.

Random cyclic-shifts of a string. Specifically, we consider distributions obtained by applying arbitrary random cyclic-shift to a fixed n-bit string as well as distributions obtained by shifting a fixed n-bit string by a uniformly distributed number of positions. The first property (i.e., random cyclic-shifts) is tested using $O(1/\epsilon)$ samples and $\widetilde{O}(\sqrt{n}/\epsilon)$ queries (see Theorem 4.3), whereas for testing the second property (i.e., uniformly random cyclic-shifts) we use $\widetilde{O}(n^{2/3}/\epsilon)$ queries (see

[8] We mention that in [3, 11] 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).

Theorem 4.4). The first tester uses a tester of cyclic-shifts (i.e., given two strings, the tester checks whether one is a cyclic shift of the other), whereas the second testing problem is reduced to testing equality between distributions (as addressed in Theorem 1.11). In the latter case, the two distribution that we test for equality are the given distribution and the distribution defined by a uniform random shift of a single sample obtained from the given distribution. (The reduction is requires a generalization of Theorem 1.11 to the case that only the support size of one of the distributions is bounded.)

Random isomorphic copies of a graph (represented by its adjacency matrix). Specifically, here the n-bit string represents the adjacency matrix of an $\sqrt{n}$-vertex graph, and the tested property is an arbitrary distribution over isomorphic copies of some fixed graph. In this case we get a tester of query complexity $\mathrm{poly}(1/\epsilon) \cdot \widetilde{O}(n^{5/8})$, by employing a known tester for graph isomorphism [13].

1.4 Orientation and Techniques

As stated upfront, we seek testers that sample the distribution but do not read any of the samples entirely (but rather probe some of their bits). In other words, we seek testers for distributions over n-bit strings of query complexity that is $o(n)$ times the sample complexity of testing these distributions in the standard model.

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.

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

1.4.1 Our testers

Some of our testers refer to label-invariant properties (see Theorems 1.6 and 1.9). The key observation here is that, in this case, it suffices to determine which samples are equal and which are different. Furthermore, considering close samples as if they were equal does not really create a problem, because we are working under Definition 1.1. Hence, checking approximate equality between strings suffices, and it can be performed by probing few random locations in the strings.

Unfortunately, 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; that is, we cannot argue that the collision pattern among the projections (at few random locations) correctly represent the distances between all pairs of strings in the support of the distribution. Instead, the analysis refers to the empirical distribution defined by a sequence of samples; that is, we show that the collision pattern among these projections correctly represent the distances between all pairs of *sampled* strings. This is done by considering a mental experiment in which the tested distribution is replaced by an imaginary distribution that is close to it. This strategy is illustrated next.

Illustration: The case of testing support size. Suppose that we wish to test whether a distribution X has support size at most m. Our starting point is the observation that *if X is ϵ-far from having support size at most m* (according to Definition 1.1), *then taking $s = O(m/\epsilon)$ samples from X yields* (with high probability) *a set that contains more than m strings that are $\epsilon/2$-H-far apart from one another*. In this case, with high probility, the projection of the s samples on $\ell = O(\epsilon^{-1} \log m)$ random locations would yield more than m different ℓ-bit strings. Hence, given oracle access to the samples $x^{(1)}, ..., x^{(s)}$, our tester proceeds as follows.

1. Selects uniformly an ℓ-subset J of $[n]$.
2. For every $i \in [s]$, queries $x^{(i)}$ at the locations $j \in J$, obtaining $y^{(i)} = x^{(i)}_J$, where $x^{(i)}_J$ denotes the restriction of $x^{(i)}$ to the locations in J.
3. Accepts if and only if $|\{y^{(i)} : i \in [s]\}| \leq m$.

Clearly if X has support size at most m, then the foregoing tester always accepts. On the other hand, as shown next, if X is ϵ-far from having support size at most m, then, with very high probability over the choice of the s samples, there exists a set $I \subset [s]$ of size $m + 1$ such that for every $i_i \neq i_2 \in I$ the relative Hamming distance between $x^{(i_1)}$ and $x^{(i_2)}$ is at least $\epsilon/2$.

To prove the foregoing claim, we consider a mental experiment in which we try to find $m + 1$ strings that are pairwise far apart in $m+1$ sampling phases. For $i \in [m]$, suppose that $w^{(1)}, \ldots, w^{(i)}$ are pairwise $\epsilon/2$-H-far apart strings that were found in the first i phases. Then, assuming that X is ϵ-far from having support size at most m, it must be that at least $\epsilon/2$ of the probability weight of X resides on strings that are $\epsilon/2$-H-far from each of the $w^{(j)}$'s (for all $j \in [i]$). Hence, a sampled string taken from X is $\epsilon/2$-H-far from all these $w^{(j)}$'s with probability at least $\epsilon/2$. Once such a sampled string is found, we denote it $w^{(i+1)}$, and proceed to the next phase. It follows that, with probability $1 - \exp(-\Omega(m))$, the process is completed successfully (i.e., $m + 1$ strings are found) while using $O(m/\epsilon)$ samples.

1.4.2 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 Σ to distributions over $\{0,1\}^n$ by using an error correcting code $C : \Sigma \to \{0,1\}^n$ 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 Σ to the random variable $C(Z)$. Clearly, if Z has support size at most m, then so does $C(Z)$. On the other hand, 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*. However, we need to show that $C(Z)$ is far from being supported on *any* (subset of at most) m strings. Specifically, we need to show that transporting a probability mass from more than m codewords to at most m strings requires a constant fraction of this mass to be transported over a constant (relative Hamming) distance. Intuitively, this is the case because trasporting a probability mass to the vicinity of a codeword does not reduce the support size unless a similar (or larger) mass is trasported from a different codeword.

1.5 Related work

Our work is informed by the vast literature on distribution testing (see [15, Chap. 11] and [8]) and on property testing at large (see [15]). Here we only mention a prior model that refers to partial

information obtained from individual samples, and a subsequent work that studies the DoHO model.

1.5.1 A different model of partial information from samples

A sequence of works, initiated by [1], studies a model in which one may obtain only partial information regarding each of the samples drawn from the distribution. Specifically, the tester (or learner) may obtain only $\ell \geq 1$ "bits of information" from each of the samples; however, unlike in the DoHO model, these ℓ bits may be an arbitrary function of the entire sample (rather than actual bits in the representation of the sample), which is actually viewed as an atom, and the distance between distributions is the total variation distance (rather than Definition 1.1).

1.5.2 Subsequent work

Following the initial posting of our work [18], a study of *index invariant* properties in the DoHO model, was initiated in [9], where a property $\mathcal{D}$ (of distributions over n-bit strings) is called index invariant if a distribution $X = X_1 \cdots X_n$ is in $\mathcal{D}$ if and only if, for every permutation $\phi : [n] \to [n]$ (on the indices of strings in $\{0,1\}^n$), it holds that $X_{\phi(1)} \cdots X_{\phi(n)}$ is in $\mathcal{D}$. Note that the class of index-invariant properties extends the class of label-invariant properties, since the invariance condition is much weaker: It only refers to relabling of strings by rearranging their bits (i.e., it only considers permutations $\pi : \{0,1\}^n \to \{0,1\}^n$ that satisfy $\pi(x_1 \cdots x_n) = x_{\phi(1)} \cdots x_{\phi(n)}$ for some permutation $\phi : [n] \to [n]$).

One notable result of [9] is that such properties can be tested within query complexity that is upper-bounded by a function of the VC-dimension (and the proximity paramter). They also show that their complexity bounds are tight, that index-invariance is essential for these bounds, and that there exists a (tight) quadratic gap between adaptive and non-adaptive testers of such properties.

1.6 Organization

We start, in Section 2, with results that refer to a few natural properties of distributions that were studied previously in the context of the standard distribution-testing model. Specifically, we present the testers asserted in Theorems 1.6 and 1.9, and establish the lower bounds claimed in Theorem 1.10.

Next, we turn to the general result captured by Theorem 1.8, and present its proof in Section 3. In Section 4 we study several types of distributions that arise naturally in the context of the DoHO model; that is, we consider distributions that capture random variations of some ideal objects. Lastly, in Section 5, we extend our treatment to testing tuples of distributions, and present a tester for the set of pairs of identical distributions (proving 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 $\pi : \{0,1\}^n \to \{0,1\}^n$ 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 testers for the standard (distribution testing) model. 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 $f : \{0,1\}^n \to \{0,1\}^n$ 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 $\pi : \{0,1\}^n \to \{0,1\}^n$, consider both the mapping π and π^{-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 standard model, and that each $\mathcal{D}_n$ is closed under mapping. Then, $\mathcal{D}$ is testable in the* DoHO *model with query complexity $\widetilde{O}(\epsilon^{-1} \cdot s(n,\epsilon/2))$. Furthermore, the resulting tester uses $3 \cdot s(n,\epsilon/2)$ samples, makes* (the same) *$O(\epsilon^{-1}\log(s(n,\epsilon/2)/\epsilon))$ 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 ℓ-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 ℓ queries per sample. Furthermore, as shown below, if X is ϵ-far from $\mathcal{D}$, then, for a random set $J \subseteq [n]$ of size $\ell = O(\epsilon^{-1}\log(s/\epsilon))$, the original tester would have rejected $X_J 0^{n-\ell}$, when invoked with proximity parameter $\epsilon/2$. The foregoing claim relies on the fact that we may assume, without loss of generality, that the original tester (which operates in the standard model) rules according to the collsion pattern among the samples that it gets. Hence, it suffices to show that the collision pattern among samples of X_J is statistically close to the collision pattern of samples drawn from a distribution X' that is $\epsilon/2$-far from $\mathcal{D}$. Let us first spell out the proposed tester.

Our starting point is the guaranteed tester T, which operates in the standard distribution-testing model. Hence, we may assume, without loss of generality, that T is *label-invariant* (see, e.g., [15, Thm. 11.12]), 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 $t \geq 2$). In particular, if the collision pattern of s samples of X_J is statistically close to the collision pattern of s samples of X' that is $\epsilon/2$-far from $\mathcal{D}$, then T will reject $X_J 0^{n-\ell}$. (Indeed, in the analysis, we shall present such an X'.)

The actual tester. Let T be the guaranteed tester of sample complexity $s : \mathbb{N} \times [0,1] \to \mathbb{N}$. Using T, on input parameters n and ϵ, when given $s = s(n, \epsilon/2)$, samples, denoted $x^{(1)}, ..., x^{(s)}$, 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. Hence, we obtain $x^{(1)}_J, ..., x^{(s)}_J$.

   (Recall that $x^{(i)}_J$ denotes the restriction of $x^{(i)}$ to J.)

2. Invoking T with proximity parameter $\epsilon/2$, we output $T'(x^{(1)}_J, ..., x^{(s)}_J)$, where

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

   That is, we invoke T on s samples of the distribution $X_J 0^{n-\ell}$, where these s samples are obtained by padding the strings $x^{(1)}_J, ..., x^{(s)}_J$ obtained in Step 1.

As observed upfront, if X is in $\mathcal{D}$, then so is $X_J 0^{n-\ell}$, 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. In particular, if T has one-sided error, then so does our tester.

We now turn to the analysis of the case that X is ϵ-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 for most J's it holds 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 $X'_J 0^{n-\ell}$ (whp), whereas Condition (i) implies that the same holds for samples of $X_J 0^{n-\ell}$, which in turn means that our tester rejects X (whp). In order to materialize the foregoing plan, we need a few definitions.

Definitions 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 definitions.

- For any fixed ℓ-subset J, we say that $\sigma \in \{0,1\}^\ell$ is J-heavy (w.r.t X) if $\Pr[X_J = \sigma] \geq 0.01/s^2$.

  (The definition of J-heavy is aimed at guaranteeing that the probability that a collision on a non-J-heavy string occurs among the J-restrictions of s samples of X is small; that is, denoting the set of non-J-heavy strings by L, it holds that $\Pr_{x^{(1)},...,x^{(s)} \sim X}[\exists i \neq j \text{ s.t. } x^{(i)}_J = x^{(j)}_J \in L] < 0.005$.)

- For any fixed ℓ-subset J, we say that a sequence of s' strings $(w^{(1)}, ..., w^{(s')}) \in (\{0,1\}^n)^{s'}$ is J-good (for X) if its J-restrictions hit all J-heavy strings; that is, for every J-heavy string σ there exists $i \in [s']$ such that $w^{(i)}_J = \sigma$.

  Note that, for every J,

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

  because the probability that some J-heavy string is not hit by any $w^{(i)}_J$ is upper-bounded by $2^\ell \cdot (1 - 0.01/s^2)^{s'} = 2^\ell \cdot \exp(-\Omega(s'/100s^2)) = o(1)$.

  (Here we used the fact that $s' = \Omega(s^2 \cdot \ell)$.)

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

  By an averaging argument,

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

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

We fix an arbitrary good (for X) sequence $(w^{(1)},....,w^{(s')})$ 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 $(w^{(1)},...,w^{(s')})$ 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 (typical) set J, let $I = I(J)$ be a maximal set of indices $i \in [s']$ such that the $w^{(i)}_J$'s are distinct; that is, $R = R(J) \stackrel{\text{def}}{=} \{w^{(i)}_J : i \in I\}$ has size $|I|$ and equals $\{w^{(i)}_J : i \in [s']\}$. We stress that R contains all J-heavy strings (w.r.t X), which means that for every $\sigma \notin R$ it holds that $\Pr[X_J = \sigma] < 0.01/s^2$, and that each J-heavy string corresponds to a unique $i \in I$. We now define X' by selecting $x \sim X$, and outputting $w^{(i)}$ if $x_J = w^{(i)}_J$ for some $i \in I$, and outputting x itself otherwise (i.e., if $x_J \notin R(J)$); that is,

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

Note that $X'_J \equiv X_J$, which establishes Condition (i) of our plan. Turning to Condition (ii), we now prove that, for a typical J, it holds that X' is $\epsilon/2$-close to X.

