Lecture Notes for Testing Properties of Distributions

Oded Goldreich*

May 26, 2016

Summary: We provide an introduction to the study of testing properties of distributions, where the tester obtains samples of an unknown distribution (resp., samples from several unknown distributions) and is required to determine whether the distribution (resp., the tuple of distributions) has some predetermined property. We focus on the problems of testing whether an unknown distribution equals a fixed distribution and of testing equality between two unknown distributions. Our presentation is based on reductions from the general cases to some seemingly easier special cases. In addition, we also provide a brief survey of general results.

The current notes are based on many sources; see Section 5.1 for details.

Teaching note: Unless one intends to devote several lectures to the current topic, one cannot hope to cover all material in this chapter in class. In such a case, we recommend focusing on Sections 1 and 2, leaving Sections 3 and 4 for optional independent reading. Note that Section 3 is quite technical, whereas Section 4 is an overview section.

Key notations: We consider *discrete* probability distributions. Such distributions have a finite *support*, which we assume to be a subset of $[n]$, where the $\mathsf{support}$ 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 mass functions like $p : [n] \to [0,1]$ that satisfy $\sum_{i\in[n]} p(i) = 1$. These two representations 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 $[n]$.)

1 The model

The difference between property testing as discussed so far and testing distributions is quite substantial. So far, we have discussed the testing of objects, viewed as functions (equiv., as sequences over some alphabet), under the uniform distribution.[1] That is, the tested object was a function,

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

[1] An extension of this study to testing properties of functions under arbitrary distributions was briefly mentioned in the first lecture, but not discussed further. A different extension, pursued in the lecture on testing in the general graph model, focused on testing propoerties of graphs that are accessible via various types of queries (without specifying their representation).

and the tested property was a property of functions (equiv., a set of functions). Furthermore, the tester was given query access to the tested object, and the (uniform) distribution was used merely as a basis for defining distance between objects.[2]

In contrast, in the context of testing distributions, the tested object is a distribution, the tested property is a propery of distributions (equiv., a set of distributions), and the tester (only) obtains samples drawn according to the tested distribution. For example, we may be given samples that are drawn from an arbitrary distribution over $[n]$, and be asked to "determine" whether the given distribution is uniform over $[n]$.

The foregoing formulation raises some concerns. We can never determine, not even with (non-trivial) error probability, whether samples that are given to us were taken from some fixed distribution. That is, given $s(n)$ (say $s(n) = 2^n$) samples from $X \in [n]$, we cannot determine whether or not X is the uniform distribution, since X may be such that $\mathbf{Pr}[X=i] = \frac{1}{n} - \frac{1}{2^n s(n)}$ if $i \in [n-1]$ and $\mathbf{Pr}[X=n] = \frac{1}{n} + \frac{n-1}{2^n s(n)}$ otherwise. Of course, what is missing is a relaxed interpretation of the term "determine" (akin to the interpretation we gave when defining approximate decision problems).

But before presenting this natural relaxation, we stress that here exact decision faces an impossiblity result (i.e., any finite number of samples does not allow to solve the exact decision problem), whereas in the context of deciding properties of functions exact decision "only" required high complexity (i.e., ruled out decision procedures of sub-linear query complexity).

The natural choice of a relaxation (for the aforementioned task) is to only require the rejection of distributions that are far from having the property, where the distance between distributions is defined 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 provide the definition of testing properties of distributions.

1.1 Testing properties of single distributions

Having specified the objects (i.e., distributions), the view obtained by the tester (i.e., samples), and the distance between objects (i.e., Eq. (1)), we can apply the "testing" paradigm and obtain the following definition. (Let us just stress that, unlike in the context of testing properties of functions, the tester is not an oracle machine but is rather an ordinary algorithm that is given a predetermimed number of samples.)[3]

Definition 1 (testing properties of distributions): *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 $[n]$, 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 $s(n,\epsilon)$ samples drawn from an unknown distribution $X \in [n]$, satisfies the following two conditions.*

[2] Actually, we also mentioned (in the first lecture) and used (in the lecture on "implicit sampling") the notion of testing functions based on random examples.

[3] Indeed, such ordinary machines are also used in the case of sample-based testing, discussed in the first lecture and defined in the lecture on "implicit sampling". In both cases, the sample complexity is stated as part of the basic definition, rather than being introduced later (as a relevant complexity measure). (We deviate from this convention in Exercise 6.)

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 are 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.

If the tester accepts every distribution in $\mathcal{D}$ with probability 1, then we say that it has one-sided error.

Note that n fully specifies the set of distributions $\mathcal{D}_n$, and we do not consider the computational complexity of obtaining an explicit description of $\mathcal{D}_n$ from n (not even when $\mathcal{D}_n$ is a singleton). For sake of simplicity, in the rest of this lecture, we will consider a generic n and present the relevant properties as properties of distributions over $[n]$.

