The uniform distribution is complete with respect to testing identity to a fixed distribution

Oded Goldreich*
Department of Computer Science,
Weizmann Institute of Science, Rehovot, Israel.

March 17, 2016

Abstract

Inspired by Diakonikolas and Kane (2016), we reduce the class of problems consisting of testing whether an unknown distribution over $[n]$ equals a fixed distribution to this very problem when the fixed distribution is uniform over $[n]$. Our reduction preserves the parameters of the problem, which are n and the proximity parameter $\epsilon > 0$, up to a constant factor.

While this reduction yields no new bounds on the sample complexity of either problems, it provides a simple way of obtaining testers for equality to arbitrary fixed distributions from testers for the uniform distribution.

Contents

1 Introduction

Inspired by Diakonikolas and Kane [5], we present, for every fixed distribution D over $[n]$, a simple reduction of the problem of testing whether an unknown distribution over $[n]$ equals D to the problem of testing whether an unknown distribution over $[n]$ equals the uniform distribution over

*Partially supported by the Israel Science Foundation (grant No. 671/13).

$[n]$. Specifically, we reduce ϵ-testing of equality to D to $\epsilon/3$-testing of equality to the uniform distribution over $[6n]$, denoted U_{6n}.

Hence, the sample (resp., time) complexity of testing equality to D, with respect to the proximity parameter ϵ, is at most the sample (resp., time) complexity of testing equality to U_{6n} with respect to the proximity parameter $\epsilon/3$. Since optimal bounds were known for both problems (cf., e.g., [9, 2, 1, 11, 4, 13]), our reduction yields no new bounds. Still, it provides a simple way of obtaining testers for equality to arbitrary fixed distributions from testers for the uniform distribution.

The setting at a glance. For any fixed distribution D over $[n]$, we consider the problem of ϵ-testing equality to D, where the tester is given samples drawn from an unknown distribution X and is required to distinguish the case that $X \equiv D$ from the case that X is ϵ-far from D, where the distance is the standard statistical distance. The sample complexity of this testing problem, depends on D, and is viewed as a function of n and ϵ. We write $D \subseteq [n]$ to denote that D ranges over $[n]$.

Wishing to present reductions between such problems, we need to spell out what we mean by such a reduction. Confining ourselves to problems of testing equality to fixed distributions, we say that ϵ-testing equality to $D \subseteq [n]$ reduces to ϵ'-testing equality to $D' \subseteq [n']$ if there exists a randomized process F that maps $[n]$ to $[n']$ such that the distribution D is mapped to the distribution D' and any distribution that is ϵ-far from D is mapped to a distribution that is ϵ'-far from D'. We say that F maps the distribution X to the distribution Y if $Y \equiv F(X)$, where here we view the distributions as random variables. Denoting the uniform distribution over n by U_n, our main result can be stated as follows.

Theorem 1 (completeness of testing equality to U_n): *For every distribution D over $[n]$ and every $\epsilon > 0$, it holds that ϵ-testing equality to D reduces to $\epsilon/3$-testing equality to U_{6n}. Furthermore, the same reduction F can be used for all $\epsilon > 0$.*

Hence, the sample complexity of ϵ-testing equality to D is upper bounded by the sample complexity of $\epsilon/3$-testing equality to U_{6n}. We mention that in some cases, testing equality to D can be easier than testing equality to U_n; such natural cases contain grained distributions (see below). (A general study of the dependence on D of the complexity of testing equality to D was undertaken in [13].)

The reduction at a glance. We decouple the reduction asserted in Theorem 1 into two steps. In the first step, we assume that the distribution D has a probability function q that ranges over multiples of $1/m$, for some parameter $m \in \mathbb{N}$; that is, $m \cdot q(i)$ is a non-negative integer (for every i). We call such a distribution *m-grained*, and reduce testing equality to any fixed m-grained distribution to testing equality to the uniform distribution over $[m]$. This reduction maps i uniformly at random to a set S_i of size $m \cdot q(i)$ such that the S_i's are disjoint. Clearly, this reduction maps the distribution q to the uniform distribution over m fixed elements, and it can be verified that this randomized mapping preserves distances between distributions.

Since every distribution over $[n]$ is $\epsilon/2$-close to a $O(n/\epsilon)$-grained distribution, it is stands to reason that the general case can be reduced to the grained case. This is indeed true, but the reduction is less obvious than the treatment of the grained case. Actually, we shall use a different "graining" procedure, which yields a better result (i.e., the result stated above). Specifically, we present a reduction of ϵ-testing equality to D to $\epsilon/3$-testing equality to some $6n$-grained distribution D', where D' depends only on D. Letting $q : [n] \to [0, 1]$ denote the probability function of D, the

reduction maps $i \in [n]$ to itself with probability $\frac{\lfloor 6n \cdot q(i) \rfloor / 6n}{q(i)}$, and otherwise maps i to $n+1$. This description works when $q(i) \geq 1/2n$ for every $i \in [n]$, and in order to guranteed this condition we use a preliminary reduction that maps $i \in [n]$ to itself with probability $1/2$ and otherwise maps it uniformlty to $[n]$. It is quite obvious that the preliminary reduction cuts the distance between distributions by a factor of two, and it can be shown that the main randomized mapping preserves distances between distributions up to a constant factor (of 2/3).

History, credits, and an acknowledgement. The study of testing properties of distributions was initiated by Batu, Fortnow, Rubinfeld, Smith and White [2]. Testers of sample complexity $\mathrm{poly}(1/\epsilon) \cdot \sqrt{n}$ for equality to U_n and for equality to an arbitrary distribution D over $[n]$ were presented by Goldreich and Ron [9] and Batu *et al.* [1], respectively, were the presentation in [9] is only implicit.[1] The tight lower and upper bound of $\Theta(\sqrt{n}/\epsilon^2)$ on the sample complexity of both problems were presented in [11, 4, 13] (see also [6, 5]). For a general survey of the areas, the interested reader is referred to Canonne [3].

As stated upfront, our reductions are inspired by Diakonikolas and Kane [5], who presented a unified approach for deriving optimal testers for various properties of distributions (and pairs of distributions) via reductions to testing the equality of two unknown distributions that have small $\mathcal{L}_2$-norm. We note that our reduction from testing equality to grained distributions to testing equality to the uniform distribution is implicit in [6].

Lastly, we wish to thank Ilias Diakonikolas for numerous email discussions, which were extremely helpful in many ways.

2 Preliminaries

We consider *discrete* probability distributions. Such distribution have a finite *support*, which we assume to be a subset of $[n]$ for some $n \in \mathbb{N}$, where the support of a distribution is the set of elements assigned positive probability mass. We represent such distributions either by random variables, like X, that are assigned values in $[n]$ (indicated by writing $X \in [n]$), or by probability functions like $p : [n] \to [0,1]$ that satisfy $\sum_{i \in [n]} p(i) = 1$. These two representation correspond via $p(i) = \mathbf{Pr}[X = i]$. At times, we also refer to distributions as such, and denote them by D. (Distributions over other finite sets can be treated analogously, but in such a case we should provide the tester with a description of the set; indeed, n serves as a concise description of $[n]$.)