Claim 2.2.1 (typically, X' is $\epsilon/2$-close to X): *With probability $1-o(1)$ over the choice of J, the corresponding $X' = X'(J)$ is $\epsilon/2$-close to X.*

Proof: The key observation is that X' differs from X only when $X_J \in \{w^{(i)}_J : i \in I(J)\} = \{w^{(i)}_J : i \in [s']\}$. In this case, strings that are $\epsilon/4$-H-close to $\{w^{(i)} : i \in I(J)\}$ contribute at most $\epsilon/4$ units (to the distance between X and X' (as in Definition 1.1)), and so we upper-bound the probability mass of strings $x \sim X$ that are $\epsilon/4$-H-far from $\{w^{(i)} : i \in I(J)\}$ but satisfy $x_J \in \{w^{(i)}_J : i \in [s']\} = R(J)$. Actually, letting $W \stackrel{\text{def}}{=} \{w^{(i)} : i \in [s']\}$ and $F_{\epsilon'}(S)$ denote the set of n-bit strings that are ϵ'-H-far from S, for every $x \in F_{\epsilon/4}(W)$, we let $\mathtt{Bad}_x(J)$ denote the event $x_J \in R(J)$. Letting $\mathtt{Bad}_x(J) = 0$ for any x that is $\epsilon/4$-H-close to W, we have

$$\begin{aligned}
\Pr_{J,X}[\mathtt{Bad}_X(J)] &= \mathrm{E}_{x\sim X}\left[\Pr_{J\in\binom{[n]}{\ell}}[\mathtt{Bad}_x(J)]\right] \\
&\leq \max_{x\in F_{\epsilon/4}(W)}\left\{\Pr_{J\in\binom{[n]}{\ell}}[\mathtt{Bad}_x(J)]\right\} \\
&\leq \max_{x\in F_{\epsilon/4}(W)}\left\{\sum_{i\in[s']}\Pr_{J\in\binom{[n]}{\ell}}[x_J = w^{(i)}_J]\right\} \\
&\leq \sum_{i\in[s']}\max_{x\in F_{\epsilon/4}(\{w^{(i)}\})}\left\{\Pr_{J\in\binom{[n]}{\ell}}[x_J = w^{(i)}_J]\right\} \\
&< s'\cdot(1-(\epsilon/4))^{\ell},
\end{aligned}$$

which is $o(\epsilon)$ by the definition of $\ell = O(\epsilon^{-1}\log(s/\epsilon))$ (and $s' = \widetilde{O}(s^2/\epsilon)$, where we actually use $s' \leq \mathrm{poly}(s/\epsilon)$). 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 $W = \{w^{(i)} : i \in [s']\}$ but satisfies $x_J \in R(J) = \{w^{(i)}_J : i \in [s']\}$ is at most $\epsilon/4$; that is,

$$\mathrm{Pr}_{J\in\binom{[n]}{\ell}}\Big[\mathrm{Pr}_{x\sim X}\left[x \in F_{\epsilon/4}(W) \text{ but } x_J \in R(J)\right] \leq \epsilon/4\Big] \;=\; 1-o(1). \tag{5}$$

It follows that, with probability $1-o(1)$ over the choice of J, it holds that X' is $\epsilon/2$-close to X, where one term of $\epsilon/4$ is due to the x's that are $\epsilon/4$-H-close to $\{w^{(i)} : i \in [s']\}$ and the other term is due to the probability mass of x's that are $\epsilon/4$-H-far to $\{w^{(i)} : i \in [s']\}$ but satisfies $x_J \in R(J)$. ■

Recalling that X is ϵ-far from $\mathcal{D}$, for a typical J, Claim 2.2.1 implies that X' is $\epsilon/2$-far from $\mathcal{D}$, which implies that X' is $\epsilon/2$-TV-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').

Claim 2.2.2 (typically, our tester rejects X): *Suppose that $(w^{(1)}, ..., w^{(s')})$ is J-good for X and that the corresponding $X' = X'(J)$ is $\epsilon/2$-far from $\mathcal{D}$. Then, our tester rejects X with probability at least* 0.66.

Proof: Recalling that $X'_J = X_J$, while relying on the hypothesis that $\overline{w} \stackrel{\text{def}}{=} (w^{(1)}, ..., w^{(s')})$ is J-good (for X), we observe that the probability that our tester rejects X equals

$$\begin{aligned}
&\mathrm{Pr}_{x^{(1)},...,x^{(s)}\sim X}[T'(x^{(1)}_J, ..., x^{(s)}_J) = 0] \\
&= \mathrm{Pr}_{x^{(1)},...,x^{(s)}\sim X'}[T'(x^{(1)}_J, ..., x^{(s)}_J) = 0] && \text{[using } X'_J = X_J\text{]} \\
&= \mathrm{Pr}_{x^{(1)},...,x^{(s)}\sim X'}[T(n, \epsilon/2; x^{(1)}_J 0^{n-\ell}, ..., x^{(s)}_J 0^{n-\ell}) = 0] && \text{[definition of } T'\text{]} \\
&= \mathrm{Pr}_{x^{(1)},...,x^{(s)}\sim X'}[T(n, \epsilon/2; x^{(1)}, ..., x^{(s)}) = 0] \pm \tfrac{\binom{s}{2}}{100\cdot s^2} && \text{[see next]}
\end{aligned}$$

where the last (approximate) equality is justified as follows (based on the definition of X' and the hypothesis that $\overline{w}$ is J-good (for 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 (i.e., for $x^{i)}_J, x^{(j)}_J \in R$, it holds that $x^{i)} = x^{(j)}$ iff $x^{i)}_J = x^{(j)}_J$). This holds because (by the definition of X'), for each $\sigma \in R$, there is a unique x in the support of X' such that $x_J = \sigma$ (i.e., $x = w^{(i)}$ such that $w^{(i)}_J = \sigma$).
- 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 $\binom{s}{2} \cdot \frac{1}{100\cdot s^2} < 0.005$. This is because (by the hypothesis that $\overline{w}$ is J-good) these J-restrictions are all non-heavy (and $\sum_{\sigma\in L} p_\sigma^2 \leq \max_{\sigma\in L}\{p_\sigma\}$, for non-negative p_σ's that sum-up to at most 1). Needless to say, the collision probability between these (other) samples themselves can only be smaller.

Indeed, the foregoing analysis establishes Condition (iii) of our plan (i.e., the collision pattern of s samples of X'_J is 0.005-close to the collision pattern of s samples of X').
It follows that our tester rejects X with probability at least $\frac{2}{3} - 0.005 > 0.66$, where the first term lower-bounds the probability that T rejects when presented with s samples of X'. ■

Conclusion. Using the hypothesis that $(w^{(1)},...,w^{(s')})$ is good (for X), with probability $1-o(1)$ over the choice of $J \in \binom{[n]}{\ell}$, it holds that $(w^{(1)},...,w^{(s')})$ is J-good (for X) and (by Claim 2.2.1) the corresponding $X' = X'(J)$ is $\epsilon/2$-close to X. Assuming that X is ϵ-far from $\mathcal{D}$, using both conclusions, it follows (by Claim 2.2.2) that our tester rejects X with probability at least $0.66-o(1)$. 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* $\mathrm{poly}(1/\epsilon) \cdot \widetilde{O}(m)$ *queries:*

1. *The set of 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 probability.

2. *The set of 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.

3. *The set of 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 a (DoHO model) tester for m-grained distributions will do.

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).[9] 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 $n = O(\epsilon^{-1}\log m)$.

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. [23]), 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 $\widetilde{O}(m/\epsilon^2)$, respectively.

Testing support size. On input parameters n and ϵ, given $s = O(m/\epsilon)$ samples, denoted $x^{(1)},....,x^{(s)}$, 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 ϵ-TV-far from having support size at most m, and note that for any set S of

[9] Saying that X is m-grained means that it is uniform on a multiset $\{x^{(1)},\ldots,x^{(m)}\}$ 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 $y^{(i)}_j$ to equal the j^{th} bit in the binary expansion of $i-1$ if $j \in [\ell]$, and $y^{(i)}_j = x^{(i)}_j$ otherwise (i.e., if $j \in \{\ell+1,...,n\}$).

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 ϵ, we set $s = O(m \log m)$ and $s' = O(\epsilon^{-2} m \log m)$. Given $s + s'$ samples, denoted $x^{(1)}, ..., x^{(s+s')}$, 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 of $|W| > m$, 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 $1/m$.

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 ϵ-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. Furthermore, for every $\ell' \in [\log_2 \log_2 n]$, every distribution over $\{0,1\}^n$ is $O(\frac{\log n}{n})$-close to being $2^{n-\ell'}$-grained.*

In contrast, in the standard (distribution testing) model, testing whether a distribution (over $\{0,1\}^n$) is $2^{n-O(1)}$-grained requires $2^{0.99 \cdot n}$ samples [17].

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 main claim follows.

In the first step, given an arbitrary distribution X, we consider the distribution X' obtained by setting the last ℓ 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 (ℓ/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 $2^{-n} \cdot \lfloor 2^n \cdot \Pr[X' = x'0^\ell] \rfloor$, 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.

The furthermore claim follows by redefining X'' such that $\Pr[X'' = x'0^\ell]$ equal $2^{-(n-\ell')} \cdot \lfloor 2^{n-\ell'} \cdot \Pr[X' = x'0^\ell] \rfloor$. In this case X'' is $2^{n-\ell'}$-grained and is at total variation distance at most $2^{n-\ell} \cdot 2^{-(n-\ell')} = 2^{-(\ell-\ell')}$ from X', which means that it is $2\frac{\ell}{n}$-close to X, since $\ell' \leq \log_2 \ell$. ■

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 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 the standard methodology by which a lower bound of L on the complexity of testing is proved by presenting two distributions X and Y that an algorithm of complexity $L-1$ cannot distinguish (with constant positive gap)[10] such that X has the property and Y is $\Omega(1)$-far from having the property (cf. [15, Thm. 7.2]). In fact, typically, at least one of the two distributions will be claimed to exist using a probabilistic argument; that is, we shall actually prove that there exists two distribution x_0 and Y_0 (over $\{0,1\}^n$) such that, for a random bijection $\pi : \{0,1\}^n \to \{0,1\}^n$, setting $X = \pi(X_0)$ and $Y = \pi(Y_0)$ will do.

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 ϵ whether a distribution is uniform over $\{0,1\}^n$ requires $2^{c \cdot n}$ samples in the* DoHO *model.*

Proof: Let S be a random $2^{2c \cdot n}$-subset of $\{0,1\}^n$, and X be uniform over S. Then, a sample of $s = o(2^{cn})$ strings does not allow for distinguishing between X and the uniform distribution over $\{0,1\}^n$; that is, for every decision procedure $D : (\{0,1\}^n)^s \to \{0,1\}$, there exists a set S such that

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

(Intuitively, this is the case because s random samples from a random set S are distributed almost identically to s random samples from the uniform distribution over n-bit strings.)[11] On the other

[10] We say that A distinguishes s samples of X from s samples of Y with gap γ if

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

[11] Formally, for every sequence $ovi = (i_1, ..., i_s) \in [N]^s$, where $N = 2^{2cn}$, let $\zeta_{\bar{i}}(S)$ denote the output of D when fed with $s_{i_1}, ..., s_{i_s}$, where s_j denotes the j^{th} element of the N-set S. Then,

$$\mu \stackrel{\text{def}}{=} \mathrm{E}_S[\zeta_{\bar{i}}(S)] = \Pr_{x^{(1)},...,x^{(s)} \in \{0,1\}^n}[D(x^{(1)}, \ldots, x^{(s)}) = 1],$$

whereas almost all paits of $\zeta_{\bar{i}}(S)$'s are pairwise independent, because $N = \omega(s^2)$. Hence,

$$\Pr_S \left[\left| \sum_{\bar{i} \in [N]^s} \zeta_{\bar{i}}(S) - N^s \cdot \mu \right| > \beta \cdot N^s \right] = \frac{O(s^2)}{\beta^2 \cdot N}$$

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 ϵ 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 ϵ 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 $\overline{S} \stackrel{\text{def}}{=} \{0,1\}^n \setminus S$.

Now, on the one hand, for a random S, no algorithm can distinguish X_S from Y_S by using $o(2^{cn})$ samples (cf. Footnote 11). On the other hand, we prove that Y_S is far from being uniform on any set. Suppose that $Y = Y_S$ is δ-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 $\overline{S} \setminus S'$ should be moved to S', whereas the average relative Hamming distance between a random element of $\overline{S} \setminus S'$ 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 $U_n \in \overline{S} \setminus S'$ is H-close to S' by noting that $|\overline{S} \setminus S'| > 2^{n-1}$, 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 $|S| = 2^{2cn} = 2^{(0.5+c) \cdot n - \Omega(n)} < 2^{-\Omega(n)} \cdot |S'|$.

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. ■

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.*

where s^2/N accounts for the fraction of non-disjoint pairs of $\bar{i}$'s.

Stronger results are presented in Propositions 2.8 and 2.9.

Proof: As in the proof of Observation 2.5, observe that no algorithm can distinguish the uniform distribution over $\{0,1\}^n$ from a distribution that is uniform over an m-subset unless its 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 $m \leq 2^{n-\Omega(n)}$. ■

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 [22] 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 [22] 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 $C : [O(m)] \to \{0,1\}^n$ 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'. Actually, as in the previous proofs, we need to consider any algorithm that takes s samples, and we identify for each such algorithm two such distributions X and Y (which are relabelings of the original X and Y) that are indistinguishable by it (cf. Footnote 11). 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 γ denote the distance between Y' and Z. By Definition 1.1, this means that there exists a "weight relocation" function $W : \{0,1\}^{2n} \to [0,1]$ 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 $\sum_{y'}\sum_z W(y',z) \cdot \mathrm{InEq}(y',z)$ is lower-bounded by the total variation distance between Y' and Z, where $\mathrm{InEq}(y',z) = 1$ if $y' \neq z$ and $\mathrm{InEq}(y',y') = 0$.

Let S denote the support of Z (so that $W(y',z) = 0$ for every $z \notin Z$), 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 $\mathtt{cc}(z) \in C$ be the codeword in C that is closest to z. By the definition of S' we have that $\delta'(z) \stackrel{\text{def}}{=} \Delta_H(\mathtt{cc}(z), z) \leq 0.4 \cdot \delta$, and for every $y' \in R \setminus \{\mathtt{cc}(z)\}$, 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 $\mathtt{cc}(z)$; that is, $\sum_{y' \in R \setminus \{\mathtt{cc}(z)\}} W(y',z) \geq \frac{1}{2} \cdot \sum_{y'} W(y',z)$. We prove that $\sum_{y' \in R \setminus \{\mathtt{cc}(z)\}} W(y',z) \geq W(\mathtt{cc}(z),z)$, 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 $\mathtt{cc}(z)$; that is, $\Pr[Z'=z]=0$ and $\Pr[Z'=\mathtt{cc}(z)] = \Pr[Z=\mathtt{cc}(z)] + \Pr[Z=z]$ (and $\Pr[Z'=z'] = \Pr[Z=z']$ for every $z' \notin \{z, \mathtt{cc}(z)\}$), 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 $W'(y',\mathtt{cc}(z)) = W(y',\mathtt{cc}(z)) + W(y',z)$ (leaving $W'(y',z') = W(y',z')$ for every $z' \notin \{z,\mathtt{cc}(z)\}$)). 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 $\sum_{y'} W(y',z) \cdot \Delta_H(y',\mathtt{cc}(z))$. Now,

$$\sum_{y' \in R} W(y',z) \cdot \Delta_H(y',z) \geq W(\mathtt{cc}(z),z) \cdot \delta'(z) + \sum_{y' \in R \setminus \{\mathtt{cc}(z)\}} W(y',z) \cdot (\delta - \delta'(z)) \,, \tag{6}$$

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

$$\sum_{y' \in R} W(y',z) \cdot \Delta_H(y',\mathtt{cc}(z)) \leq \sum_{y' \in R \setminus \{\mathtt{cc}(z)\}} W(y',z) \cdot \delta \,. \tag{7}$$

Using the counter hypothesis (i.e., $W(\mathtt{cc}(z),z) > \sum_{y' \in R \setminus \{\mathtt{cc}(z)\}} W(y',z)$), we lower-bound Eq. (6) by $\sum_{y' \in R \setminus \{\mathtt{cc}(z)\}} W(y',z) \cdot \delta$, 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 $\sum_{y'} \sum_{z \in (S' \cap C)} W(y',z) \cdot \mathrm{InEq}(y',z) \cdot \delta$.