We comment that testers of one-sided error are quite rare in the context of testing properties of distributions (unlike in the context of testing properties of functions). This phenomenon seems rooted in the following fact. *If there exist a X is in $\mathcal{D}$ and a distribution Y that is not in $\mathcal{D}$ such that the support of Y is a subset of the support of X, then $\mathcal{D}$ has no one-sided error tester* (regardless of the sample complexity).[4]

Relation to learning. As in the context of testing properties of functions, it is possible to reduce testing to learning, alas in the context of testing properties of distributions the cost of such a reduction is larger. Nevertheless, let us outline this reduction.

1. When using proximity parameter ϵ, the tester uses part of the sample in order to learn a distribution in $\mathcal{D}$ such that if the input distribution X is in $\mathcal{D}$ then, with high probability, the learning algorithm outputs a description of a distribution Y in $\mathcal{D}$ that is $\epsilon/2$-close to X.

2. The tester uses a different part of the sample in order to check whether X is $\epsilon/2$-close to Y or is ϵ-far from it.

The problem with this reduction is that, in general, Step 2 has almost linear complexity (i.e., it has complexity $\Omega(n/\log n)$). In contrast, recall that in the context of testing properties of functions, the analogous step has extremely low complexity.[5] Furthermore, in many natural cases (of distribution testing) the cost of Step 2 is significantly higher than the cost of Step 1 (e.g., Step 2 may require $\Omega(n/\log n)$ samples also when Step 1 is trivial, as in the case that $\mathcal{D}$ is the singleton containing the uniform distribution). Hence, like in the context of testing properties of functions, we shall seek to outperform this reduction; however, unlike in the case of testing functions, typically this will

[4] This is because, for some ϵ, the distribution Y is ϵ-far from $\mathcal{D}$, whereas rejecting Y with positive probability implies rejecting X with positive probability, since any sample of Y occurs also as a sample of X. We mention that the condition for the non-existence of one-sided error testers is indeed necessary (see Exercise 1).

[5] Recall that $O(1/\epsilon)$ samples suffice in order to determine whether an unknown input function is $\epsilon/2$-close to a fixed function or is ϵ-far from it.

not be because learning (i.e,., Step 1) is too expensive but rather because Step 2 is too expensive. Nevertheless, in some cases, this reduction or variants of it (cf., e.g., [23, 1]) are very useful. Finally, we note that Step 2 can always be performed by using $O(n/\epsilon^2)$ samples, and the same holds for Step 1 (see [9, Lem. 3]).[6]

Notations: In order to simplify some of the discussion, 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. This refers also to testers of properties of pairs of distributions, defined next.

1.2 Testing properties of pairs of distributions

Definition 1 generalizes naturally to testing properties of m-tuples of distributions, where the cases of $m = 1$ and $m = 2$ are most popular. By a property of m-tuples of distributions, we mean a set of such m-tuples, and in the testing problem we are given samples from the m distributions being tested. For example, given samples from two distributions, one may be asked to test whether they are identical.

Definition 2 (testing properties of m-tuples of distributions):[7] *Let $\mathcal{D}$ be a property of m-tuples 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 m sequences each consisting of $s(n,\epsilon)$ samples drawn from m unknown distributions $X_1, ..., X_m \in [n]$, satisfies the following two conditions.*

1. The tester accepts tuples that belong to $\mathcal{D}$: *If $(X_1, ..., X_m)$ is in $\mathcal{D}$, then*

$$\mathbf{Pr}_{i_1^{(1)},...,i_s^{(1)} \sim X_1; ...; i_1^{(m)},...,i_s^{(m)} \sim X_m}[T(n,\epsilon; i_1^{(1)}, ..., i_s^{(1)}; \ldots; i_1^{(m)}, ..., i_s^{(m)}) = 1] \geq 2/3,$$

   where $s = s(n,\epsilon)$ and $i_1^{(j)}, ..., i_s^{(j)}$ are drawn independently from the distribution X_j.

2. The tester rejects tuples that are far from $\mathcal{D}$: *If $(X_1, ..., X_m)$ is ϵ-far from any tuple in $\mathcal{D}$* (i.e., for every $(Y_1, ..., Y_m)$ in $\mathcal{D}$ the average variation distance between X_j and Y_j, where $j \in [m]$, is greater than ϵ)*, then*

$$\mathbf{Pr}_{i_1^{(1)},...,i_s^{(1)} \sim X_1; ...; i_1^{(m)},...,i_s^{(m)} \sim X_m}[T(n,\epsilon; i_1^{(1)}, ..., i_s^{(1)}; \ldots; i_1^{(m)}, ..., i_s^{(m)}) = 0] \geq 2/3,$$