Recall that the study of "distribution testing" refers to testing properties of distributions. That is, the object being testing is a distribution, and the property it is tested for is a property of distributions (equiv., a set of distributions). The tester itself is given samples from the distribution and is required to distinguish the case that the distribution has the property from the case that the distribution is far from having the property, where the distance between distributions is defined as as the total variation distance between them (a.k.a the statistical difference). That is, X and Y are said to be ϵ-close if

$$\frac{1}{2} \cdot \sum_i |\mathbf{Pr}[X=i] - \mathbf{Pr}[Y=i]| \leq \epsilon, \tag{1}$$

and otherwise they are deemed ϵ-far. With this definition in place, we are ready to recall the standard definition of testing distributions.

[1] Testing equality to U_n is implicit in a test of the distribution of the endpoint of a realtively short random walk on a bounded-degree graph.

Definition 2 (testing properties of distributions): *Let $\mathcal{D} = \{\mathcal{D}_n\}_{n\in\mathbb{N}}$ be a property of distributions 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 a sequence of $s(n)$ samples drawn from an unknown distribution $X \in [n]$, satisfies the following two conditions.*

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

$$\mathbf{Pr}_{i_1,...,i_s\sim X}[T(n,\epsilon;i_1,...,i_s)=1] \geq 2/3,$$

   where $s = s(n,\epsilon)$ and $i_1,...,i_s$ are drawn independently from the distribution X.

2. The tester rejects distributions that far from $\mathcal{D}$: *If X is ϵ-far from any distribution in $\mathcal{D}_n$* (i.e., X is ϵ-far from $\mathcal{D}$)*, then*

$$\mathbf{Pr}_{i_1,...,i_s\sim X}[T(n,\epsilon;i_1,...,i_s)=0] \geq 2/3,$$

   where $s = s(n,\epsilon)$ and $i_1,...,i_s$ are as in the previous item.

Our focus is on "singleton" properties; that is, the property is $\{D_n\}_{n\in\mathbb{N}}$, where D_n is a fixed distribution over $[n]$. Note that n fully specifies the distribution D_n, and we do not consider the complexity of obtaining an explicit description of D_n from n. For sake of simplicity, we will consider a generic n and omit it from the notation (i.e., use D rarher than D_n). Furthermore, we refer to ϵ-testers derived by setting the proximity parameter to ϵ. Nevertheless, all testers discussed here are actually uniform with respect to the proximity parameter ϵ (and also with respect to n, assuming they already derived or obtained an explicit description of D_n).

Confining ourselves to problems of testing equality to distributions, we formally restate the notion of a reduction used in the introduction. In fact, we explicitly refer to the randomized mapping at the heart of the reduction, and also define a stronger (i.e., uniform over ϵ) notion of a reduction that captures the furthermore part of Theorem 1.

Definition 3 (reductions via filters): *We say that a randomized process F, called a* filter*,* reduces ϵ-testing equality to $D \subseteq [n]$ to ϵ'-testing equality to $D' \subseteq [n']$ *if the distribution D is mapped to the distribution D' and any distribution that is ϵ-far from D is mapped to a distribution that is ϵ'-far from D'. We say that F* reduces testing equality to $D \subseteq [n]$ to testing equality to $D' \subseteq [n']$ *if, for some constant c and every $\epsilon > 0$, it holds that F reduces ϵ-testing equality to D to ϵ/c-testing equality to D'.*

Recall that we say that F maps the distribution X to the distribution Y if Y and $F(X)$ are identically distributed (i.e., $Y \equiv F(X)$), where we view the distributions as random variables. We stress that if F_q is invoked t times on the same i, then the t outcomes are (identically and) independently distributed. Hence, a sequence of samples drawn independently from a distribution X is mapped to a sequence of samples drawn independently from the distribution $F(X)$.

3 The reduction

The following description was reproduced (and slightly adapted) from lecture notes for an introductory cource on property testing, which are currently in preparation [8]. Hence, the style is

somewhat different from the one in the introduction. In particular, the notion of a reduction is not used explicitly (except in Propositions 6 and 9 and in Section 3.3).[2] Instead, the randomized mapping is called a filter, and reductions are presented in the form of testers that use filters in order to transform samples given to them into samples for sub-testers that they invoke. Typically, the sub-tester is invoked once, but at one occasion we allow the tester not to invoke the sub-tester at all (but rather reject based on a simple test). This exception can be avoided, as discussed in Section 3.3.

Recall that testing equality to a fixed distribution D means testing the property $\{D\}$; that is, testing whether an unknown distribution equals the fixed distribution D. For any distribution D over $[n]$, we present a reduction of the task of ϵ-testing $\{D\}$ to the task of $\epsilon/3$-testing the uniform distribution over $[O(n)]$.

We decouple the reduction into two steps. In the first step, we assume that the distribution D has a probability function q that ranges over multiples of $1/m$, for some parameter $m \in \mathbb{N}$; that is, $m \cdot q(i)$ is a non-negative integer (for every i). We call such a distribution *m-grained*, and reduce testing equality to any fixed m-grained distribution to testing uniformity (over $[m]$). Since every distribution over $[n]$ is $\epsilon/2$-close to a $O(n/\epsilon)$-grained distribution, it stands to reason that the general case can be reduced to the grained case. This is indeed true, but the reduction is less obvious than the treatment of the grained case. (Actually, we shall use a different "graining" procedure, which yields a better result.)

Definition 4 (grained distributions): *We say that a probability distribution over $[n]$ having a probability function $q : [n] \to [0,1]$ is m-*grained *if q ranges over multiples of $1/m$; that is, if for every $i \in [n]$ there exists a non-negative integer m_i such that $q(i) = m_i/m$.*

Note that the uniform distribution over $[n]$ is n-grained, and that an m-grained distribution must have support size at most m. In particular, if a distribution D results from applying some function to the uniform distribution over $[m]$, then D is m-grained.

3.1 Testing equality to a fixed grained distribution

Fixing any m-grained distribution (represented by a probbaility function) $q : [n] \to \{j/m : j \in \mathbb{N} \cup \{0\}\}$, we consider a randomized transformation (or "filter"), denoted F_q, that maps the support of q to $S = \{\langle i,j\rangle : i \in [n] \wedge j \in [m_i]\}$, where $m_i = m \cdot q(i)$. (We stress that, as with any randomized process considered so far (e.g., any type of randomized algorithm including any tester), invoking the filter several times on the same input yields independently and identically distributed outcomes.) Specifically, for every i in the support of q, we map i uniformly to $S_i = \{\langle i,j\rangle : j \in [m_i]\}$; that is, $F_q(i)$ is uniformly distributed over S_i. If i is outside the support of q (i.e., $q(i) = 0$), then we map it to $\langle i, 0\rangle$. Note that $|S| = \sum_{i\in[n]} m_i = \sum_{i\in[n]} m \cdot q(i) = m$. The key observations about this filter are:

1. *The filter F_q maps q to a uniform distribution*: If Y is distributed according to q, then $F_q(Y)$ is distributed uniformly over S; that is, for every $\langle i,j\rangle \in S$, it holds that

$$\begin{aligned} \mathbf{Pr}[F_q(Y) = \langle i,j\rangle] &= \mathbf{Pr}[Y = i] \cdot \mathbf{Pr}[F_q(i) = \langle i,j\rangle] \\ &= q(i) \cdot \frac{1}{m_i} \end{aligned}$$

[2] Propositions 6 and 9 were added in the current write-up, and Section 3.3 was significantly revised.

$$= \frac{m_i}{m} \cdot \frac{1}{m_i}$$

which equals $1/m = 1/|S|$.