To summarize, the distance γ between Y' and Z, under Definition 1.1, is at least a 0.3δ 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 [17]. See Section 2.4 for further discussion.

Proof: We use the $\Omega(m^c)$ lower bound of [17] that refers to testing whether a distribution over $[O(m)]$ is m-grained, under the total variation distance. This lower bound is proved in [17] 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 $C : [O(m)] \to \{0,1\}^n$ 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).[12] 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.)[13] ■

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.[14]

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.[15] 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. ■

2.4 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. The following lower bounds on the complexity of testing in the DoHO model rely on a conjecture regarding the sample complexity of testing grained distributions in the standard model.

Conjecture 2.11 (on the complexity of testing the set of m-grained distributions in the standard distribution-testing model): *In the standard distribution-testing model, the sample complexity of testing m-grained distributions over a domain of size $O(m)$ is $\Omega(m/\log m)$.*

Theorem 2.12 (on testing the set of m-grained distributions in the DoHO model): *Assuming Conjecture 2.11, for every $m \leq 2^{n-\Omega(n)}$, testing that a distribution over $\{0,1\}^n$ is m-grained requires $\Omega(m/\log m)$ samples in the* DoHO *model.*

---

[12] In fact, as in the proof of Proposition 2.8, we actually consider adequare relabelings of X and Y.

[13] 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.

[14] We actually use $m \log m = o(n^{1/c})$, which follows from $m \leq n$.

[15] Saying that X' is m-grained means that it is uniform on a multiset $\{x^{(1)}, \ldots, x^{(m)}\}$ 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 $\ell_1 < \ell_2 < \cdots < \ell_{\log_2 m}$ denote the chosen locations, we set $y^{(i)}_{\ell_j}$ to equal the j^{th} bit in the binary expansion of $i-1$ and set $y^{(i)}_\ell = x^{(i)}_\ell$ if $\ell \in [n] \setminus \{\ell_1, \ell_2, \ldots, \ell_{\log_2 m}\}$.

We mention that the proof would remain unchanged if the lower bound in Conjecture 2.11 is replaced by $s(m)$; that is, any lower bound of the form $\Omega(s(m))$ on the sample complexity of testing m-grained distributions in the standard distribution testing model translates to an $\Omega(s(m))$ lower bound in the DoHO model. A similar comment refers to Theorem 2.13.

Proof: We would have liked to argue that the proof is analogous to the proof of Proposition 2.9, except that here we assume the existence of two distributions, X and Y, over $[O(m)]$ that cannot be distinguished by a label-invariant algorithm that gets $o(m/\log m)$ samples, where X is m-grained and Y is far (in total variation distance) from being m-grained. However, since Conjecture 2.11 does not quite imply the existence of such distributions X and Y, we apply a slightly more complex argument. Our starting point is the observation that Conjecture 2.11 implies the existence of multisets of distributions[16], denoted $\mathcal{X}$ and $\mathcal{Y}$, such that the following holds:

1. Each distribution in $\mathcal{X}$ is m-grained.
2. Each distribution in $\mathcal{Y}$ is TV-far from being m-grained.
3. No algorithm can distinguish between $s = o(m/\log m)$ samples taken from a distribution X that is selected uniformly in $\mathcal{X}$ and s samples taken from a distribution Y that is selected uniformly in $\mathcal{Y}$.

The foregoing observation is proved by applying the MiniMax Principle (cf. [14, Apdx A.1]). Specifically, we consider deterministic algorithms that, given s samples from a distribution Z, try to distinguish between the case that Z is m-grained and the case that Z is TV-far from being m-grained, and denote by $c(A, Z)$ the probability that algorithm A is correct on Z (i.e., it correctly identifies Z's type). Then, Conjecture 2.11 asserts that, for every distribution $\mathcal{A}$ of algorithms (i.e., a randomized algorithm) that get s samples, there exists a distribution Z (which is either m-grained or far from m-grained) such that $\mathcal{A}$ errs on Z with probability greater than $1/3$ (i.e., $\mathrm{E}_{A\sim\mathcal{A}}[c(A, Z)] < 2/3$). The minimax principle then implies that there exists a multiset $\mathcal{Z}$ of such distributions (which are each either m-grained or far from m-grained) on which each algorithm A that takes s samples errs on the average with probability greater than $1/3$ (i.e., $\mathrm{E}_{Z\in\mathcal{Z}}[c(A, Z)] < 2/3$). Analogously to [15, Exer. 7.3], we obtain $\mathcal{X}$ and $\mathcal{Y}$ as desired, where the indistinguishability gap is less than $1/2$.

Consider the corresponding multisets $\mathcal{X}'$ and $\mathcal{Y}'$, which are obtained by applying a (constant-distance) error correcting code C to the elements of each distribution in $\mathcal{X}$ and $\mathcal{Y}$, respectively. we conclude that no algorithm that takes s samples can distinguish X' from Y', where X' (resp., Y') is selected uniformly in $\mathcal{X}'$ (resp., $\mathcal{Y}'$), where the indistinguishability gap is less than $1/2$. (As in the proof of Proposition 2.9, we show that each Y' is far (under Definition 1.1) from being m-grained.) Observing that a distinguishing gap of less than $1/2$ means that no algorithm (of low complexity) constitutes a tester with error probability at most $1/4$ (rather than at most $1/3$), the claim follows (using error reduction). ■

Theorem 2.13 (on testing parameterized uniformity in the DoHO model): *Assuming Conjecture 2.11, for every $m \leq n$, testing that a distribution over $\{0,1\}^n$ is uniform over some m-subset requires $\Omega(m/\log m)$ queries in the* DoHO *model.*

[16] Actually, $\mathcal{X}$ and $\mathcal{Y}$ are distributions of distributions. However, to avoid confusion, we preferred to present them as multi-set and consider a uniformly selected element in them.

We stress that, unlike Theorem 2.12, which lower-bounds the sample complexity of testers, in Theorem 2.13 we only lower-bound their query complexity.

Proof: Let $\mathcal{X}'$ and $\mathcal{Y}'$ denote the multisets of distributions derived in the proof of Theorem 2.12. Recall that each distribution in $\mathcal{X}'$ is m-grained, whereas each distribution in $\mathcal{Y}'$ is far from being m-grained (under Definition 1.1). As in the proof of Proposition 2.10, note that each distribution in $\mathcal{Y}'$ is $\Omega(1)$-far from being uniform over a set of size m, and observe that each distribution in $\mathcal{X}'$ is $\frac{\log_2 m}{n}$-close to being uniform over a set of size m. Specifically, we can make each distribution in $\mathcal{X}'$ uniform by modifying only the bits that reside in $\log_2 m$ locations, where the choice of these locations is arbitrary. 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 $o(m/\log m)$ queries cannot distinguish between distributions in the modified multiset $\mathcal{X}'$ and distributions in the multiset $\mathcal{Y}'$, which implies that it fails to test uniformity in the DoHO model. ■

3 Distributions on self-correctable/testable sets

In this section we prove Theorem 1.8, which refers to properties of distributions that are supported on a set of strings $\Pi \subseteq \{0,1\}^n$ that has an efficient self-correction/testing procedure. In this case, label-invariance actually means being label-invariant when restricted to Π; that is, for every bijection $\pi : \Pi \to \Pi$ and every distribution X, it holds that X is in the property if and only if $\pi(X)$ is in the property.

Our starting point is a label-invariant property of distributions, denoted $\mathcal{D}$, and a property of strings, denoted Π, that has a relatively efficient tester and local self-corrector. Actually, we use a *relaxed definition of self-correction*, which allows to output a special failure symbol in case the input (oracle) is not in Π (but close to Π).[17] Indeed, proper behavior of the self-corrector is required only up to a specified distance from the set Π. Combining $\mathcal{D}$ and Π, we get a property of distributions, denoted $\mathcal{D}_\Pi$, that consists of all distributions in $\mathcal{D}$ that are supported by Π. Note that $\mathcal{D}_\Pi$ is label-invariant over its support. We prove that the query complexity of testing $\mathcal{D}_\Pi$ in the DoHO model is related to the sample complexity of testing $\mathcal{D}$ (in the standard model) and to the query complexity of the two foregoing procedures.

Theorem 3.1 (from standard distribution testing of $\mathcal{D}$ to testing $\mathcal{D}_\Pi$ in the DoHO model, when Π is efficiently testable and self-correctable):

- *Let $\mathcal{D}$ be a label-invariant property of distributions over $\{0,1\}^n$, and suppose that $\mathcal{D}$ is testable in the standard model using $s(n,\epsilon) = \Omega(1/\epsilon)$ samples.*[18]
- *Let $\Pi \subseteq \{0,1\}^n$ be a property of strings that is testable with query complexity $q_{\mathtt{T}}(n,\epsilon)$ and self-correctable up to distance $\delta(n)$ with $q_{\mathtt{C}}(n)$ queries; that is, there exists an oracle machine C that makes at most $q_{\mathtt{C}}(n)$ queries such that for every $x \in \{0,1\}^n$ and $i \in [n]$ the following two conditions hold:*
  1. *If $x \in \Pi$, then $\Pr[C^x(i) = x_i] \geq 2/3$.*

[17] The notion of relaxed self-correction was introduced in [19], by analogy to the notion of relaxed LDCs [7, Sec. 4.2].

[18] Indeed, it would have been more consistent with the literature to denote the sample complexity by $s(2^n,\epsilon)$, since the domain size is 2^n.

2. *If x is $\delta(n)$-close to $x' \in \Pi$, then $\Pr[C^x(i) \in \{x'_i, \perp\}] \geq 2/3$, where $\perp$ is a special symbol indicating failure.*

- *Suppose that every distribution in $\mathcal{D}$ is supported by a subset of size at most $|\Pi|$, and let $\mathcal{D}_\Pi$ denote the set of all distributions in $\mathcal{D}$ that have a support that is a subset of Π.*[19]

Then, $\mathcal{D}_\Pi$ is testable with query complexity

$$q(n,\epsilon) = \widetilde{O}(s(n,\epsilon'/2)) \cdot \frac{q_{\mathtt{T}}(n,\delta(n)) + q_{\mathtt{C}}(n)}{\delta(n)}$$

and sample complexity $s(n,\epsilon'/2)$, where $\epsilon' = \min(\epsilon, \delta(n))$.

Note that Π must have relative (Hamming) distance greater $\delta(n)$, since otherwise we reach a contradiction by considering two strings in Π that are $\delta(n)$-H-close to one another.[20] We also note that Π is testable with query complexity $q'_{\mathtt{T}}(n,\epsilon) = O(q_{\mathtt{T}}(n,\delta(n))) + \widetilde{O}(1/\epsilon) \cdot q_{\mathtt{C}}(n)$, by first invoking the original tester with proximity parameter set to $\delta(n)$, and then selecting $O(1/\epsilon)$ random locations $i \in [n]$ and comparing their self-corrected value (i.e., $C^x(i)$) to their given value (i.e., x_i).[21] (The extra $O(\log(1/\epsilon))$ factor accounts for reducing the error of the self-corrector to $o(1/\epsilon)$.)

In some natural applications (e.g., low-degree multi-variate polynomials), we may use $\delta(n) \geq 1/\mathrm{poly}(\log n)$ and $q_{\mathtt{T}}(n,\delta(n)) + q_{\mathtt{C}}(n) \leq \mathrm{poly}(\log n)$, but for relaxed locally correctable (and testable) code [7, 19, 10] we may have $\delta(n) = \Omega(1)$ and $q_{\mathtt{T}}(n,\delta(n)) + q_{\mathtt{C}}(n) = O(1)$.

Proof: We first recall that, we may assume, without loss of generality, that when testing a label-invariant property, in the standard model, the tester is label-invariant [2] (see also [15, Thm. 11.12]). Such a tester, denoted T', actually rules according to the collision pattern that it sees in the sequence of samples; that is, the number of i-way collisions in the sequence, for each $i \in \mathbb{N}$. Specifically, the collision pattern of the sequence $\overline{x} = (x^{(1)}, \ldots, x^{(s)})$, denoted $\mathtt{CP}(\overline{x})$, is the sequence $(c_1, \ldots, c_s)$ such that $c_i = |\{v \in \Omega : \#_v(x^{(1)}, \ldots, x^{(s)}) = i\}|$ is the number of i-way collisions, where $\#_v(x^{(1)}, \ldots, x^{(s)}) = |\{j \in [s] : x^{(j)} = v\}|$ is the number of elements (in the sequence) that equal v. Hence, we claim that

$$T'(n,\epsilon; x^{(1)}, \ldots, x^{(s)}) = T''(n,\epsilon; \mathtt{CP}(x^{(1)}, \ldots, x^{(s)})),$$

for some randomized decision procedure T''.

Warm-up (or a first attempt). Using the fact that Π has relative distance $\delta = \delta(n)$, let us consider a tester that, given the sample-sequence $\overline{x} = (x^{(1)}, \ldots, x^{(s)})$, where $s = s(n,\epsilon)$, picks uniformly at random an $O(\delta^{-1} \log s)$-subset of $[n]$, denoted I, and outputs $T''(n,\epsilon; \mathtt{CP}(x^{(1)}_I, \ldots, x^{(s)}_I))$, where $x^{(i)}_I$ is the restriction of $x^{(i)}$ to the coordinates in I. This tester works well when the tested distribution X is supported on Π. In this case, for any sequence of s samples $\overline{x} = (x^{(1)}, \ldots, x^{(s)})$

[19] The condition (regarding the size of the support of distributions in $\mathcal{D}$) implies that *if X' is supported by a subset of Π and is TV-close to $\mathcal{D}$, then it is TV-close to $\mathcal{D}_\Pi$*. Note that this implication does not hold without the condition. (Consider, for example, $\Pi = \{0^n, 1^n\}$ and $\mathcal{D}$ that contains all distributions that are uniform on some 2-subset as well as all distribution that have support size 2^n. Then, $X' \equiv 0^n$ is very TV-close to $\mathcal{D}$, although it is TV-far from $\mathcal{D}_\Pi$ (which consists of a single distribution).

[20] Let x and x' be two distinct strings that are at relative distance at most $\delta(n)$, and suppose than $x_i \neq x'_i$. Then, by the first condition $\Pr[C^x(i) = x_i] \geq 2/3$, but by the second condition $\Pr[C^x(i) \in \{x'_i, \perp\}] \geq 2/3$.

[21] This idea is implicit in the proof of [15, Thm. 5.12].

of X, with high probability, $\mathtt{CP}(x_I^{(1)},\ldots,x_I^{(s)})$ equals $\mathtt{CP}(x^{(1)},\ldots,x^{(s)})$, since $x^{(i)} \neq x^{(j)}$ implies $\Pr_I[x_I^{(i)} = x_I^{(j)}] < (1-\delta)^{O(\delta^{-1}\log s)} = o(1/s^2)$. Hence, in this case we correctly distinguish X in $\mathcal{D}_\Pi$ from X that is ϵ-far from $\mathcal{D}_\Pi$ although it is supported by Π. Of course, we can easily test that X is supported by Π (using the tester for Π – see Theorem 1.4), but the problem is that our samples may be close to Π and yet not reside in it. This is a problem because the foregoing analysis presupposed that the inequality between samples is reflected in their restrictions to a small subset I (i.e., that $x^{(i)} \neq x^{(j)}$ typically implies $x_I^{(i)} = x_I^{(j)}$]).

We address this problem by using the hypothesis regarding Π; that is, not only is Π testable (with proximity parameter ϵ using $q_{\mathtt{T}}(n,\epsilon)$ queries), but it is also self-correctable to distance δ by using $q_{\mathtt{C}}(n)$ queries. In particular, combining the tester for Π (applied with proximity parameter $\delta(n)$), and the self-corrector (and employing error reduction[22] we can obtain an oracle machine C that satisfies the following for every $x \in \{0,1\}^n$ and $i \in [n]$ (for any integer parameter s):

1. If $x \in \Pi$, then $\Pr[C^x(i) = x_i] \geq 1 - o(1/s^2)$.
2. If x is δ-H-close to $x' \in \Pi$, then $\Pr[C^x(i) \in \{x'_i, \perp\}] \geq 1 - o(1/s^2)$.
3. If x is δ-H-far from Π, then $\Pr[C^x(i) = \perp] \geq 1 - o(1/s^2)$.

This combined machine has query complexity $q'_{\mathtt{C}}(n,s) = O(\log s) \cdot (q_{\mathtt{T}}(n,\delta) + q_{\mathtt{C}}(n))$. We are now ready to present an analyze our tester.

The proposed tester, T. On input n, ϵ and $s = s(n, \epsilon/2)$ samples drawn from a tested distribution X, and denoted $x^{(1)},\ldots,x^{(s)}$, we proceeds as follows (assuming, for simplicity (and the most part), that $\epsilon \leq \delta(n)/2$).

1. We test whether X is supported by Π with proximity parameter $\epsilon/2$, where the distance here (and throughout the proof, unless stated explicitly otherwise) is according to Definition 1.1. If this test rejects, then we reject.

   (Here and below, "supported by Π" means having a support that is a subset of Π.)

   We use the tester provided by Theorem A.2 (with proximity parameter $\epsilon/2$), while noting that we can reuse some of the samples provided to T for this purpose (since $s = \Omega(1/\epsilon)$). As noted following the statement of Theorem 3.1, the query complexity of testing Π is $q'_{\mathtt{T}}(n,\epsilon) = O(q_{\mathtt{T}}(n,\delta(n))) + \widetilde{O}(1/\epsilon) \cdot q_{\mathtt{C}}(n))$. We infer that the query complexity of the current step is $\max(\widetilde{O}(q'_{\mathtt{T}}(n,\epsilon/4)), O(q_{\mathtt{T}}(n,\delta(n))/\delta(n)))$.[23]

2. For each $i \in [s]$, we test whether $x^{(i)}$ is in Π, when setting the proximity parameter to $\delta = \delta(n)$ and the error bound to $o(1/s)$. If any of these checks rejects, then we reject. Otherwise, we may assume that each sample is $\delta(n)$-H-close to Π.

   Note that typically $s \gg O(1/\epsilon)$ and $\delta \gg \epsilon$, which implies that the current step is incomparable to Step 1.[24] The query complexity of the current step is $\widetilde{O}(s) \cdot q_{\mathtt{T}}(n,\delta)$.

[22] In the context of self-correction, performing error reduction means that if a strict majority of the invocations return a Boolean value, then we can use that value (which happens w.h.p. when $x \in \Pi$). Otherwise, we output $\perp$, since the lack of a strict majority indicates error.

[23] The first term account for the case of $\epsilon \leq \delta(n)/2$, whereas the second term accounts for the opposite case.

[24] On the one hand, we test that all samples are in Π (rather than only testing the first $O(1/\epsilon)$ samples. On the other hand, the proximity parameter we use is larger; that is, the test is more crude.

3. For $\ell = O(\delta^{-1} \log s)$, select a random ℓ-subset of $[n]$, denoted I, and obtain $(y^{(1)}, \ldots, y^{(s)})$ such that $y^{(i)}$ is the self-correction of $x^{(i)}$ for locations I; that is, letting $I = \{p_1, \ldots, p_\ell\}$ such that $p_j < p_{j+1}$ (for all $j \in [s-1]$), we let $y^{(i)}_j \leftarrow C^{x^{(i)}}(p_j)$. (Recall that C has error probability $o(1/s^2)$.)

   If any of these correction attempts fails (i.e., if any $y^{(i)}_j$ equals $\perp$), then we reject. Otherwise, we output the verdict of $T''(n, \epsilon/2; \mathtt{CP}(y^{(1)}, \ldots, y^{(s)}))$. We may assume, without loss of generality, that T' (and so T'') errs with probability at most 0.1.

Typically (i.e., when $s \gg 1/\epsilon$), the query complexity of T is dominated by the last step and equals

$$s \cdot \ell \cdot q'_{\mathtt{C}}(n, s) = O(s \log^2 s) \cdot \frac{q_{\mathtt{T}}(n, \delta(n)) + q_{\mathtt{C}}(n)}{\delta(n)}.$$

(Recall that the complexity of Step 1 is $\max(\widetilde{O}(q'_{\mathtt{T}}(n, \epsilon/4)), O(q_{\mathtt{T}}(n, \delta(n))/\delta(n)))$, where $q'_{\mathtt{T}}(n, \epsilon/4) = O(q_{\mathtt{T}}(n, \delta)) + \widetilde{O}(1/\epsilon) \cdot q_{\mathtt{C}}(n)$, whereas the complexity of Step 2 is $\widetilde{O}(s) \cdot q_{\mathtt{T}}(n, \delta)$.)

Analysis of the proposed tester T. We start with the case that X is in $\mathcal{D}_\Pi$. In this case, the first two steps cause rejection with probability $o(1)$, since all $x^{(i)}$'s are in Π. Furthermore, in this case, with probability $1 - o(1)$, each $y^{(i)}$ equals the restriction of $x^{(i)}$ to the locations in I (i.e., $y^{(i)} = x^{(i)}_I$). As argued in the motivational discussion, if $x^{(i)} \neq x^{(j)}$, then $x^{(i)}$ and $x^{(j)}$ are δ-H-far apart from one another, and so $\Pr_{p \in [n]}[x^{(i)}_p = x^{(j)}_p] < 1 - \delta$. This implies that $\Pr_I[x^{(i)}_I = x^{(j)}_I] < (1-\delta)^\ell = o(1/s^2)$, by our choice of ℓ. We conclude that $\Pr_I[\mathtt{CP}(y^{(1)}, \ldots, y^{(s)})) = \mathtt{CP}(x^{(1)}, \ldots, x^{(s)}))] = 1 - o(1)$, which implies that our tester accepts X with probability at least $0.9 - o(1)$.

We now turn to the case that X is ϵ-far from $\mathcal{D}_\Pi$ (according to Definition 1.1). The easy case is that X is $\epsilon/2$-far from being supported by Π, and this case leads Step 1 to reject with very high probability. We thus assume that X is $\epsilon/2$-close to being supported by Π, and let $\mathtt{corr}(x)$ denotes the string in Π that is closest to x. Then, in expectation, $\mathtt{corr}(X)$ is $\epsilon/2$-H-close to X, since the Hamming distance between x and $\mathtt{corr}(x)$ equals the Hamming distance between x and Π. Hence, $\mathtt{corr}(X)$ is $\epsilon/2$-close to X, which implies that $\mathtt{corr}(X)$ is $\epsilon/2$-far from $\mathcal{D}_\Pi$. By the next claim, this implies that the total variation distance between $\mathtt{corr}(X)$ and $\mathcal{D}$ is greater than $\epsilon/2$.

Claim 3.1.1 (distance to $\mathcal{D}_\Pi$ vs TV-distance to $\mathcal{D}$): *Let X' be a distribution supported by Π such that X' is ϵ'-far from $\mathcal{D}_\Pi$* (according to Definition 1.1)*. Then, the total variation distance between X' and $\mathcal{D}$ is greater than ϵ'.*

Proof: Assume, contrary to the claim, that X' is ϵ'-TV-close to some distribution Y' in $\mathcal{D}$. If Y' in $\mathcal{D}_\Pi$, then we immediately reach a contradiction to the hypothesis of the claim by which X' is at distance greater than ϵ' from $\mathcal{D}_\Pi$ (according to Definition 1.1). This is the case because (as noted in the introduction), the total variation distance between distributions upper-bounds the distance according to Definition 1.1. Hence, Y' in $\mathcal{D} \setminus \mathcal{D}_\Pi$. We claim that in such a case, based on Y', we can define a distribution Y'' in $\mathcal{D}_\Pi$ such that X' is ϵ'-TV-close to Y'', resulting once again in a contradiction. Thus, it remains to establish the existence of such a distribution Y''.

Recall that by the premise of Theorem 3.1, the support size of Y' is at most $|\Pi|$. Let S' denote the support of Y', and $S'' = S' \setminus \Pi$. Consider any subset Π'' of $\Pi \setminus S'$ such that $|\Pi''| = |S''|$ (such a subset must exist because $|S'| \leq |\Pi|$ and hence $|S''| = |S' \setminus \Pi| \leq |\Pi \setminus S'|$). Selecting any bijection ϕ between S'' and Π'', we set $\Pr[Y'' = j] = \Pr[Y' = \phi^{-1}(j)]$ for every $j \in \Pi''$, and

$\Pr[Y''=j]=\Pr[Y'=j]$ for every $j\in\Pi\cap S'$. Note that the total variation distance between X' and Y'' is upper-bounded by the total variation between X' and Y', because the probability mass assigned by Y'' to Π'' is already charged to the TV-distance between Y' and X' (since $\Pr[Y'\in S'']=\Pr[Y''\in\Pi'']$ and $\Pr[X'\in S'']=0$). ■

By applying Claim 3.1.1, we get that the total variation distance between $\mathtt{corr}(X)$ and $\mathcal{D}$ is greater than $\epsilon/2$. It follows that, with probability at least 0.9, the (standard) tester for $\mathcal{D}$ (i.e., T'), rejects when given $s=s(n,\epsilon/2)$ samples of $\mathtt{corr}(X)$. Hence, *with probability at least 0.9, a sequence of s samples of $\mathtt{corr}(X)$ yields a collision pattern that leads T'' to reject.* Recall, however, that we invoke T'' on s samples of X, not of $\mathtt{corr}(X)$. Nevertheless, we show that our tester (i.e., T) will reject with high probability also in this case.

Claim 3.1.2 (the distance of $\mathtt{corr}(X)$ from $\mathcal{D}$): *If $\mathtt{corr}(X)$ is $\epsilon/2$-TV-far from $\mathcal{D}$, then T rejects with probability $0.9-o(1)$.*

Proof: We consider s samples $x^{(1)},\ldots,x^{(s)}$ taken from X. On the one hand, if any of these $x^{(i)}$'s is δ-H-far from Π, then Step 2 rejects with very high probability. On the other hand, if $x^{(i)}$ is δ-H-close to Π, then $\Pr[C^{x^{(i)}}(j)\in\{\mathtt{corr}(x^{(i)})_j,\perp\}]=1-o(1/s^2)$ for every $j\in[n]$, which means that T either obtains s samples of $\mathtt{corr}(X)_I$ or rejects. Recall that, with very high probability, a sequence of s samples of $\mathtt{corr}(X)_I$ has the same collision pattern as a sequence of s samples of $\mathtt{corr}(X)$, since $\mathtt{corr}(X)$ is supported by strings that are pairwise δ-H-far apart. Lastly, recall that the collision pattern of a sequence of s samples of $\mathtt{corr}(X)$ causes T'' to reject (whp). To summarize, letting $C^x(p_1,\ldots,p_\ell)=(C^x(p_1),\ldots,C^x(p_\ell))$, we have

$$\begin{aligned}
&\Pr_{x^{(1)},\ldots,x^{(s)}\sim X}[T^{x^{(1)},\ldots,x^{(s)}}(n,\epsilon)=0]\\
&\geq \Pr_{x^{(1)},\ldots,x^{(s)}\sim X\,;\,I\in\binom{[n]}{\ell}}[T''(n,\epsilon;\mathtt{CP}(C^{x^{(1)}}(I),\ldots,C^{x^{(s)}}(I))=0]\\
&\geq \Pr_{x^{(1)},\ldots,x^{(s)}\sim X\,;\,I\in\binom{[n]}{\ell}}[T''(n,\epsilon;\mathtt{CP}(\mathtt{corr}(x^{(1)})_I,\ldots,\mathtt{corr}(x^{(s)})_I)=0]-o(1)\\
&\geq \Pr_{x^{(1)},\ldots,x^{(s)}\sim X}[T''(n,\epsilon;\mathtt{CP}(\mathtt{corr}(x^{(1)}),\ldots,\mathtt{corr}(x^{(s)}))=0]-o(1)
\end{aligned}$$

which is $0.9-o(1)$. We stress that the foregoing inequalities hold since we have ignored cases that cause rejection (e.g., $x^{(i)}$ being δ-H-far from Π and other cases in which C outputs $\perp$). ■

Combining Claims 3.1.1 and 3.1.2, we infer that if $\mathtt{corr}(X)$ is $\epsilon/2$-far from $\mathcal{D}_\Pi$, then T rejects X with high probability. Recalling that if X is ϵ-far from $\mathcal{D}_\Pi$, then either X is $\epsilon/2$-far from being supported on Π (which causes Step 1 to reject (whp)) or $\mathtt{corr}(X)$ is $\epsilon/2$-far from $\mathcal{D}_\Pi$, it follows that T rejects (whp) in any case. ■

4 Distributions as Materialization of an Ideal Object

As stated in the introduction, we consider three types of random variations of an ideal object: random noise applied to bits of a string (a.k.a perturbations), random cyclic-shifts of a string, and random isomorphic copies of a graph represented by a string. These types are studied in the following three subsections.