2. *The filter preserves the variation distance between distributions*: The total variation distance between $F_q(X)$ and $F_q(X')$ equals the total variation distance between X and X'. This holds since, for $S' = S \cup \{\langle i, 0\rangle : i \in [n]\}$, we have

$$\begin{aligned}
&\sum_{\langle i,j\rangle \in S'} \left|\mathbf{Pr}[F_q(X) = \langle i,j\rangle] - \mathbf{Pr}[F_q(X') = \langle i,j\rangle]\right| \\
&= \sum_{\langle i,j\rangle \in S'} \left|\mathbf{Pr}[X = i] \cdot \mathbf{Pr}[F_q(i) = \langle i,j\rangle] - \mathbf{Pr}[X' = i] \cdot \mathbf{Pr}[F_q(i) = \langle i,j\rangle]\right| \\
&= \sum_{\langle i,j\rangle \in S'} \mathbf{Pr}[F_q(i) = \langle i,j\rangle] \cdot \left|\mathbf{Pr}[X = i] - \mathbf{Pr}[X' = i]\right| \\
&= \sum_{i \in [n]} \left|\mathbf{Pr}[X = i] - \mathbf{Pr}[X' = i]\right| .
\end{aligned}$$

   Indeed, this is a generic statement that applies to any filter that maps i to a pair $\langle i, Z_i\rangle$, where Z_i is an arbitrary distibution that only depends on i. (Equivalently, the statement holds for any filter that maps i to a random variable Z_i that only depends on i such that the supports of the different Z_i's are disjoint.)

Noting that a knowledge of q allows to implement F_q as well as to map S to $[m]$, yields the following reduction.

Algorithm 5 (reducing testing equality to m-grained distributions to testing uniformity over $[m]$): *Let D be an m-grained distribution with probability function $q : [n] \to \{j/m : j \in \mathbb{N} \cup \{0\}\}$. On input $(n, \epsilon; i_1, ..., i_s)$, where $i_1, ..., i_s \in [n]$ are samples drawn according to an unknow distribution p, invoke an ϵ-tester for uniformity over $[m]$ by providing it with the input $(m, \epsilon; i'_1, ..., i'_s)$ such that for every $k \in [s]$ the sample i'_k is generated as follows:*

1. *Generate $\langle i_k, j_k\rangle \leftarrow F_q(i_k)$.*

   Recall that if $m_{i_k} \stackrel{\text{def}}{=} m \cdot q(i_k) > 0$, then j_k is selected uniformly in $[m_k]$, and otherwise $j_k \leftarrow 0$. We stress that if F_q is invoked t times on the same i, then the t outcomes are (identically and) *independently distributed. Hence, the s samples drawn independently from p are mapped to s samples drawn independently from p' such that $p'(\langle i, j\rangle) = p(i)/m_i$ if $j \in [m_i]$ and $p'(\langle i, 0\rangle) = p(i)$ if $m_i = 0$.*

2. *If $j_k \in [m_{i_k}]$, then $\langle i_k, j_k\rangle \in S$ is mapped to its rank in S* (according to a fixed order of S)*, where $S = \{\langle i, j\rangle : i \in [n] \wedge j \in [m_i]\}$, and otherwise $\langle i_k, j_k\rangle \notin S$ is mapped to $m + 1$.*

(Altertaively, the reduction may just reject if any of the j_k's equals 0.)[3]

The forgoing description presumes that the tester for uniform distributions over $[m]$ also operates well when given arbitrary distributions (which may have a support that is not a subset of $[m]$). However, any tester for uniformity can be easily extended to do so (see discussion in Section 3.3). In any case, we get

[3] See farther discussion in Section 3.3.

Proposition 6 (Algorithm 5 as a reduction): *The filter F_q used in Algorithm 5 reduces ϵ-testing equality to an m-grained distribution D* (over $[n]$) *to ϵ-testing equality to the uniform distribution over $[m]$, where the distributions tested in the latter case are over $[m+1]$. Furthermore, if the support of q equals $[n]$, which may happen only if $m \geq n$, then the reduction is to testing whether a distribution over $[m]$ is uniform on $[m]$.*

Using any of the known uniformity tests that have sample complexity $O(\sqrt{n}/\epsilon^2)$,[4] we obtain –

Corollary 7 (testing equality to m-grained distributions): *For any fixed m-grained distribution D, the property $\{D\}$ can be ϵ-tested in sample complexity $O(\sqrt{m}/\epsilon^2)$.*

We note that the foregoing *tester for equality to grained distributions* is of independent interest, which extends beyond its usage towards testing equality to arbitrary distributions.

3.2 From arbitrary distributions to grained ones

We now turn to the problem of testing equality to an arbitrary known distribution, represented by $q : [n] \to [0,1]$. The basic idea is to round all probabilities to multiples of γ/n, for an error parameter γ (which will be a small constant). Of course, this rounding should be performed so that the sum of probabilities equals 1. For example, we may use a randomized filter that, on input i, outputs i with probability $\frac{m_i \cdot \gamma/n}{q(i)}$, where $m_i = \lfloor q(i) \cdot n/\gamma \rfloor$, and outputs $n+1$ otherwise. Hence, if i is distributed according to p, then the output of this filter will be i with probability $\frac{\gamma m_i/n}{q(i)} \cdot p(i)$. This works well if $\gamma m_i/n \approx q(i)$, which is the case if $q(i) \gg \gamma/n$ (equiv., $m_i \gg 1$), but may run into trouble otherwise.

For starters, we note that if $q(i) = 0$, then we should take $\frac{\gamma m_i/n}{q(i)} = 1$, beacuse otherwise we may not distinguish between distributions that are identical when conditioned on i's such that $q(i) > 0$ (but differ significantly on i's on which $q(i) = 0$).[5] Similar effects occur when $q(i) \in (0, \gamma/n)$: In this case $m_i = 0$ and so the proposed filter ignores the probability assigned by the distribution p on this i. Hence, we modify the basic idea such to avoid this problem.

Specifically, we first use a filter that averages the input distribution p with the uniform distribution, and so guarantees that all elements occur with probability at least $1/2n$, while preseving distances between different input distributions (up to a factor of two). Only then, do we apply the foregoing proposed filter (which outputs i with probability $\frac{m_i \cdot \gamma/n}{q(i)}$, where $m_i = \lfloor q(i) \cdot n/\gamma \rfloor$, and outputs $n+1$ otherwise). Details follow.

1. We first use a filter F' that, on input $i \in [n]$, outputs i with probability $1/2$, and outputs the uniform distribution (on $[n]$) otherwise. Hence, if i is distributed according to the distribution p, then $F'(i)$ is distributed according to $p' = F'(p)$ such that

$$p'(i) = \frac{1}{2} \cdot p(i) + \frac{1}{2} \cdot \frac{1}{n}. \tag{2}$$

[4] Recall that the alternatives include the tests of [11] and [4] or the collision probability test (of [9]), per its improved analysis in [7].

[5] Consider for example the case that $q(i) = 2/n$ on every $i \in [n/2]$ and a distribution X that is uniform on $[n]$. Then, $\mathbf{Pr}[X = i | q(X) > 0] = q(i)$ for every $i \in [n/2]$, but $\mathbf{Pr}[X = i | q(X) = 0] = 2/n$ for every $i \in [(n/2)+1, n]$. Hence, X and the uniform distribution on $[n/2]$ are very different, but are identical when conditioned on i's such that $q(i) > 0$.