4.1 Perturbation

For two constant parameters $\eta \in [0, 0.5)$ and $\delta \in [0, 1]$, and every string $x^* \in \{0,1\}^n$, we consider all distributions in which each bit of x^* is flipped with probability at most η and the outcome is at Hamming distance at most $\delta \cdot n$ from x^*. That is, $\mathcal{D}^{\mathrm{per}}_{\eta,\delta}(x^*)$ contains the distribution X if

1. For every $i \in [n]$, it holds that $\Pr[X_i \neq x^*_i] \leq \eta$.
2. $\Pr[|\{i \in [n] : X_i \neq x^*_i\}| \leq \delta \cdot n] = 1$.

Indeed, setting $\delta = 1$ trivializes the second condition, whereas setting $\delta = 0$ mandates $X \equiv x^*$. Letting $\mathcal{D}^{\mathrm{per}}_{\eta,\delta} = \bigcup_{x^* \in \{0,1\}^n} \mathcal{D}^{\mathrm{per}}_{\eta,\delta}(x^*)$, we prove the following.

Theorem 4.1 (testing noisy versions of a string): *For two constant parameters $\eta \in [0, 0.5)$ and $\delta \in [0, 1]$, the property $\mathcal{D}^{\mathrm{per}}_{\eta,\delta}$ can be tested with $\mathrm{poly}(1/\epsilon)$ queries.*

Proof: The key observation is that if X is in $\mathcal{D}^{\mathrm{per}}_{\eta,\delta}(x^*)$, for some string $x^* \in \{0,1\}^n$, then each bit of x^* can be recovered with probability $1 - 2^{-t}$ by querying $O(t)$ samples of X (at the corresponding location). This allows to estimate the flipping probability of individual bits in X as well as the distribution of the Hamming distance between X and x^*. In view of this observation, the tester proceeds as follows (assuming $\eta + 0.1\epsilon < 0.5$, or else ϵ is set so that it satisfies this constraints - recall that η is a constant).

1. The tester selects uniformly, independently at random $\widetilde{O}(1/\epsilon^2)$ indices in $[n]$ and lets the resulting set be denoted by I.
2. For each $i \in I$, the tester estimates the probability $\Pr[X_i = 1]$ by taking $\widetilde{O}(1/\epsilon^2)$ samples of X and querying each sample at location i. If the estimated value is in $[\eta + 0.2\epsilon, 1 - \eta - 0.2\epsilon]$, then the tester *rejects*. Otherwise it determines $\hat{x}_i$ accordingly; that is, $\hat{x}_i = 1$ if the estimate is at most $\eta + 0.2\epsilon$, and $\hat{x}_i = 0$ otherwise.

   (Note that the same samples are used for all $i \in I$.)
3. The tester takes $m = \widetilde{O}(1/\epsilon)$ samples of X, denoted $x^{(1)}, \ldots, x^{(m)}$. If, for any $j \in [m]$, it holds that $|\{i \in I : x^{(j)}_i \neq \hat{x}_i\}| > (\delta + 0.1\epsilon) \cdot |I|$, then the tester *rejects*. Otherwise, the tester accepts.

Suppose X belongs to $\mathcal{D}^{\mathrm{per}}_{\eta,\delta}(x^*)$ for some $x^* \in \{0,1\}^n$. First observe that for any choice of the subset I (in the first step of the algorithm), the following holds by applying the additive Chernoff bound and a union bound: With high constant probability, taken over the choice of the sampled strings selected in the second step, the tester does not reject in this step, and furthermore, $\hat{x}_i = x^*_i$ for every $i \in I$. Next observe that for any choice of $x^{(1)}, \ldots, x^{(m)}$ (as selected in the third step of the algorithm), the following also holds by applying the additive Chernoff bound and a union bound: The probability, taken over the choice of I, that for some j we have that $|\{i \in I : x^{(j)}_i \neq x^*_i\}| > (\delta + 0.1\epsilon) \cdot |I|$, is a small constant. (Note that here we are referring to x^* and not $\hat{x}$). By combining the two observations we get that the tester accepts with high constant probability (taken both over the choice of I and over the choice of the sample selected in the second step).

Now suppose that X is ϵ-far from $\mathcal{D}^{\mathrm{per}}_{\eta,\delta}$. For each $i \in [n]$, let x'_i denote the more likely value of X_i. Then one of the following two conditions must hold (or else we get that X is ϵ-close to $\mathcal{D}^{\mathrm{per}}_{\eta,\delta}(x')$).

1. $\sum_{i\in[n]} \min(\Pr[X_i \neq x'_i] - \eta, 0) > \epsilon n/2$. This implies that $\Pr[X_i \neq x'_i] > \eta + \epsilon/4$ for at least $\epsilon/4$ fraction of the indices $i \in [n]$.

2. The probability that X is $(\delta + 0.2\epsilon)$-H-far from $x' = x'_1 \cdots x'_n$ is at least 0.3ϵ.

Suppose that the first condition holds. Then with high constant probability over the choice of I, for at least one of the indices $i \in I$, we have that $\Pr[X_i \neq x'_i] > \eta + \epsilon/4$. Assuming this event holds, with high constant probability over the choice of the sample selected in the second step of the algorithm, the algorithm rejects in this step. Furthermore, for any choice of I, if it does not contain any i for which $\Pr[X_i \neq x'_i] > \eta + \epsilon/4$, then with high constant probability, $\hat{x}_i = x'_i$ for every $i \in I$. Now suppose that the second condition holds. Then with high constant probability, for at least one of the sample strings $x^{(j)}$ selected in the third step of the algorithm, $x^{(j)}$ is $(\delta + 0.2\epsilon)$-H-far from x'. Conditioned on this event, with high constant probability over the choice of I, we have that $|\{i \in I : x^{(j)}_i \neq x'_i\}| > (\delta + 0.1\epsilon) \cdot |I|$. We hence conclude that if X is ϵ-far from $\mathcal{D}^{\mathrm{per}}_{\eta,\delta}$, then with high constant probability, the algorithm rejects (either in the second step or in the third step). ■

Properties of the ideal object. For η and δ as above, and for a property of n-bit long strings Π, we let $\mathcal{D}^{\mathrm{per},\Pi}_{\eta,\delta} = \bigcup_{x^*\in\Pi} \mathcal{D}^{\mathrm{per}}_{\eta,\delta}(x^*)$. Building on the proof of Theorem 4.1, we get

Theorem 4.2 (testing noisy versions of a string in a predetermined set): *Let $\eta \in [0, 0.5)$ and $\delta \in [0, 1]$ be constants, and Π be a property of n-bit strings that can be tested using $Q(n, \epsilon)$ queries. Then, the property $\mathcal{D}^{\mathrm{per},\Pi}_{\eta,\delta}$ can be tested using $\mathrm{poly}(1/\epsilon) + \widetilde{O}(Q(n, e/2))$ queries.*

Proof: We combine the tester presented in the proof of Theorem 4.1 with an emulation of the tester for Π. Specifically, each query made by the latter tester is emulated by making corresponding queries to $O(\log Q(n, e/2))$ samples of the tested distribution (and taking a majority vote).

Evidently, any distribution X in $\mathcal{D}^{\mathrm{per},\Pi}_{\eta,\delta}$ is accepted with high probability, and in case X is $\epsilon/2$-far from $\mathcal{D}^{\mathrm{per}}_{\eta,\delta}$ it is rejected with high probability (by the first step). Hence, we are left with the case that X is $\epsilon/2$-close to $\mathcal{D}^{\mathrm{per}}_{\eta,\delta}(x^*)$ for some s that is $\epsilon/2$-H-far from Π (since otherwise X is ϵ-close to $\mathcal{D}^{\mathrm{per},\Pi}_{\eta,\delta}$).[25] Consequently, the emulated tester of Π will rejected with high probability. ■

4.2 Random cyclic shifts

For any string $x^* \in \{0,1\}^n$, we consider all distributions that are obtained by random (cyclic) shifts of the string x^*; that is, $\mathcal{D}^{\mathrm{cyc}}(x^*)$ contains the distribution X if there exists a (related) random variable $J \in \{0, 1, \ldots, n-1\}$ such that, for every j, with probability $\Pr[J=j]$ it holds that $X_i = x^*_{(i+j)_n}$ for every $i \in [n]$, where $(i+j)_n$ denotes $i+j$ if $i+j \in [n]$ and $i+j-n$ otherwise (i.e., $i+j > n$).

Theorem 4.3 (testing random shifts of a string): *The property $\mathcal{D}^{\mathrm{cyc}} \stackrel{\mathrm{def}}{=} \bigcup_{x^*\in\{0,1\}^n} \mathcal{D}^{\mathrm{cyc}}(x^*)$ can be tested using $O(1/\epsilon)$ samples and $\widetilde{O}(\sqrt{n}/\epsilon)$ queries.*

[25] Specifically, suppose that X is $\epsilon/2$-close to $\mathcal{D}^{\mathrm{per}}_{\eta,\delta}(x^*)$ for some x^* that is $\epsilon/2$-H-close to $x' \in \Pi$. Then, X is ϵ-close to $\mathcal{D}^{\mathrm{per}}_{\eta,\delta}(x')$.

Analogously to Theorem 4.2, we can also test the ideal string for a predetermined property provided that this property is invariant under cyclic shifts.

Proof: For the sake of the presentation, we describe a slightly simpler tester that makes $\widetilde{O}(\sqrt{n}/\epsilon^2)$ queries; the claimed tester can be obtained by employing Levin's Economical Work Investment Strategy [15, Sec. 8.2.4].

The tester is given oracle access to $t = O(1/\epsilon)$ samples, denoted $x^{(1)}, \ldots, x^{(t)}$, and consists of checking that each $x^{(i)}$ is a cyclic shift of $x^{(1)}$. Denoting the two strings by x and y, we check whether y is a cyclic shift of x by selecting $m = O(\sqrt{n \cdot \log t})$ random position indices, denoted $p_1, \ldots, p_m$, and $\ell = O(\epsilon^{-1} \cdot \log(n/\epsilon))$ offsets, denoted $o_1, \ldots, o_\ell$, querying both strings at locations $(p_j + o_k)_n$ for every $j \in [m]$ and $k \in [\ell]$, and accepting if and only if there exists $j, j' \in [m]$ such that $x_{(p_j+o_k)_n} = y_{(p_{j'}+o_k)_n}$ for every $k \in [\ell]$.

We first consider the case that X is in $\mathcal{D}^{\mathrm{cyc}}$; that is, suppose that X is in $\mathcal{D}^{\mathrm{cyc}}(x^*)$ for some $x^* \in \{0,1\}^n$. In this case, each of the samples (i.e., $x^{(i)}$) is a cyclic shift of x^*; that is, for each $i \in [t]$, there exists a shift σ_i such that $x^{(i)}_k = x^*_{(k+\sigma_i)_n}$ for every $k \in [n]$. Hence, for every $i \in \{2, \ldots, t\}$ and every pair $j, j' \in [m]$, with probability at least $1/n$ over the choice of $p_j, p_{j'} \in [n]$, it holds that $x^{(1)}_{(k+p_j)_n} = x^{(i)}_{(k+p_{j'})_n}$ for every $k \in [n]$. Since the events that correspond to different pairs of samples are pairwise independent, it follows that, for every $i \in \{2, \ldots, t\}$, with probability at least $1 - O(n/m^2)$ over the choice of $p_1, \ldots, p_m \in [n]$, it holds that $x^{(1)}_{(k+p_j)_n} = x^{(i)}_{(k+p_{j'})_n}$ for some $j, j' \in [m]$ and every $k \in [n]$. We conclude that, in this case (regardless of the choice of the $x^{(i)}$'s and the o_k's), the tester accepts with probability at least $2/3$.

Suppose, on the other hand, that X is ϵ-far from $\mathcal{D}^{\mathrm{cyc}}$. Fixing the first sample, denoted $x^{(1)}$, it follows that with probability at least $\epsilon/2$ it holds that (a sample of) X is $(\epsilon/2)$-H-far from being a shift of $x^{(1)}$. Hence, with probability at least 0.9 over the choice of the $x^{(i)}$'s, there exists an $i \in [t]$ such that $x^{(i)}$ is $(\epsilon/2)$-H-far from being a shift of $x^{(1)}$. It follows that, for each choice of $p_1, \ldots, p_m \in [n]$ and every $j, j' \in [m]$, it holds that $|\{k \in [n] : x^{(1)}_{(k+p_j)_n} \neq x^{(i)}_{(k+p_{j'})_n}\}| > \epsilon n/2$, and consequently

$$\mathrm{Pr}_{o_1,\ldots,o_\ell \in [n]} \left[(\exists k \in [\ell])\ x^{(1)}_{(p_j+o_k)_n} \neq x^{(i)}_{(p_{j'}+o_k)_n} \} | \right] \ > \ 1 - (1 - 0.5\epsilon)^\ell.$$

Recalling that $m = O(\sqrt{n \cdot \log t})$ and using a suitable $\ell = O(\epsilon^{-1} \cdot \log(n/\epsilon))$, it follows that with probability at least $1 - m^2 \cdot \exp(-\epsilon \cdot \ell) > 0.9$ (over the choice of $o_1, \ldots, o_k$) the tester detects that $x^{(i)}$ is not a cyclic shift of $x^{(1)}$. Therefore, in this case (i.e., X ϵ-far from $\mathcal{D}^{\mathrm{cyc}}$), the tester rejects with probability at least $2/3$. This completes the analysis of the slightly simpler tester, which performs $t \cdot m \cdot \ell = \widetilde{O}(\sqrt{n}/\epsilon^2)$ queries.

The claimed tester (which performs $\widetilde{O}(\sqrt{n}/\epsilon$ queries), follows by observing that if X is ϵ-far from $\mathcal{D}^{\mathrm{cyc}}$, then, for some $r \in [\log(1/\epsilon)]$, with probability at least $2^r \cdot \epsilon / O(\log(1/\epsilon))$ it holds that (a sample of) X is 2^{-r}-H-far from being a shift of $x^{(1)}$. Hence, it suffices to have $O(\log(1/\epsilon))$ iterations such that in the r-th iteration we use $t = \widetilde{O}(1/\epsilon)/2^r$ and $\ell = O(2^r \log(n/\epsilon))$. ■

Testing specific distributions over cyclic shifts. The property $\mathcal{D}^{\mathrm{cyc}}$ does not impose any constraint on the distribution over shifts. We next consider a natural variant, where this distribution is uniform.

Theorem 4.4 (testing uniformly random shifts of a string): *Let $\mathcal{D}^{\mathrm{Ucyc}}(x^*)$ denote the uniform distribution over the cyclic shifts of a string x^*. Then, the property $\mathcal{D}^{\mathrm{Ucyc}} \stackrel{\mathrm{def}}{=} \bigcup_{x^* \in \{0,1\}^n} \mathcal{D}^{\mathrm{Ucyc}}(x^*)$ can be tested using $\widetilde{O}(n^{2/3}/\epsilon^3)$ queries.*

Theorem 4.4 is proved by a reduction to a more general problem, and it is indeed possible that a more efficient tester exists. For every fixed distrution I over $\{0, 1, ..., n-1\}$, the foregoing reduction and result extends to testing the property $\mathcal{D}^{\mathrm{cyc}_I} \stackrel{\mathrm{def}}{=} \bigcup_{x^* \in \{0,1\}^n} \mathcal{D}^{\mathrm{cyc}_I}(x^*)$ such that $\mathcal{D}^{\mathrm{cyc}_I}(x^*)$ denote the distribution obtained by shifting x^* by I positions (to the right).