(Indeed, we denote by $F'(p)$ the probability function of the distribution obtained by selecting i according to the probability function p and outputing $F'(i)$.)

Let $q' = F'(q)$; that is, $q'(i) = 0.5 \cdot q(i) + (1/2n) \geq 1/2n$.

2. Next, we apply a filter $F''_{q'}$, which is related to the filter F_q used in Algorithm 5. Letting $m_i = \lfloor q'(i) \cdot n/\gamma \rfloor$, on input $i \in [n]$, the filter outputs i with probability $\frac{m_i \cdot \gamma/n}{q'(i)}$, and outputs $n+1$ otherwise (i.e., with probability $1 - \frac{m_i \gamma/n}{q'(i)}$).

   Note that $\frac{m_i \gamma/n}{q'(i)} \leq 1$, since $m_i \leq q'(i) \cdot n/\gamma$. On the other hand, recalling that $q'(i) \geq 1/2n$ and observing that $m_i \cdot \gamma/n > ((q'(i) \cdot n/\gamma) - 1) \cdot \gamma/n = q'(n) - (\gamma/n)$, it follows that $\frac{m_i \gamma/n}{q'(i)} > 1 - 2\gamma$.

   Now, if i is distributed according to the distribution p', then $F''_{q'}(i)$ is distributed according to $p'' : [n+1] \to [0,1]$ such that, for every $i \in [n]$, it holds that

$$p''(i) = p'(i) \cdot \frac{m_i \cdot \gamma/n}{q'(i)} \tag{3}$$

   and $p''(n+1) = 1 - \sum_{i \in [n]} p''(i)$.

   Let q'' denote the probability function related to q'. Then, for every $i \in [n]$, it holds that $q''(i) = q'(i) \cdot \frac{m_i \gamma/n}{q'(i)} = m_i \cdot \gamma/n \in \{j \cdot \gamma/n : j \in \mathbb{N} \cup \{0\}\}$ and $q''(n+1) = 1 - \sum_{i \in [n]} m_i \cdot \gamma/n < \gamma$, since $m \stackrel{\text{def}}{=} \sum_{i \in [n]} m_i > \sum_{i \in [n]} ((n/\gamma) \cdot q'(i) - 1) = (n/\gamma) - n$. Note that if n/γ is an integer, then q'' is n/γ-grained, since in this case $q''(n+1) = 1 - m \cdot \gamma/n = (n/\gamma - m) \cdot \gamma/n$. Furthermore, if $m = n/\gamma$, which happens if and only if $q'(i) = m_i \cdot \gamma/n$ for every $i \in [n]$, then q'' has support $[n]$, and otherwise it has support $[n+1]$.

Combining these two filters, we obtain the desired reduction.

Algorithm 8 (reducing testing equality to a general distribution to testing equality to a $O(n)$-grained distributions): *Let D be an arbitrary distribution with probability function $q : [n] \to [0,1]$, and T be an ϵ'-tester for m-grained distributions having sample complexity $s(m, \epsilon')$. On input $(n, \epsilon; i_1, ..., i_s)$, where $i_1, ..., i_s \in [n]$ are $s = s(O(n), \epsilon/3)$ samples drawn according to an unknow distribution p, the tester proceeds as follows:*

1. *It produces a s-long sequence $(i''_1, ..., i''_s)$ by applying $F''_{F'(q)} \circ F'$ to $(i_1, ..., i_s)$, where F' and $F''_{q'}$ are as in* Eq. (2)&(3)*; that is, for every $k \in [s]$, it produces $i'_k \leftarrow F'(i_k)$ and $i''_k \leftarrow F''_{F'(q)}(i'_k)$.*

   (Recall that $F''_{q'}$ depends on a universal constant γ, which we shall set to 1/6.)

2. *It invokes the $\epsilon/3$-tester T for q'' providing it with the sequence $(i''_1, ..., i''_s)$. Note that this is a sequence over $[n+1]$.*

Using the notations as in Eq. (2)&(3), we first observe that the total variation distance between $p' = F'(p)$ and $q' = F'(q)$ is half the total variation distance between p and q (since $p'(i) = 0.5 \cdot p(i) + (1/2n)$ and ditto for q'). Next, we observe that the total variation distance between $p'' = F''_{q'}(p')$ and $q'' = F''_{q'}(p')$ is lower bounded by a constant fraction of the total variation distance

between p' and q'. To see this, let X and Y be distributed according to p' and q', respectively, and observe that

$$\begin{aligned}\sum_{i\in[n]}\left|\mathbf{Pr}[F_{q'}(X)=i]-\mathbf{Pr}[F_{q'}(Y)=i]\right| &= \sum_{i\in[n]}\left|p'(i)\cdot\frac{m_i\gamma/n}{q'(i)}-q'(i)\cdot\frac{m_i\gamma/n}{q'(i)}\right| \\ &= \sum_{i\in[n]}\frac{m_i\gamma/n}{q'(i)}\cdot\left|p'(i)-q'(i)\right| \\ &\geq \min_{i\in[n]}\left\{\frac{m_i\gamma/n}{q'(i)}\right\}\cdot\sum_{i\in[n]}\cdot\left|p'(i)-q'(i)\right|.\end{aligned}$$

As stated above, recalling that $q'(i)\geq 1/2n$ and $m_i=\lfloor(n/\gamma)\cdot q'(i)\rfloor>(n/\gamma)\cdot q'(i)-1$, it follows that

$$\frac{m_i\gamma/n}{q'(i)}>\frac{((n/\gamma)\cdot q'(i)-1)\cdot\gamma/n}{q'(i)}=1-\frac{\gamma/n}{q'(i)}\geq 1-\frac{\gamma/n}{1/2n}=1-2\gamma.$$

Hence, if p is ϵ-far from q, then p' is $\epsilon/2$-far from q', and p'' is $\epsilon/3$-far from q'', where we use $\gamma\leq 1/6$. On the other hand, if $p=q$, then $p''=q''$. Noting that q'' is an n/γ-grained distribution, provided that n/γ is an integer (as is the case for $\gamma=1/6$), we complete the analysis of the reduction. Hence,

Proposition 9 (Algorithm 8 as a reduction): *The filter $F''_{F'(q)}\circ F'$ used in Algorithm 8 reduces ϵ-testing equality to any fixed distribution D* (over $[n]$) *to ϵ-testing equality to an $6n$-grained distribution over $[n']$, where $n'\in\{n,n+1\}$ depends on q.*[6] *Furthermore, the support of $F''_{F'(q)}\circ F'(q)$ equals $[n']$.*

Hence, *the sample complexity of ϵ-testing equality to arbitrary distributions over $[n]$ equals the sample complexity of $\epsilon/3$-testing equality to $O(n)$-grained distributions* (which is essentially a special case).

Digest. One difference between the filter underlying Algorithm 5 and the one underlying Algorithm 8 is that the former preserves the exact distance between distributions, whereas the later only preseves them up to a constant factor. The difference is reflected in the fact that the first filter maps the different i's to distributions of disjoint support, whereas the second filter (which is composed of the filters of Eq. (2)&(3)) maps different i's to distributions of non-disjoint support. (Specifically, the filter of Eq. (2) maps every $i\in[n]$ to a distribution that assigns each $i'\in[n]$ probability at least $1/2n$, whereas the filter of Eq. (3) typically maps each $i\in[n]$ to a distribution with a support that contains the element $n+1$.)