Proof: We reduce the current problem to testing the equality between two distributions over $\{0,1\}^n$ such that one of the distributions has support size at most n, while noting that a tester for the latter problem is provided in Theorem 5.2. Specifically, given s samples, denoted $x^{(1)}, \ldots, x^{(s)}$, of a distribution X over n-bit strings, we consider the distribution $Y \stackrel{\mathrm{def}}{=} \mathcal{D}^{\mathrm{Ucyc}}(x^{(1)})$, and test equality between X and Y, where we emulate samples to X by using $x^{(2)}, \ldots, x^{(s)}$, and emulate samples to Y by using (random shifts of) $x^{(1)}$. Note that Y has support of size at most n, which suffices when using the furthermore clause of Theorem 5.2.

The complexity of our tester equals the complexity of the tester of Theorem 5.2, and its analysis reduces to the latter. Specifically, if X is in $\mathcal{D}^{\mathrm{Ucyc}}$, then, for every possible $x^{(1)}$ drawn from X, it holds that $X \equiv \mathcal{D}^{\mathrm{Ucyc}}(x^{(1)})$, and it follows that our tester accepts (whp). On the other hand, if X is ϵ-far from $\mathcal{D}^{\mathrm{Ucyc}}$, then for every x^* it holds that X is ϵ-far from $\mathcal{D}^{\mathrm{Ucyc}}(x^*)$, and it follows that our tester rejects (whp). ■

4.3 Random isomorphic copies of a graph

Using a sublinear-query tester for graph isomorphism, we can adapt the ideas underlying the proof of Theorem 4.3 to test distributions of strings that describe the adjacency matrices of random isomorphic copies of a graph. That is, we consider n-bit long strings that describe the adjacency matrices of $\sqrt{n}$-vertex graphs. Specifically, for every string $x^* \in \{0,1\}^n$, we consider the graph G_{x^*} described by x^* and any distribution on isomorphic copies of G_{x^*}; that, $\mathcal{D}^{\mathrm{iso}}(x^*)$ contains the distribution X if X is a distribution over strings that describe graphs that are isomorphic to G_{x^*}.

Recall that testing isomorphism of k-vertex graphs in the dense graph model, which uses the adjacency matrix representation, has query complexity $\mathrm{poly}(1/\epsilon) \cdot \widetilde{O}(k^{5/4})$; see [13], where the dependence on ϵ is mentioned at the end of Section 1. In contrast, the query complexity of the tester of [20] is $k^{1+o(1)}$ provided that $\epsilon = \omega((\log\log k)/(\log k)^{1/2})$.

Theorem 4.5 (testing random isomorphic copies of a graph): *The property $\mathcal{D}^{\mathrm{iso}} \stackrel{\mathrm{def}}{=} \bigcup_{x^* \in \{0,1\}^n} \mathcal{D}^{\mathrm{iso}}(x^*)$ can be tested using $O(1/\epsilon)$ samples and $\mathrm{poly}(1/\epsilon) \cdot \widetilde{O}(n^{5/8})$ queries.*

Note that testing isomorphism in the dense graph model is reducible to testing $\mathcal{D}^{\mathrm{iso}}$ in the DoHO model. We also mention that, analogously to Theorem 4.2, one can also test the ideal string for a predetermined graph property (since a graph property is invariant under graph isomorphism).

Proof: Analogously to the proof of Theorem 4.3, the tester takes $t = O(1/\epsilon)$ samples, denoted $x^{(1)}, \ldots, x^{(t)}$, and checks whether all $x^{(i)}$'s describe graphs that are isomorphic to the graph described by $x^{(1)}$. Hence, for each $i \in \{2, \ldots, t\}$, we check whether $G_{x^{(i)}}$ is isomorphic to $G_{x^{(1)}}$, by invoking a graph isomorphism tester for the dense graph model. Specifically, we use the tester

presented in [13], while setting the proximity parameter to $\epsilon/2$ (and the error probability of the test to $o(\epsilon)$).

Note that if X is in $\mathcal{D}^{\text{iso}}$, then each invocation of the isomorphism test accept with probability $1 - o(\epsilon)$. On the other hand, if X is ϵ-far from $\mathcal{D}^{\text{iso}}$, then, for any choice of $x^{(1)}$ and every $i \in \{2, \ldots, t\}$, with probability at least $\epsilon/2$ it holds that $G_{x^{(i)}}$ is $\epsilon/2$-far from being isomorphic to $G_{x^{(1)}}$, where the latter distance is in the dense graph model. Hence, the corresponding invocation of the graph isomorphism tester rejects (whp), and so does our tester. ■

What about the bounded-degree graph model? We could have adapted the proof strategy of Theorem 4.5 to bounded-degree graphs that are represented by their incidence functions. However, unfortunately, we do not know of a sublinear-query tester for graph isomorphism in that model (see [16]).

5 Tuples of Distributions

Our notion of testing properties of distributions over huge objects (as captured by Definition 1.2), extends easily to testing tuples of such distributions.

5.1 The definition

Following the convention stated in Section 1.4, we refer to distributions via the corresponding random variables.

Definition 5.1 (testing properties of t-tuples of huge distributions): *Let $\mathcal{D}$ be a property of t-tuples of distributions, which are each as in Definition 1.2, and $s : \mathbb{N} \times (0,1] \to \mathbb{N}$. A* tester*, denoted T,* of sample complexity s for the property $\mathcal{D}$ *is a probabilistic machine that, on input parameters n and ϵ, and oracle access to a sequence of $s(n,\epsilon)$ samples drawn from each of the t unknown distributions $X^{(1)}, \ldots, X^{(t)} \in \{0,1\}^n$, satisfies the following two conditions.*

1. The tester accepts tuples that belong to $\mathcal{D}$: *If $(X^{(1)}, \ldots, X^{(t)})$ is in $\mathcal{D}$, then*

$$\Pr_{x^{(1,1)},\ldots,x^{(1,s)} \sim X^{(1)};\ldots;x^{(t,1)},\ldots,x^{(t,s)} \sim X^{(t)}}[T^{x^{(1,1)},\ldots,x^{(1,s)},\ldots,x^{(t,1)},\ldots,x^{(t,s)}}(n,\epsilon)=1] \geq 2/3,$$

   where $s = s(n,\epsilon)$ and $x^{(i,1)}, \ldots, x^{(i,s)}$ are drawn independently from the distribution $X^{(i)}$.

2. The tester rejects tuples that are far from $\mathcal{D}$: *If $(X^{(1)}, \ldots, X^{(t)})$ is ϵ-far from $\mathcal{D}$* (i.e., for every $(Y_1, \ldots, Y_t)$ in $\mathcal{D}$ the average distance (according to Definition 1.1) between X_j and Y_j, where $j \in [t]$, is greater than ϵ)*, then*

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

   where $s = s(n,\epsilon)$ and $x^{(i,1)}, \ldots, x^{(i,s)}$ are as in the previous item.

The query complexity of such tester is defined as in the case of testing a single distribution (i.e., $t = 1$). Indeed, Definition 1.2 is a special case of Definition 5.1 (i.e., $t = 1$).

5.2 Testing equality

This is indeed the archetypal example for the case of $t = 2$. Using any tester for the standard model, we obtain a tester for the DoHO model by querying all samples at a logarithmic (in the support size) number of locations. Hence, this tester requires an upper bound on the size of the supports of the tested distributions.

Theorem 5.2 (testing equality of distributions in the DoHO model): *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 they are identical and the case that they are ϵ-far from one another* (according to Definition 1.1) *using $\widetilde{O}(m^{2/3}/\epsilon^3)$ queries and $O(m^{2/3}/\epsilon^2)$ samples. Furthermore, the claim holds even if only the support size of one of the distributions is upper-bounded by m.*

Actually, the sample complexity is $s = O(\max(\epsilon^{-4/3}m^{2/3}, \epsilon^{-2}m^{1/2}))$, and the query complexity is $\widetilde{O}(s/\epsilon)$.

Proof: The key observation is that if X is ϵ-far from Y (according to Definition 1.1), then, with high probability over the choice of a random $O(\epsilon^{-1}\log m)$-subset $J \subset [n]$, the total variation distance between X_J and Y_J is at least 0.3ϵ. This observation is proved next.

Claim 5.2.1 (typically, the distance between X and Y is preserved by the distance between X_J and Y_J): *Suppose that X is ϵ-far from Y, and that both distributions have support size at most m. Then, with probability $1 - o(1)$ over the choice of $J \in \binom{[n]}{O(\epsilon^{-1}\log m)}$, the total variation distance between X_J and Y_J is greater than 0.3ϵ. Actually, X_J is 0.3ϵ-far from Y_J, even according to Definition 1.1.*

Proof: We start by letting $x^{(1)}, \ldots, x^{(m')}$ (resp., $y^{(1)}, \ldots, y^{(m'')}$) denote the elements in the support of X (resp., Y), where $m' \leq m$ (resp., $m'' \leq m$). Next, we note that for every $i \in [m']$ and $k \in [m'']$, when selecting uniformly an $O(t/\epsilon)$-subset J, with probability at least $1-2^{-t}$, the relative Hamming distance between $x_J^{(i)}$ and $y_J^{(k)}$ it at least $\min(0.5 \cdot \Delta_H(x^{(i)}, y^{(k)}), \Delta_H(x^{(i)}, y^{(k)}) - 0.2\epsilon)$, where the first (resp., second) term refers to the case that $\Delta_H(x^{(i)}, y^{(k)}) > 0.2\epsilon$ (resp., $\Delta_H(x^{(i)}, y^{(k)}) \leq 0.2\epsilon$) and followed by a multiplicative Chernoff bound (resp., by triviality).

Now, letting $t = O(\log m)$ and using a union bound (over all $(i,k) \in [m'] \times [m'']$), with probability $1 - o(1)$ over the choice of $J \in \binom{[n]}{O(t/\epsilon)}$, for every $i \in [m']$ and every mapping $\mu : [m'] \to [m'']$, it holds that

$$\Delta_H(x_J^{(i)}, y_J^{(\mu(i))}) \geq 0.5 \cdot \Delta_H(x^{(i)}, y^{(\mu(i))}) - 0.2\epsilon. \tag{8}$$

(We stress the order of quantifiers: With high probability over the choice of J, Eq. (8) holds for every $i \in [m']$ and $\mu : [m'] \to [m'']$.)[26] Hence, with probability $1 - o(1)$ over the choice of J, for every mapping $\mu : [m'] \to [m'']$ and every probability distribution $p : [m'] \to [0,1]$, it holds that

$$\begin{aligned} \sum_{i \in [m']} p(i) \cdot \Delta_H(x_J^{(i)}, y_J^{(\mu(i))}) &\geq \sum_{i \in [m']} p(i) \cdot (0.5 \cdot \Delta_H(x^{(i)}, y^{(\mu(i))}) - 0.2\epsilon) && (9) \\ &\geq 0.5 \cdot \sum_{i \in [m']} p(i) \cdot \Delta_H(x^{(i)}, y^{(\mu(i))}) \;-\; 0.2\epsilon\,. && (10) \end{aligned}$$

[26] Indeed, although there are $(m'')^{m'}$ different μ's, the union bounds is only over the pairs $(i, \mu(i)) \in [m'] \times [m'']$.

Suppose now that instead of a (deterministic) mapping $\mu : [m'] \to [m'']$, we consider a randomized process $\mu : [m'] \to [m'']$, where for each $i \in [m']$ the random variable $\mu(i)$ represents a distribution over $[m'']$. Then, Eq. (9)&(10) extends to any random process μ, where we consider expected distances (with expectation taken over the random choices of μ). In particular, letting $p(i) = \Pr[X = x^{(i)}]$, with probability $1 - o(1)$ over the choice of J, for any randomized process $\mu : [m'] \to [m'']$, it holds that

$$\sum_{i \in [m']} p(i) \cdot \mathrm{E}_\mu \left[\Delta_H(x_J^{(i)}, y_J^{(\mu(i))}) \right] \geq 0.5 \cdot \sum_{i \in [m']} p(i) \cdot \mathrm{E}_\mu \left[\Delta_H(x^{(i)}, y^{(\mu(i))}) \right] - 0.2\epsilon. \tag{11}$$

We observe that for any choice of μ that maps X to Y (i.e., $\sum_{i \in [m']} p(i) \cdot \Pr[\mu(i) = k] = \Pr[Y = y^{(k)}]$ for every $k \in [m'']$), the main sum in the r.h.s of Eq. (11) is lower-bounded by the distance between X and Y (according to Definition 1.1; cf. Eq. (15)). Recalling that the latter distance is greater than ϵ, it follows that (for any μ that maps X to Y) the l.h.s of Eq. (11) is greater than $0.5 \cdot \epsilon - 0.2\epsilon = 0.3\epsilon$. On the other hand, we observe that the minimum over μ's that map X_J to Y_J of the l.h.s of Eq. (11) captures the distance between X_J and Y_J (according to Definition 1.1), which lower-bounds the total variation distance between X_J and Y_J. Hence, with probability $1 - o(1)$ over the choice of J, the total variation distance between X_J and Y_J is greater than 0.3ϵ. ■

The tester. In light of the above, our tester proceeds as follows. For $s = O(\max(\epsilon^{-4/3}m^{2/3}, \epsilon^{-2}m^{1/2}))$, given oracle access to s samples, denoted $u^{(1)}, \ldots, u^{(s)}$ and $v^{(1)}, \ldots, v^{(s)}$, of each of the two distributions, the tester selects an $O(\epsilon^{-1} \log m)$-subset $J \subset [n]$ uniformly at random, and queries each sample at the bits in J. Denoting the resulting strings (i.e., the restrictions of the sampled strings to J) by $u_J^{(1)}, \ldots, u_J^{(s)}$ and $v_J^{(1)}, \ldots, v_J^{(s)}$, our tester invokes the standard tester (with proximity parameter 0.3ϵ), and provides these strings as the expected samples.

Note that if $X \equiv Y$, then $X_J \equiv Y_J$ always holds, and the standard tester accepts (whp). On the other hand, by the foregoing observation, if X is ϵ-far from Y (according to Definition 1.1), then, with high probability over the choice of J, it holds that X_J is 0.3ϵ-far from Y_J (in total variation distance), and in this case the standard tester rejects (whp).