3.3 From arbitrary distributions to the uniform one

Combining the reductions stated in Propositions 6 and 9, we obtain a proof of Theorem 1.

Theorem 10 (Theorem 1, restated) *For every probability function $q:[n]\to[0,1]$ the filter $F_{q''}\circ F''_{F'(q)}\circ F'$, where $q''=F''_{F'(q)}\circ F'(q)$ is as in Algorithm 8 and $F_{q''}$ is as in Algorithm 5, reduces ϵ-testing equality to q to $\epsilon/3$-testing equality to the uniform distribution over $[6n]$.*

[6] Typically, $n'=n+1$. Recall that $n'=n$ if and only if D itself is $6n$-grained, in which case the reduction is not needed anyhow.

Proof: First, setting $\gamma = 1/6$ and using the filter $F''_{F'(q)} \circ F'$, we reduce the problem of ϵ-testing equality to q to the problem of $\epsilon/3$-testing equality to the $6n$-grained distribution q'', while noting that the distribution q'' has support $[n']$, where $n' \in \{n, n+1\}$ (depending on q). Note that the latter assertion relies on the furthermore part of Proposition 9. Next, using the furthermore part of Proposition 6, we note that $F_{q''}$ reduces $\epsilon/3$-testing equality to q'' to $\epsilon/3$-testing equality to the uniform distribution over $[6n]$. ■

Observe that the proof of Theorem 10 avoids the problem discussed right after the presentation of Algorithm 5, which refers to the fact that testing equality to an m-grained distribution $q : [n] \to [0,1]$ is reduced to testing whether distributions over $[n']$ are uniform over $[m]$, where in some cases $n' \in [n, n+m]$ rather than $n' = m$. These bad cases arise when the support of q is a strict subset of $[n]$, and it was avoided since we applied the filter of Algorithm 5 to distributions $q'' : [n'] \to [0,1]$ that have support $[n']$. Nevertheless, it is nice to have a reduction from the general case of "testing uniformity" to the special case, where the general case refers to testing whether distributions over $[n]$ are uniform over $[m]$, for any n and m, and the special case mandates that $m = n$. Such a reduction is provided next.

Theorem 11 (testing uniform distributions, a reduction between two versions): *There exists a simple filter that maps U_m to U_{2m}, while mapping any distribution X that is ϵ-far from U_m to a distribution over $[2m]$ that is $\epsilon/2$-far from U_{2m}. We stress that X is not necessarily distributed over $[m]$ and remind the reader that U_n denotes the uniform distribution over $[n]$.*

Thus, this filter reduces ϵ-testing whether distributions over $[n]$ are uniform over $[m]$ to $\epsilon/2$-testing whether distributions over $[2m]$ are uniform over $[2m]$.

Proof: The filter, denoted F, maps $i \in [m]$ uniformly at random to an element in $\{i, m+i\}$, while mapping any $i \notin [m]$ uniformly at random to an element in $[m]$. Observe that any distribution over $[n]$ is mapped to a distribution over $[2m]$ and that $F(U_m) \equiv U_{2m}$. Note that F does not necessarily preserve distances between arbitrary distributions over $[n]$ (e.g., both the uniform distribution over $[2m]$ and the uniform distribution over $[m] \cup [2m+1, 3m]$ are mapped to the same distribution), but (as shown next) F preserves distances to the relevant uniform distributions up to a constant factor. Specifically, note that

$$\sum_{i \in [m+1,2m]} |\mathbf{Pr}[F(X)=i] - \mathbf{Pr}[U_{2m}=i]| = \frac{1}{2} \cdot \sum_{i \in [m]} |\mathbf{Pr}[X=i] - \mathbf{Pr}[U_m=i]|$$

and

$$\begin{aligned}
\sum_{i \in [m]} |\mathbf{Pr}[F(X)=i] - \mathbf{Pr}[U_{2m}=i]| &\geq \mathbf{Pr}\left[F(X) \in [m]\right] - \mathbf{Pr}\left[U_{2m} \in [m]\right] \\
&= \left(\frac{1}{2} \cdot \mathbf{Pr}[X \in [m]] + \mathbf{Pr}[X \notin [m]]\right) - \frac{1}{2} \\
&= \frac{1}{2} \cdot \sum_{i \notin [m]} |\mathbf{Pr}[X=i] - \mathbf{Pr}[U_m=i]| \,.
\end{aligned}$$

Hence, the total variation distance between $F(X)$ and U_{2m} is at least half the total variation distance between X and U_m. ■

Playing with the parameters. The filter of Eq. (2) can be generalized by introducing a parameter $\beta \in (0,1)$ and letting $p'(i) = (1-\beta) \cdot p(i) + \beta/n$. Picking $\gamma \in (0,\beta)$ such that n/γ is an integer, we get a trade-off between the loss in the proximity parameter ϵ and the blow-up in the size parameter n. Specifically, this reduces ϵ-testing equality to q to ϵ'-testing equality to the uniform distribution over $[n/\gamma]$, where $\epsilon' = (1-\beta) \cdot (1-(\gamma/\beta)) \cdot \epsilon$. Recalling that the complexity of the latter problem is proportional to $\sqrt{n/\gamma}/(\epsilon')^2$, it seems that setting $\beta = 1/2$ and $\gamma = 1/6$ is quite good (alas not optimal).

4 On the complexity of testing whether a distribution is grained

A natural question that arises from the interest in grained distributions refers to the complexity of testing whether an unknown distribution is grained. Specifically, given n and m (and a proximity parameter ϵ), how many samples are required in order to determine whether an unknown distribution X over $[n]$ is m-grained or ϵ-far from any m-grained distribution. This question can be partially answered by invoking the results of Valiant and Valiant [12]. Specifically, for an upper bound we use their "learning up to relabelling" algorithm, which may be viewed as a learner of histograms (which is what it actually does). Recall that the histogram of the probability function p is defined as the multiset $\{p(i) : i \in [n]\}$ (equiv., as the set of pairs $\{(v,m) : m = |\{i \in [n] : p(i) = v\}| > 0\}$).

Theorem 12 (learning the histogram [12, Thm. 1]):[7] *There exists an $O(\epsilon^{-2} \cdot n/\log n)$ time algorithm that, on input n, ϵ and $O(\epsilon^{-2} \cdot n/\log n)$ samples drawn from an unknown distribution $p : [n] \to [0,1]$, outputs, with probability $1 - 1/\mathrm{poly}(n)$, a histogram of a distribution that is ϵ-close to p.*

The implication of this result on testing any label-invariant property of distributions is immediate. In our case, the tester consists of employing the algorithm of Theorem 12 with proximity parameter $\epsilon/2$ and accepting if and only if the output fits a histogram of a distribution that is $\epsilon/2$-close to being m-grained. The same holds with respect to estimating the distance from the set of m-grained distributions (which can be captured as a special case of label-invariant properties). Hence, we get

Corollary 13 (testing whether a distribution is grained): *For every $n, m \in \mathbb{N}$, the set of m-grained distributions over $[n]$ has a tester of sample complexity $O(\epsilon^{-2} \cdot n/\log n)$. Furthermore, the distance of an unkown distribution to the set of m-grained distributions over $[n]$ can be approximated up to an additive error of ϵ using the same number of samples.*

We comment that it seems that using the techniques of [12] one can reduce the complexity to $O(\epsilon^{-2} \cdot n'/\log n')$, where $n' = \min(n,m)$. (For the case of testing, this is shown in the appendix, using a reduction.) On the other hand, for $m \in [\Omega(n), O(n)]$, the above distance approximator is optimal, whereas it makes no sense to consider $m > n/\epsilon$ (since any distribution over $[n]$ is ϵ-close to being n/ϵ-grained). The negative result follows from the corresponding result of Valiant and Valiant [12].