The foregoing establishes the main claim. Turning to the furthermore claim (i.e., only assuming that Y has support size at most m), note that we cannot afford a union bound over $[m'] \times [m'']$. Still, letting $t = O(\log(m/\epsilon))$ and assuming only $m'' \leq m$, we replace the assertion regarding Eq. (8) by the assertion that, for every $i \in [m']$, with probability $1 - o(\epsilon)$ over the choice of $J \in \binom{[n]}{O(t/\epsilon)}$, for every mapping $\mu : [m'] \to [m'']$ it holds that

$$\Delta_H(x_J^{(i)}, y_J^{(\mu(i))}) \geq 0.5 \cdot \Delta_H(x^{(i)}, y^{(\mu(i))}) - 0.2\epsilon. \tag{12}$$

Fixing any probability distribution $p : [m'] \to [0, 1]$, we call J good if for every mapping $\mu : [m'] \to [m'']$ it holds that Eq. (12) is satisfies for a set of i's that has weight at least $1 - 0.1\epsilon$ under p (i.e., letting I denote the set of these i's, it holds that $\sum_{i \in I} p(i) \geq 1 - 0.1\epsilon$). Using an averaging argument, it follows that $1 - o(1)$ of the J's are good. Hence, with probability $1 - o(1)$ over the choice of J, for every $\mu : [m'] \to [m'']$ it holds that

$$\begin{aligned} \sum_{i \in [m']} p(i) \cdot \Delta_H(x_J^{(i)}, y_J^{(\mu(i))}) &\geq \sum_{i \in [m']} p(i) \cdot (0.5 \cdot \Delta_H(x^{(i)}, y^{(\mu(i))}) - 0.2\epsilon) - 0.05\epsilon & (13) \\ &\geq 0.5 \cdot \sum_{i \in [m']} p(i) \cdot \Delta_H(x^{(i)}, y^{(\mu(i))}) - 0.25\epsilon. & (14) \end{aligned}$$

where the term 0.05ϵ accounts for contribution of the i's that do not satisfy Eq. (12). That is, Eq. (9)&(10) is replaced by Eq. (13)&(14). Proceeding as in the proof of the Claim 5.2.1, we get

Claim 5.2.2 (Claim 5.2.1, extended): *Suppose that X is ϵ-far from Y, and that Y has support size at most m. Then, with probability $1 - o(1)$ over the choice of $J \in \binom{[n]}{O(\epsilon^{-1} \log m)}$, the total variation distance between X_J and Y_J is greater than 0.25ϵ. Actually, X_J is 0.25ϵ-far from Y_J, even according to Definition 1.1.*

The furthermore claim follows by observing that the equality tester (for the standard model) of [12] works also when the support size of only one of the tested distributions is upper-bounded.[27] Specifically, using the presentation of [15, Sec. 11.2–11.3], we observe that the support size is only used in the proof of [15, Cor. 11.21], when upper-bounding the total variation distance between two distributions by the norm-2 of their difference. But essentially the same upper bound (on the total variation distance) holds also if only the support of one of the distributions is upper-bounded.[28] (For more details, see Appendix A.4.) ■

Acknowledgements

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

References

[1] Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Inference under information constraints I: lower bounds from chi-square contraction. *IEEE Transactions on Information Theory*, 66(12):7835–7855, 2020.

[2] Tugkan Batu. *Testing properties of distributions*. PhD thesis, Computer Science department, Cornell University, 2001.

[3] 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.

[27] This is not a generic claim regarding any such tester: See Footnote 32 in Appendix A.4.

[28] Specifically, let $p : S \to [0,1]$ be the probability function representing one distribution and $q : U \to [0,1]$ be the function representing the other distribution, where $S \subseteq U$. Then,

$$\begin{aligned}
\sum_{i \in U} |p(i) - q(i)| &= 2 \cdot \sum_{i \in U : p(i) > q(i)} |p(i) - q(i)| \\
&\leq 2 \cdot \sum_{i \in S} |p(i) - q(i)| \\
&\leq 2 \cdot \sqrt{|S|} \cdot \left(\sum_{i \in S} |p(i) - q(i)|^2 \right)^{1/2} \\
&\leq 2 \cdot \sqrt{|S|} \cdot \left(\sum_{i \in U} |p(i) - q(i)|^2 \right)^{1/2}
\end{aligned}$$

which equals $2\sqrt{|S|} \cdot \|p - q\|_2$.

[4] 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.

[5] 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.

[6] 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 [5].

[7] Eli Ben-Sasson, Oded Goldreich, Prahladh Harsha, Madhu Sudan, and Salil Vadhan. Robust PCPs of proximity, shorter PCPs, and applications to coding. *SIAM Journal on Computing*, 36(4):889–974, 2006.

[8] 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.

[9] Sourav Chakraborty, Eldar Fischer, Arijit Ghosh, Gopinath Mishra, and Sayantan Sen. Testing of index-invariant properties in the huge object model. Technical Report TR22-155, Electronic Colloquium on Computational Complexity (ECCC), 2022.

[10] Alessandro Chiesa, Tom Gur, and Igor Shinkar. Relaxed locally correctable codes with nearly-linear block length and constant query complexity. In *Proceedings of the Thirty-First Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1395–1411, 2020.

[11] Ilias Diakonikolas, Daniel Kan, and Alistair Stewart. Sharp bounds for generalized uniformity testing. Technical Report TR17-132, Electronic Colloquium on Computational Complexity (ECCC), 2017.

[12] Ilias Diakonikolas and Daniel Kane. A new approach for testing properties of discrete distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 685–694, 2016.

[13] Eldar Fischer and Arieh Matsliah. Testing graph isomorphism. *SIAM Journal on Computing*, 38(1):207–225, 2008.

[14] Oded Goldreich. On multiple input problems in property testing. In *Proceedings of the Eighteenth International Workshop on Randomization and Computation (RANDOM)*, pages 704–720, 2014.

[15] Oded Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017.

[16] Oded Goldreich. Testing isomorphism in the bounded-degree graph model. Technical Report TR19-102, Electronic Colloquium on Computational Complexity (ECCC), 2019.

[17] 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.

[18] Oded Goldreich and Dana Ron. Testing distributions of huge objects. Technical Report TR21-133, Electronic Colloquium on Computational Complexity (ECCC), 2021.

[19] Tom Gur, Govind Ramnarayan, and Ron Rothblum. Relaxed locally correctable codes. In *Proceedings of the Ninth Innovations in Theoretical Computer Science conference (ITCS)*, pages 27:1–27:11, 2018.

[20] Krzysztof Onak and Xiaorui Sun. The query complexity of graph isomorphism: Bypassing distribution testing lower bounds. In *Proceedings of the Fifteeth Annual ACM Symposium on the Theory of Computing (STOC)*, pages 165–171, 2018.

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

[22] 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.

[23] Gregory Valiant and Paul Valiant. Estimating the unseen: Improved estimators for entropy and other properties. *Journal of the ACM*, 64(6), 2017.

[24] Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927—-1968, 2011.

Appendices

The appendices vary in nature. Appendix A.1 presents a proof of a well-known fact, Appendix A.2 presents ramifications on a simple result stated in the main text, Appendix A.3 presents a possible avenue towards a stronger version of Theorem 1.6, and Appendix A.4 presents a result that is only implicit in prior work (and may be of independent interest).

A.1 Earth mover distance with inequality measure

A general definition of the earth model distance associates a distance function f with the domain, and considers the cost of the best randomized process that transforms one distribution to another, where the cost of the (randomized) process M, which moves the distribution X to the distribution $M(X)$, is defined as

$$\sum_x \Pr[X=x] \cdot \mathrm{E}[f(x, M(x))]. \tag{15}$$

Definition 1.1 is derived by letting f be the relative Hamming distance between strings.[29] Here, we consider the crude inequality function; that is, $f(x,y) = 1$ if $x \neq y$ and $f(x,x) = 0$. We prove that the earth mover's distance with respect to the inequality function equals the total variation distance, where the total variation distance between X and Y equals $\max_S\{\Pr[X \in S] - \Pr[Y \in S]\}$, which equals half $\sum_z |\Pr[X=z] - \Pr[Y=z]|$.

[29] Here the random process M replaces the $w_{x,y}$'s that appear in Eq. (2); specifically, $\Pr[M(x)=y] = \frac{w_{x,y}}{\Pr[X=x]}$.

Claim A.1 (on the earth mover's distance with the inequality measure): *The earth mover's distance with respect to the inequality measure* (i.e., $f(x,y)=1$ if $x \neq y$ and $f(x,x)=0$) *coincides with the total variation distance.*

Proof: For $S = \{z : \Pr[Y=z] \geq \Pr[X=z]\}$, consider the randomized process M such that $M(z)=z$ if $z \in S$ and $\Pr[M(z)=z] = \frac{\Pr[Y=z]}{\Pr[X=z]}$ otherwise, where the excess probability mass is distributed among the strings in S so that $M(X) \equiv Y$. Note that the cost of this M equals

$$\sum_{x \notin S} \Pr[X=x] \cdot \left(1 - \frac{\Pr[Y=z]}{\Pr[X=z]}\right)$$

which equals the total variation distance. Hence, the earth mover distance (w.r.t inequality) is upper-bounded by the total variation distance. On the other hand, the earth mover distance (w.r.t inequality) is lower-bounded by the total variation distance, since the latter measures the probability mass that has to be moved from $S = \{z : \Pr[X=z] \geq \Pr[Y=z]\}$ to $\overline{S}$. ■

A.2 Ramifications regarding Theorem 1.4

We first restate the basic claim of Theorem 1.4 and improve it in the special case of "nice" query complexity bounds. Specifically, we prove the following.

Theorem A.2 (Theorem 1.4, restated and improved): *Let Π and $\mathcal{D}_\Pi$ be as in Theorem 1.4. If the query complexity of testing Π is q, then the query complexity of testing $\mathcal{D}_\Pi$ is at most q' such that $q'(n,\epsilon) = \widetilde{O}(1/\epsilon) \cdot q(n,\epsilon/2)$. Furthermore, if $q(n,\epsilon) \geq 2^c \cdot q(n,2\epsilon)$ holds for some constant $c \geq 1$ and all $\epsilon \in [O(1/n), \Omega(1)]$, then $q'(n,\epsilon) = O(q(n,\epsilon))$ if $c > 1$ and $q'(n,\epsilon) = O(\log(1/\epsilon))^3 \cdot q(n,\epsilon)$ otherwise* (i.e., if $c=1$)*. In both cases, the tester uses $O(1/\epsilon)$ samples.*

Proof: Recall that the proof of the main claim relied on the observation that if the tested distribution P is ϵ-far from $\mathcal{D}_n$ (according to Definition 1.1), then, $x \sim P$ is $\epsilon/2$-H-far from Π_n with probability at least $\epsilon/2$. (This is the case, since otherwise, letting $f(x)$ be a string in Π_n that is closest to x in Hamming distance yields a distribution $Q(y) = \sum_{x \in f^{-1}(y)} P(x)$ that is in $\mathcal{D}_\Pi$ and is $(\frac{\epsilon}{2} \cdot 1 + (1-\frac{\epsilon}{2}) \cdot \frac{\epsilon}{2})$-close to P.)

The furthermore claim is proved by employing Levin's Economical Work Investment Strategy [15, Sec. 8.2.4]. Specifically, the key observation is that there exists $i \in [\lceil \log_2(16/\epsilon) \rceil]$ such that with probability at least $2^{-i}/(i+3)^2$ it holds that $x \sim X$ is $2^{i-3} \cdot \epsilon$-H-far from Π_n. In this case, the query complexity is $\sum_{i \leq \ell} O(i^2 \cdot 2^i) \cdot q(n, 2^{i-3}\epsilon)$, where $\ell = \lceil \log_2(16/\epsilon) \rceil$. Using $q(n, 2^{i-3}\epsilon) \leq (2^{i-3})^{-c} \cdot q(n,\epsilon)$, the foregoing sum is upper-bounded by is $\sum_{i \leq \ell} i^2 \cdot 2^{-(c-1)\cdot i} \cdot O(q(n,\epsilon))$, and the claim follows. ■

Generalization. Towards the following generalization of Theorem A.2, we consider a generalization of property testing of strings. In this generalization the property Π_n is partitioned into $m = m(n)$ parts and, when accepting, the tester also indicates the index of the part in which the object resides. For example, the set of low-degree multi-variate polynomials can be partitioned according to their value at a fixed point, and coupled with a generalized tester of low complexity. Generalizing Theorem A.2, we get –

Theorem A.3 (Theorem A.2, generalized): *For $m : \mathbb{N} \to \mathbb{N}$, let $\Pi = \{\Pi_n\}_{n\in\mathbb{N}}$ be an m-partitioned property that is testable in the generalized sense within query complexity $q(n,\epsilon)$. Let $\mathcal{D} = \{\mathcal{D}_n\}_{n\in\mathbb{N}}$ be a property of distributions over $[m(n)]$ that is testable in the standard model with sample complexity $s(m,\epsilon) = \Omega(1/\epsilon)$, and let $\mathcal{C} = \mathcal{C}_{\Pi,\mathcal{D}}$ be a property of distributions such that X is in $\mathcal{C}$ if and only if X consists of selecting an index $i \in [m]$ according to some distribution in $\mathcal{D}$ and outputting an element selected according to an arbitrary distribution that is supported by a subset of $\Pi_n^{(i)}$. Then, the query complexity of testing $\mathcal{C}$ is at most q' such that $q'(n,\epsilon) = \widetilde{O}(s(m(n),0.3\epsilon)) \cdot q(n,0.3\epsilon)$.*

In particular, if $\Pi_n = \cup_{i\in[m]}\Pi_n^{(i)}$ such that each $\Pi_n^{(i)}$ is testable with $q(n,\epsilon)$ queries and the $\Pi_n^{(i)}$'s are δ-H-far apart, then we can obtain a generalized tester of query complexity $\widetilde{O}(m(n)) \cdot q(n,\delta) + O(q(n,\epsilon))$ for Π_n.

Proof: We combine the tester for Π, denoted T, with the tester for $\mathcal{D}$, while invoking both with proximity parameter $\epsilon/2$, and reducing the error probability of T to $o(1/s(m(n),0.3\epsilon))$. Hence, when invoked on input $(n,0.3\epsilon)$ and given oracle access to $x \in \{0,1\}^n$, with probability at least $1 - o(1/s(m(n),0.3\epsilon))$, the tester T outputs i if $x \in \Pi_n^{(i)}$, where $\Pi_n^{(i)}$ is the i^{th} part of Π_n, and rejects (with output 0) if x is 0.3ϵ-H-far from Π_n. Furthermore, with probability at least $1 - o(1/s(m(n),0.3\epsilon))$, the tester T does not output i if x is 0.3ϵ-H-far from $\Pi_n^{(i)}$. Denoting the (majority) output of T by $\chi(x)$, we may assume that either $\chi(x) = 0$ (indicating rejection) or x is $\epsilon/2$-H-close to $\Pi_n^{(\chi(x))}$. The key observation is that if X is ϵ-far from $\mathcal{C}$ (according to Definition 1.1), then either X is 0.7ϵ-far from being distributed over Π_n (according to Definition 1.1) or $\chi(X)$ is 0.3ϵ-TV-far from $\mathcal{D}$. Hence, we get an adequate tester that, on access to the samples $x^{(1)},\ldots,x^{(s)}$, where $s = s(m(n),0.3\epsilon)$, invokes T on each of these samples, obtaining the answers $a_1,\ldots,a_s \in \{0,1,\ldots,m(n)\}$, rejects if any of these a_i's equals 0, and outputs the verdict of the distribution tester (i.e., the $\mathcal{D}$-tester) on $(a_1,\ldots,a_s) \in [m(n)]^s$ otherwise.