[7] Valiant and Valiant [12] stated this result for the "relative earthmover distance" (REMD) and commented that the total variation distance up to relabelling is upper bounded by REMD. This claim appears as a special case of [14, Fact 1] (using $\tau = 0$), and a detailed proof appears in [10].

Theorem 14 (optimality of Theorem 12, [12, Thm. 2]):[8] *For every sufficiently small $\epsilon > 0$, there exist two distributions $p_1, p_2 : [n] \to [0,1]$ that are indistinguishable by $O(\epsilon^{-1} n/\log n)$ samples although p_1 is ϵ-close to the uniform distribution over $[n]$ and p_2 is ϵ-close to the uniform distribution over some set of $n/2$ elements.*

Corollary 15 (optimality of Corollary 13): *For any $m \in [\Omega(n), O(n)]$, estimating the distance to the set of m-grained distributions over $[n]$ up to a sufficiently small additive constant requires $\Omega(n/\log n)$ samples.*

Similarly, tolerant testing in the sense of distinguishing distributions that are ϵ_1-close to being m-grained from distributions that are ϵ_2-far from being m-grained requires $\Omega(n/\log n)$ samples, for any constant $\epsilon_2 \in (0, 1/(2 \cdot \lfloor 2m/n \rfloor))$ and $\epsilon_1 \in (0, \epsilon_2)$.

Proof Sketch: The case of $m = n/2$ follows by invoking Theorem 14 (with $\epsilon = \delta$), while observing that the uniform distribution over $[n/2]$ is m-grained whereas the uniform distribution over $[n]$ is $(0.499 - \delta)$-far from the the set of distributions that are δ-close to beging m-grained.[9] Hence, distinguishing the distributions p_2 and p_1 (of Theorem 14) is reducible to $(0.499 - 2\delta)$-testing the set of distributions that are δ-close to be m-grained, which implies that the latter task has sample complexity $\Omega(n/\log n)$. For $m < n/2$, we invoke Theorem 14 while resetting n to $2m$, which means that we consider distributions over $[n]$ with a support that is a subset of $[2m]$. (So the lower bound is $\Omega(m/\log m) = \Omega(n/\log n)$, where the inequality uses $m = \Omega(n)$.)

For $m > n/2$, we show a reduction of the distinguishing task underlying Theorem 14 to the testing problem at hand. Specifically, let $t = \lceil 2m/n \rceil$, and assume that t divides m (otherwise use $\lfloor m/t \rfloor$ instead of m/t, and reduce this special case to the general case).[10] Consider a randomized filter, denoted $F_{m,t}$, that with probability $1/t$ maps $i \in [m/t]$ to $(m/t) + i$, otherwise maps i to itself, and always maps $i \notin [m/t]$ to $i - (m/t)$. Then, $F_{m,t}$ maps the uniform distribution over $[m/t]$ to a distribution q_2 such that $q_2(i) = (t-1)/m$ if $i \in [m/t]$ and $q_2(i) = 1/m$ if $i \in [(m/t)+1, 2m/t]$, which is m-grained. On the other hand, $F_{m,t}$ maps the uniform distribution over $[2m/t]$ to a distribution q_1 such that $q_1(i) = (2t-1)/2m$ if $i \in [m/t]$ and $q_1(i) = 1/2m$ if $i \in [(m/t)+1, 2m/t]$, which is $0.999/2t$-far from being m-grained. Applying the same filter to the distributions p_1 and p_2 of Theorem 14 (while setting $n = 2m/t$ and $\epsilon = \delta$), we obtain distributions p_2' and p_1' such that p_2' is δ-close to being m-grained whereas p_1' is $((0.999/2t) - \delta)$-far from being m-grained, since filters can only decarese the distance between distributions. Hence, distinguishing the distributions p_2 and p_1 (over $[2m/t]$) is reducible to $(0.999/2t - 2\delta)$-testing the set of distributions that are δ-close to being m-grained, which implies that the latter task has sample complexity $\Omega((2m/t)/\log(2m/t))$. (The claim follows by recalling that $1/t = \Omega(1)$, since $m = O(n)$.) ■

Open Problems. Note that Corollary 15 does not refer to testing, but rather to distance approximation, and there are natural cases in which the complexity of testing a property of distributions is

[8] Like in Footnote 7, we note that Valiant and Valiant [12] stated this result for the "relative earthmover distance" (REMD) and commented that the total variation distance up to relabelling is upper bounded by REMD. This claim appears as a special case of [14, Fact 1] (using $\tau = 0$), and a detailed proof appears in [10].

[9] The constant 0.499 stands for an arbitrary large constant that is smaller than 0.5. Recall that the definition of δ-far mandates that the relevant distance be *greater* than δ.

[10] For example, the reduction may use a filter that maps $i \in [n]$ with itself with probability $t \cdot \lfloor m/t \rfloor / m$ and maps it to n otherwise.

significantly lower than the corresponding distance approximation task (cf. [9] versus [12]). Hence, we ask –

Open Problem 16 (the sample complexity of testing whether a distribution is grained): *For any m and n, what is the sample complexity of testing the property that consists of all m-grained distributions over $[n]$.*

This question can be generalized to properties that allow m to reside in some predetermined set M, where the most natural case is that M is an interval, say of the form $[m', 2m']$.

Open Problem 17 (Problem 16, generalized): *For any finite set $M \subset \mathbb{N}$ and $n \in \mathbb{N}$, what is the sample complexity of testing the property that consists of all distributions over $[n]$ that are each m-grained for some $m \in M$.*

Appendix

Recall that Corollary 13 asserts that *for every $n, m \in \mathbb{N}$, the set of m-grained distributions over $[n]$ has a tester of sample complexity $O(\epsilon^{-2} \cdot n/\log n)$*. As commented in the main text, we believe that using the techniques of [12] one can reduce the complexity to $O(\epsilon^{-2} \cdot n'/\log n')$, where $n' = \min(n, m)$. Here we show an alternative proof of this result. Specifically, we shall reduce ϵ-testing m-grained distributions over $[n]$ to $\Omega(\epsilon)$-testing m-grained distributions over $[O(m)]$, and apply Corollary 13.

The reduction will consist of using a deterministic filter $f : [n] \to [k]$, where $k = O(m)$, which will be selected uniformly at random among all such filters. We stress that this is fundamentally different from the randomized filters F used in the main text. Specifically, when applying F several times to the same input, we obtained outcomes that are independently and identically distributed, whereas when we apply a function f (which is selected at random) several times to the same input we obtain the same output.

Note that applying any function $f : [n] \to [k]$ to any m-grained distribution yields an m-grained distribution. Our main result is that, for any distribution X over $[n]$ that is ϵ-far from being m-grained, for almost all functions $f : [n] \to [O(m)]$, the distribution $f(X)$ is $\Omega(\epsilon)$-far from being m-grained.

Lemma 18 (relative preservation of distance from m-grained distributions): *For all sufficiently small $c > 0$, the following holds. For all sufficiently large n and $m < c \cdot n$, any $\epsilon > 0$ and any distribution X over $[n]$ that is ϵ-far from being m-grained, with probability at least $1 - 32c$ over the choice of a function $f : [n] \to [m/c]$, the distribution $f(X)$ is $0.01 \cdot \epsilon$-far from being m-grained.*

Proof: Let $k = m/c$ and let $p : [n] \to [0, 1]$ denote the probability function that describes X. Define $r : [n] \to [0, 1/m)$ such that $r(i) = p(i) - \lfloor m \cdot p(i) \rfloor / m$. Denoting by $\Delta_G(p)$ the statistical distance between p and the set of m-grained distributions (i.e., half the norm-1 distance), we have

$$\begin{aligned} 2 \cdot \Delta_G(p) &\geq \sum_{i \in [n]} \min(r(i), (1/m) - r(i)) \\ 2 \cdot \Delta_G(p) &\leq 2 \cdot \sum_{i \in [n]} \min(r(i), (1/m) - r(i)) \end{aligned}$$

where the first inequality is due to the need to transform each $p(i)$ to a multiple of $1/m$ and the second inequality is justified by a two-stage correction process: first we round each $p(i)$ to the closest multiple of $1/m$, and then we correct the resulting function so that it sums up to 1 (while keeping its values as multiples of $1/m$). Hence, the lemma's hypothesis implies that $\sum_{i\in[n]} \min(r(i), (1/m) - r(i)) > \epsilon$, and we shall prove the lemma by lower-bounding (w.h.p.) the corresponding sum that refers to the distribution $f(X)$, when f is selected at random. (Specifically, for $p'(j) = \sum_{i:f(i)=j} p(i)$, we shall lower-bound the probability that $\sum_{j\in[k]} \min(r'(j), (1/m) - r'(j)) > c \cdot \epsilon$, where $r'(j) = p'(j) - \lfloor m \cdot p'(j) \rfloor$.)

Before doing so, we introduce a few notations. Firstly, we let $s(i) = \min(r(i), (1/m) - r(i))$, and let $\sigma = \sum_{i\in[n]} s(i)$, which is greater than ϵ by the hypothesis. Next, we let $H = \{i \in [n] : p(i) \geq 1/3m\}$ denote the set of "heavy" elements in X. We observe that $|H| \leq 3m$ and that for every $i \in \overline{H} \stackrel{\text{def}}{=} [n] \setminus H$ it holds that $s(i) = r(i)$, since $r(i) < 1/3m < 1/2m$. We consider two cases, according to whether or not the sum $\sum_{i\in\overline{H}} r(i)$ is smaller than $0.5 \cdot \sigma$.

The first case is that $\sum_{i\in\overline{H}} r(i) < 0.5 \cdot \sigma$, and in this case $\sum_{i\in H} s(i) > 0.5 \cdot \sigma$. Consider a uniformly selected function $f : [n] \to [k]$. We consider two good events w.r.t this probability space.

1. The first event is the event that the function f maps at least 0.4σ of the $s(i)$-mass of the i's in H to distinct images. Intuitively, this is very likely given the total $s(i)$-mass of i's in H is greater than 0.5σ and that $|H| \ll k$. Formally, denoting by H_f the (random variable that represents the) set of $i \in H$ that satisfy $f(i) \notin f(H \setminus \{i\})$ (i.e., for every $i \in H_f$ it holds that $f^{-1}(f(i)) \cap H = \{i\}$), we claim that $\mathbf{Pr}[\sum_{i\in H_f} s(i) > 0.4\sigma] > 1 - c$.

   To see this, we first note that, for every $i \in H$, conditioned on the values assigned to $H \setminus \{i\}$, the probability that $f(i) \notin f(H \setminus \{i\})$ is at least $\frac{k-(|H|-1)}{k} > 1 - |H|/k \geq 0.9$, where the inequality is due to $3m \leq 0.1k$. Hence, each $i \in H$ contributes $s(i) \leq 1/2m$ to the sum with probability at least 0.9, also when conditioned on all other values assigned by f. It follows that $\mathbf{Pr}[\sum_{i\in H_f} s(i) > 0.4\sigma] > 1 - c$, where the case of $\sigma = \omega(1/m)$ is straightforward.[11]

2. The second event is the event that the function f does not map much $s(i)$-mass of i's in $\overline{H}$ to the images occupied by H. Again, this is very likely given that $|H| \ll k$. Specifically, observe that $\mathbb{E}[\sum_{i\in\overline{H}:f(i)\in f(H)} r(i)] \leq \frac{|H|}{k} \cdot \sigma \leq 3c \cdot \sigma$, since $p(i) = r(i)$ for every $i \in \overline{H}$ (and $|H| \leq 3m$ and $k = m/c$). Letting $R = \sum_{i\in\overline{H}:f(i)\in f(H)} r(i)$, we get $\mathbf{Pr}[R < 0.2\sigma] > 1 - 15c$.

Assuming that the two good events occur (which happens with probability at least $1 - 16c$), it follows that at least 0.4σ of the $s(i)$-mass of H is mapped by f to distinct images and at most 0.2σ of the rest of the probability space is mapped to these images. Hence, $f(X)$ corresponds to a probability function p' such that $\sum_{i\in H_f} \min(r'(i), (1/m) - r'(i)) > 0.4\sigma - 0.2\sigma = 0.2\sigma$, where $r'(i) = p'(i) - \lfloor m \cdot p'(i) \rfloor$. Hence, with probability at least $1 - 6c$ over the choice of f, it holds that $f(X)$ is 0.2ϵ-far from being m-grained.

The second case is that $\sigma' \stackrel{\text{def}}{=} \sum_{i\in\overline{H}} r(i) \geq 0.5 \cdot \sigma$. In this case, we first show that much of the

[11] Specifically, letting ζ_i denote the contribution of $i \in H$ to the said sum, we have $\mathbb{E}[\zeta_i] > 0.9s(i)$ and $\mathbb{V}[\zeta_i] \leq \mathbb{E}[\zeta_i^2] \leq s(i)/2m$. Hence, $\mathbf{Pr}[\sum_{i\in H} \zeta_i \leq 0.4\sigma] < \frac{\sigma/2m}{(0.5\sigma)^2}$. This suffices for $\sigma = \omega(1/m)$. Actually, the same argument holds if $\sum_{i\in H} s(i)^2 = o(\sigma^2)$. Finally, if $\sum_{i\in H} s(i)^2 = \Omega(\sigma^2)$, then $O(1)$ of the $s(i)$'s sum-up to more than 0.4σ, and in this case the claim follows by focusing on these i's.

probability mass of $\overline{H}$ is mapped disjointly of H. That is,

$$\mathbf{Pr}_{f:[n]\to[k]}\left[\sum_{i\in\overline{H}:f(i)\notin f(H)} r(i) > 0.5\cdot\sigma'\right] > 1-7c \tag{4}$$

where the probability is taken uniformly over all possible choices of f. The proof is similar to the analysis of the foregoing second event. Specifically, we consider random variables ζ_i's such that $\zeta_i = r(i)$ if $f(i)\notin f(H)$ and $\zeta_i = 0$ otherwise, and observe that $\mathbb{E}[\zeta_i] \geq \frac{k-|H|}{k}\cdot r(i) \geq (1-3c)\cdot r(i)$ (since $|H|\leq 3m$ and $m=ck$). Thus, $\mathbb{E}[\sum_{i\in\overline{H}}\zeta_i] \geq (1-3c)\cdot\sigma'$ and Eq. (4) follows by Markov Inequality while using $\zeta_i \leq r(i)$. This holds also if we fix the values of f on H and condition on it, which is what we do from this point on. Hence, we fix an arbitrary sequence of value for $f(H)$, and consider the uniform distribution of f conditioned on this fixing as well as on the event in Eq. (4). In other words, we consider a random choice of $f:\overline{H}\to[k]$ such that $\sum_{i\in\overline{H}:f(i)\in J} r(i) > 0.5\cdot\sigma'$, where $J \stackrel{\text{def}}{=} [k]\setminus f(H)$. Considering the random variable $J_f = \{j\in J : \sum_{i:f(i)=j} p(i) \leq 0.8/m\}$, our aim is to show that $\sum_{i\in\overline{H}:f(i)\in J_f} r(i) > 0.1\cdot\sigma'$.