To see that the foregoing tester is correct, note that if X in $\mathcal{C}$, then $X = Y_I$ such that I is in $\mathcal{D}$ and each Y_i is supported by $\Pi_n^{(i)}$. It follows that, in this case, X is accepted with high probability. On the other hand, if X is accepted with high probability, then $\chi(X)$ is 0.3ϵ-TV-close to a distribution in $\mathcal{D}$, and, with probability at least $1 - 0.3\epsilon$ over the choice of $x \sim X$, it holds that x is 0.3ϵ-H-close to $\Pi_n^{(\chi(x))}$. It follows that X is ϵ-close to $\mathcal{C}$. ■

A.3 Towards a stronger version of Theorem 1.6

Recall that, for any property $\mathcal{D}$ that is closed under mapping, Theorem 1.6 upper-bounds the query complexity of testing $\mathcal{D}$ in the DoHO model in terms of the sample complexity of testing $\mathcal{D}$ in the standard model. This leaves open the question of whether the query complexity of testing $\mathcal{D}$ in the DoHO model can be similarly upper-bounded in terms of the sample complexity of testing $\mathcal{D}$ in the DoHO model, which may be lower than the sample complexity of testing $\mathcal{D}$ in the standard model. A possible avenue towards establishing such a result is resolving positively the following open problem.

Open Problem A.4 (preservation of distances under a random relabeling): *Suppose that $\mathcal{D}$ is a property of distributions over n-bit strings that is closed under mapping. Is it the case that if X is ϵ-far from $\mathcal{D}$, then, with high probability over the choice of a random bijection $\pi : \{0,1\}^n \to \{0,1\}^n$, it holds that $\pi(X)$ is $\Omega(\epsilon)$-far from $\mathcal{D}$? We stress that the distances here are according to Definition 1.1 and that the hidden constant in the Ω-notation is universal.*

A positive answer to Problem A.4 would allow to convert a tester for $\mathcal{D}$ in the DoHO model into one that only considers the collision pattern among the samples. Specifically, given a collision pattern among s samples, the latter tester will generate at random a sequence of s samples that fits the given collision pattern, and invoke the original tester on this sequence of samples. In such a case, we can apply the strategy used in the proof of Theorem 1.6 to the resulting tester.

We were able to establish a positive answer to Problem A.4 in the special case that the support of X has size at most $2^{(0.5-\Omega(1))\cdot n}$. In fact, in that case, we prove a stronger result (where, for simplicity, 0.49 stands for $0.5-\Omega(1)$).

Proposition A.5 (a partial answer to Problem A.4): *Suppose that $\mathcal{D}$ is a property of distributions over n-bit strings that is closed under mapping, and that X has support size at most $2^{0.49n}$. Then, if X is ϵ-far from $\mathcal{D}$ in total variation distance, then, with high probability over the choice of a random bijection $\pi:\{0,1\}^n\to\{0,1\}^n$, it holds that $\pi(X)$ is $\Omega(\epsilon)$-far from $\mathcal{D}$ according to Definition 1.1.*

The restriction on X is essential; see Section 2.2.

Proof: The key observation is that, for some constant $\delta>0$, with high probability over the choice of a random bijection $\pi:\{0,1\}^n\to\{0,1\}^n$, it holds that the elements in the support of $\pi(X)$ are at relative Hamming distance at least δ. Fixing any such π, we let C denote the support of $\pi(X)$ and note that $\min_{w\neq w'\in C}\{\Delta_H(w,w')\}\geq\delta$. Assuming that $X'=\pi(X)$ is ϵ'-close to $\mathcal{D}$ *according to Definition 1.1*, we shall show that X' is $\frac{2}{\delta}\cdot\epsilon'$-close to $\mathcal{D}$ *in total variation distance*. (It follows that X is $\frac{2}{\delta}\cdot\epsilon'$-close to $\mathcal{D}$ *in total variation distance*.) Specifically, we consider a distribution Y in $\mathcal{D}$ such that X' is ϵ'-close to Y according to Definition 1.1, and show that a related distribution Y' that is also in $\mathcal{D}$ is $\frac{2}{\delta}\cdot\epsilon'$-close to X' in total variation distance. In particular, we shall replace Y by the distribution Y' of the strings in C that are closest to Y.

Claim A.5.1 (the effect of correction to the closest element of C): *Suppose that X' is supported on a set C such that $\min_{w\neq w'\in C}\{\Delta_H(w,w')\}\geq\delta$, and that Y is ϵ'-close to X' according to Definition 1.1. Then, $Y'=\mathtt{corr}(Y)$ is $\frac{2}{\delta}\cdot\epsilon'$-close to X' in total variation distance, where $\mathtt{corr}(y)$ denotes a string in C that is closest to y.*

Recalling that in our application Y is in $\mathcal{D}$, it follows that $\mathtt{corr}(Y)$ is in $\mathcal{D}$, since $\mathcal{D}$ is closed under mapping. Hence, X' is $\frac{2}{\delta}\cdot\epsilon'$-close to $\mathcal{D}$.

Proof: Intuitively, replacing Y by $\mathtt{corr}(Y)$ may increase the distance from X' according to Definition 1.1, but not too much (i.e., for $x'\in C$, it holds that $\Delta_H(x',\mathtt{corr}(y))\leq 2\cdot\Delta_H(x',y)$). The key observation is that the distance of $Y'=\mathtt{corr}(Y)$ to X' (according to Definition 1.1) is due solely to strings that are at Hamming distance at least δ. This implies that the total variation distance between Y' and X' is at least a δ fraction of the distance between Y' and X' according to Definition 1.1. Furthermore, we shall show that the total variation distance between Y' and X' is at least a $\delta/2$ fraction of the distance between Y and X' according to Definition 1.1. The actual proof follow.

For $w_{x',y}$'s as in Definition 1.1 (i.e., the minimum sequence of non-negative numbers that satisfies $\sum_y w_{x',y}=\Pr[X'=x']$ and $\sum_{x'} w_{x',y}=\Pr[Y=y]$), the hypothesis means that

$$\sum_{x',y\in\{0,1\}^n} w_{x',y}\cdot\Delta_H(x',y)\leq\epsilon'.$$

Recall that $w_{x',y} > 0$ only if $x' \in C$, and that $\texttt{corr}(y)$ denote a string in C that is closest to y. Then, the foregoing sum equals

$$\sum_{y\in\{0,1\}^n} w_{\texttt{corr}(y),y} \cdot \Delta_H(\texttt{corr}(y),y) + \sum_{y\in\{0,1\}^n} \sum_{x'\in C\setminus\{\texttt{corr}(y)\}} w_{x',y} \cdot \Delta_H(x',y)$$

which is lower-bounded by

$$\sum_{y\in\{0,1\}^n} \sum_{x'\in C\setminus\{\texttt{corr}(y)\}} w_{x',y} \cdot \frac{\delta}{2}$$

because $\Delta_H(x',y) \geq \Delta_H(x',\texttt{corr}(y)) - \Delta_H(\texttt{corr}(y),y)$ and $\Delta_H(x',y) \geq \Delta_H(\texttt{corr}(y),y)$ for any $x' \in C$, which implies $\Delta_H(x',y) \geq \Delta_H(x',\texttt{corr}(y))/2$, whereas $\Delta_H(x',\texttt{corr}(y)) \geq \delta$ (for $x' \in C \setminus \{\texttt{corr}(y)\}$). Hence,

$$\sum_{y\in\{0,1\}^n} \sum_{x'\in C\setminus\{\texttt{corr}(y)\}} w_{x',y} \cdot \frac{\delta}{2} \;\leq\; \epsilon'. \tag{16}$$

Next, we observe that the total variation distance between X' and $\texttt{corr}(Y)$ is upper-bounded by

$$\sum_{y'\in C} \sum_{x'\in C\setminus\{y'\}} \sum_{y:\texttt{corr}(y)=y'} w_{x',y} = \sum_{y\in\{0,1\}^n} \sum_{x'\in C\setminus\{\texttt{corr}(y)\}} w_{x',y}, \tag{17}$$

since the l.h.s represents the cost of a possible way of transforming X' into $\texttt{corr}(Y)$.[30] Combining Eq. (16) and Eq. (17), it follows that X' is $\frac{2}{\delta}\cdot\epsilon'$-close to $Y' = \texttt{corr}(Y)$ in total variation distance. ■

Recalling that $\texttt{corr}(Y)$ is in $\mathcal{D}$ and using Claim A.5.1, it follows that X' is $\frac{2}{\delta}\cdot\epsilon'$-close to $\mathcal{D}$ in total variation distance. (Recall that it follows that $X = \pi^{-1}(X')$ is $\frac{2}{\delta}\cdot\epsilon'$-close to $\mathcal{D}$ in total variation distance, since $\mathcal{D}$ is label invariant.)[31] The claim follows, since if X is ϵ-far from $\mathcal{D}$ in total variation distance, then X' must be $\delta\epsilon/2$-far from $\mathcal{D}$ according to Definition 1.1. ■

A.4 On standard testing of equality of distributions

Recall that when proving the furthermore clause of Theorem 5.2 we use the fact that the equality tester (for the standard model) of [12] works also when only the support size of one of the tested distributions is upper-bounded.[32] Here, we provide more details about the proof of this fact.

Theorem A.6 (testing equality of two unknown distributions, revised): *Suppose that X and Y are distributed over U and that one of them has support size at most n. Then, distinguishing between the case that $X \equiv Y$ and the case that X is ϵ-far from Y is possible in time $O(\max(n^{2/3}/\epsilon^{4/3}, \sqrt{n}/\epsilon^2))$.*

[30] That is, we map to y' the mass of X' that was mapped to $\{y : \texttt{corr}(y) = y'\}$ (i.e., $\sum_{x'\in C\setminus\{y'\}} \sum_{y:\texttt{corr}(y)=y'} w_{x',y}$), where we pay 1 unit per each $x' \neq y'$. Recall that the total variation distance equals the earth mover's distance with respect to the inequality function (i.e., $\mathrm{InEq}(x',y') = 1$ if $x' \neq y$ and $\mathrm{InEq}(y',y') = 0$).

[31] Indeed, unlike Definition 1.1, the total variation distance (between a pair of distributions) is preserved under relabeling.

[32] This is not a generic claim regarding any such tester. Consider, for example, a modification of any tester such that the modified algorithm always accepts if all samples that are taken from one of the distributions are distinct. The latter event is unlikely when the support size of both distributions is smaller than the square of the number of samples, but is extremely likely if one distribution is uniform over a sufficiently large set (i.e., much larger than the square of the number of samples).

This is a generalization of [15, Thm. 11.24], which originates in [12], where the special case mandates that $U = [n]$ (and in that case the said algorithm is a tester of equality).

Proof: We follow the presentation of [15, Sec. 11.2–11.3], and observe that the support size is only used in the proof of [15, Cor. 11.21], when upper-bounding the total variation distance between two distributions by the norm-2 of their difference, whereas essentially the same upper bound holds also if only the support of one of the distributions is upper-bounded. Details follow.

Our starting point is [15, Alg. 11.17], which is stated as referring to distributions over $[n]$ but can be restated as referring to distributions over U. Recall that the actions of this algorithm only depend on the s samples it obtains from each distribution, whereas s is a free parameter. The same holds with respect to the analysis of this algorithm as an $\mathcal{L}_2$-distance approximator, which is provided in [15, Thm. 11.20].

The key point is that the analysis of [15, Alg. 11.17] as a very crude $\mathcal{L}_1$-distance approximator, provided in [15, Cor. 11.21], remains valid under the relaxed hypothesis (i.e., when only one of the two distributions is guaranteed to have support size at most n). This is because this upper bound (on the support size) is only used when upper-bounding the norm-1 (of the difference between the two distributions) by the norm-2 of the same difference. We observe that we only lose a factor of two when performing the argument on the smaller of the two supports, because at least half of the norm-1 of the difference is due to this smaller support. Specifically, let $p : S \to [0,1]$ be the probability function representing one distribution and $q : U \to [0,1]$ be the function representing the other distribution, where $S \subseteq U$. Then,

$$\begin{aligned}
\sum_{i\in U} |p(i)-q(i)| &= 2 \cdot \sum_{i\in U: p(i)>q(i)} |p(i)-q(i)| \\
&\leq 2 \cdot \sum_{i\in S} |p(i)-q(i)| \\
&\leq 2 \cdot \sqrt{|S|} \cdot \left(\sum_{i\in S} |p(i)-q(i)|^2 \right)^{1/2} \\
&\leq 2 \cdot \sqrt{|S|} \cdot \left(\sum_{i\in U} |p(i)-q(i)|^2 \right)^{1/2}
\end{aligned}$$

where the first inequality is due to $\{i \in U : p(i) > q(i)\} \subseteq \{i \in U : p(i) > 0\} = S$. (Indeed, the first and last inequalities are the place where we go beyond the original proof of [15, Cor. 11.21].) Hence, $\|p-q\|_1 \leq 2\sqrt{|S|} \cdot \|p-q\|_2$, where $|S| \leq n$ by our hypothesis. (In the original proof of [15, Cor. 11.21], which refers to $p, q : [n] \to [0,1]$, one gets $\|p-q\|_1 \leq \sqrt{n} \cdot \|p-q\|_2$, but the difference is immaterial.)

Next, we note that [15, Cor. 11.22(2)] remains valid under the relaxed hypothesis (i.e., when only one of the two distributions is guaranteed to have support size at most n).[33] We stress that this result will only be used when $\beta \geq n^{-1/2}$ (as presumed in the original text).

Lastly, we turn to [15, Alg. 11.24], which is stated as referring to distributions over $[n]$ but can be restated as referring to distributions over U, while making n a free parameter (just as m in the original text). When analyzing this algorithm, we let n denote an upper bound on the size

[33] We mention that [15, Cor. 11.22(1)] also remains valid (even when both distributions have support of unbounded size), but it is not used here.

of the support of one of the two distributions, and apply the revised [15, Cor. 11.22(2)] (which holds in this case). Using $m = \min(n^{2/3}/\epsilon^{4/3}, n)$ (as in the original text), the current claim follows (analogously to establishing [15, Thm. 11.26]). ■