Towards this end, we consider an iterative process of selecting random values for $f:\overline{H}\to[k]$ such that in the i^{th} step a random value is assigned to the i^{th} element of $\overline{H}$. For every $j\in[k]$ and $t\in[n']$, where $n'=|\overline{H}|$, let $\eta_{j,t}$ denote weight of $r(i)$'s that correspond to i's that assigned to j (i.e., $f(i)=j$) in the first t steps. Let $\eta_j = \eta_{j,n'}$. Define $\eta'_j = 0$ if $\eta_j < 1/3m$, and $\eta'_j = \eta_{j,t}$ if $\eta_{j,t}\geq 1/3m$ and $\eta_{j,t-1} < 1/3m$. Then, there are at most $3m\cdot\sigma'$ indices j such that $\eta'_j > 0$, and each of these j's satisfies $\eta'_j \leq 2/3m$ (since $\eta_{j,t}-\eta_{j,t-1} \leq \max_{i\in\overline{H}}\{r(i)\} \leq 1/3m$). On the other hand, $\mathbb{E}[\eta_j - \eta'_j|\eta'_j > 0] \leq 1/k$, since this expectation refers to the weight assigned to j in steps following the step that causes $\eta'_j > 0$ (and so can be bounded just as $\mathbb{E}[\eta_j]\leq 1/k$). Hence,

$$\begin{aligned}
\mathbb{E}_f\left[\sum_{j\in J\setminus J_f}\eta_j-\eta'_j\right] &\leq \sum_{j\in[k]}\mathbf{Pr}[\eta'_j>0]\cdot\mathbb{E}[\eta_j-\eta'_j|\eta'_j>0]\\
&\leq \sum_{j\in[k]}\mathbf{Pr}[\eta'_j>0]/k\\
&\leq 3m\cdot\sigma'/k
\end{aligned}$$

which is bounded by $3c\cdot\sigma'$ (since $k=m/c$). Using $\mathbb{E}[\eta_j-\eta'_j|\eta'_j>0]\leq 1/k$, we can also infer that $\mathbf{Pr}[\eta_j > 0.8/m|\eta'_j>0] \leq 10m/k = 10c$, because this event requires $\eta_j-\eta'_j > 0.1/m$ (since $\eta'_j\leq 2/3m$). It follows that $\mathbb{E}[\eta'_j|\eta_j > 0.8/m] \leq 10c\cdot 2/3m$, and so

$$\begin{aligned}
\mathbb{E}_f\left[\sum_{j\in J\setminus J_f}\eta'_j\right] &\leq \sum_{j\in[k]}\mathbf{Pr}[\eta_j>0.8/m]\cdot\mathbb{E}[\eta'_j|\eta_j>0.8/m]\\
&\leq \sum_{j\in[k]}\mathbf{Pr}[\eta'_j>0]\cdot 20c/3m\\
&< 7c\cdot\sigma'
\end{aligned}$$

Combining the two bounds, we have $\mathbb{E}_f\left[\sum_{j\in J\setminus J_f}\eta_j\right] \leq 10c\sigma'$. Hence, $\mathbf{Pr}_f\left[\sum_{j\in J\setminus J_f}\eta_j \geq 0.4\sigma'\right] \leq$

$10c/0.4 = 25c$, and (recalling that $\mathbf{Pr}[\sum_{i\in\overline{H}:f(i)\in J} r(i) > 0.5\cdot\sigma'] > 1-7c$) it follows that

$$\mathbf{Pr}_f\left[\sum_{i\in\overline{H}:f(i)\in J_f} r(i) > 0.1\sigma'\right] > 1-32c.$$

This means that at least $0.1\sigma' \geq 0.05\sigma$ of the $s(\cdot)$-mass of $\overline{H}$ is mapped by f to a set of $j \in [k]\setminus f(H)$ such that $\sum_{i:f(i)=j} p(i) \leq 0.8/m$, which means that the $s(\cdot)$-mass mapped to such j's may go down from at most $0.5/m$ to at least $0.2/m$, which constitutes a loss of a 2.5-factor. It follows that, with probability at least $1-32c$ over the choice of f, the distribution $f(X)$ corresponds to a probability function p' such that $\sum_{i\in\overline{H}:f(i)\in J_f} \min(r'(i),(1/m)-r'(i)) > 0.05\sigma/2.5 = 0.02\sigma$, where $r'(i) = p'(i) - \lfloor m\cdot p'(i)\rfloor$. The claim follows. ■

References

[1] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *42nd FOCS*, pages 442–451, 2001.

[2] T. Batu, L. Fortnow, R. Rubinfeld, W.D. Smith, and P. White. Testing that Distributions are Close. In *41st FOCS*, pages 259–269, 2000.

[3] C.L. Canonne. A Survey on Distribution Testing: Your Data is Big. But is it Blue? *ECCC*, TR015-063, 2015.

[4] S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal Algorithms for Testing Closeness of Discrete Distributions. In *25th ACM-SIAM Symposium on Discrete Algorithms*, pages 1193–1203, 2014.

[5] I. Diakonikolas and D. Kane. A New Approach for Testing Properties of Discrete Distributions. `arXiv:1601.05557 [cs.DS]`, 2016.

[6] I. Diakonikolas, D. Kane, V. Nikishkin. Testing Identity of Structured Distributions. In *26th ACM-SIAM Symposium on Discrete Algorithms*, pages 1841–1854, 2015.

[7] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. In preparation.

[8] O. Goldreich. *Introduction to Property Testing: Lecture Notes*. In preparation. Drafts are available from `http://www.wisdom.weizmann.ac.il/~oded/pt-ln.html`

[9] O. Goldreich and D. Ron. On Testing Expansion in Bounded-Degree Graphs. *ECCC*, TR00-020, March 2000.

[10] O. Goldreich and D. Ron. On the relation between the relative earth mover distance and the variation distance (an exposition). Available from `http://www.wisdom.weizmann.ac.il/~oded/p_remd.html`

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

[12] G. Valiant and P. Valiant. Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new CLTs. In *43rd ACM Symposium on the Theory of Computing*, pages 685-694, 2011.

[13] G. Valiant and P. Valiant. Instance-by-instance optimal identity testing. *ECCC*, TR13-111, 2013.

[14] G. Valiant and P. Valiant. Instance Optimal Learning. `CoRR abs/1504.05321`, 2015.