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

[6] It turns out that approximating an unknown distribution $X \in [n]$ by the "empirical distribution" of $O(n/\epsilon^2)$ samples will do. The analysis, presented in Exercise 3, is highly recommended. As a motivation, we point out that naive attempts at such an analysis do not yield the desired result. For example, one may seek to approximate each $p(i)$ up to an additive term of $\epsilon/4n$ (or so), but this will require $\Omega(n/\epsilon)^2$ samples. A less naive attempt is based on the observation that it suffices to have a $1 + 0.1\epsilon$ factor approximation of each $p(i) \geq 0.1\epsilon/n$ (as well as a list containing all i's such that $p(i) < 0.1\epsilon/n$). Such an approximation can be obtained, with high probability, using a sample of size $\widetilde{O}(n)/\epsilon^2$. That is, for each i, using a sample of such size, with probability at least $1/3n$, we either provide a $1 + 0.1\epsilon$ factor approximation of $p(i)$ or detect that $p(i) < 0.1\epsilon/n$. As stated upfront, a better approach is presented in Exercise 3. Furthermore, as discussed in Section 4, relaxed forms of both tasks, which suffice for many testing problems, can be performed using $O(\epsilon^{-2} \cdot n/\log n)$ samples (see [23, Thm. 1]).

[7] The current definition mandates that the same number of samples are given for each of the m distributions. A more flexible definition that allows a different sample size for each distribution is natural and has been used in several studies.

We stress that the property that consists of pairs of identical distributions (i.e., $\{(D_1, D_2) : D_1 = D_2\}$) is a property of pairs of distributions. In contrast the property that consists of being identical to a fixed distribution D (i.e., the property $\{D\}$) is a property of (single) distributions. In the former case, the tester is given samples from two unknown distributions, whereas in the latter case the tester is given samples from one unknown distribution (whereas the fixed distribution D is a ("massive") parameter of the testing problem).

Note that, for any $m > 1$, testing m-tuples of distributions includes testing $(m-1)$-tuples of distributions as a special case (e.g., by just ignoring the last distribution). On the other hand, testing m-tuples of distributions reduces to testing the single distribution that corresponds to the Cartesian product of the m distributions, but this (single distribution) testing task may be harder than the original testing task (for m-tuples), because the tester also has to deal with the case that the input distribution is not a product of m distributions. (In contrast, when testing an m-tuple of distributions, the tester is guaranteed that the samples provided for the various m distributions are independent.)[8]

1.3 Label-invariant properties

A very natural class of properties of distributions consists of *label invariant* properties: For a distribution $X \in [n]$ and a permutation $\pi : [n] \to [n]$, we let $Y = \pi(X)$ be the distribution obtained by sampling X and applying π to the outcome; that is, $\mathbf{Pr}[Y = \pi(i)] = \mathbf{Pr}[X = i]$. A property $\mathcal{D}$ of distributions is label invariant if for every distribution $X \in [n]$ in $\mathcal{D}$ and for every permutation $\pi : [n] \to [n]$ the distribution $\pi(X)$ is in $\mathcal{D}$. Likewise, a property $\mathcal{D}$ of m-tuples of distributions is label invariant if for every tuple $(X_1, ..., X_m)$ in $\mathcal{D}$ and for every permutation $\pi : [n] \to [n]$ the tuple $(\pi(X_1), ..., \pi(X_m))$ is in $\mathcal{D}$.

Note that the property that consists of the uniform distribution over $[n]$ and the property that consists of pairs of identical distributions are both label-invariant. On the other hand, the property that consists of a single distribution D that is not uniform over $[n]$ is not label-invariant. Other label-invariant properties include the set of distributions over $[n]$ having support that is smaller than some threshold, and the set of distributions having entropy greater than some threshold.

In general, properties of distributions that only depend on the histograms of the distributions are label-invariant, and vice versa. The histogram of a distribution D over $[n]$ is a multiset of all the probabilities in the distribution D (sorted according to these probabilities); that is, the histogram of the distribution represented by the probability function $p : [n] \to [0, 1]$ is the multiset $\{p(i) : i \in [n]\}$. Equivalently, the histogram of p is the set of pairs $\{(v, m) : m = |\{i \in [n] : p(i) = v\}| > 0\}$.

1.4 Organization

We focus on the problems of testing whether an unknown distribution equals a fixed distribution and of testing equality between two unknown distributions: Solutions to these problems are presented in Sections 2 and 3, respectively. The testers presented have complexity $\mathrm{poly}(1/\epsilon) \cdot n^{1/2}$ and $\mathrm{poly}(1/\epsilon) \cdot n^{2/3}$, respectively, which is the best possible.

[8] Let $\mathcal{D}_1, ..., \mathcal{D}_m$ be properties of distributions. When testing whether the m-tuple of distributions $(X_1, ..., X_m)$ is in $\mathcal{D}_1 \times \cdots \times \mathcal{D}_m$, we are given a sequence $(i_1^{(1)}, ..., i_s^{(1)}; \ldots; i_1^{(m)}, ..., i_s^{(m)})$ such that the $i_k^{(j)}$'s are drawn from X_j independently of all other $i_k^{(j')}$'s (for $j' \neq j$). But when testing whether the distribution $\overline{X} \in [n]^m$ is in $\{\overline{D} : \overline{D} \equiv D_1 \times \cdots \times D_m \ \wedge \ (\forall j) D_j \text{ in } \mathcal{D}_j\}$, we are given a sequence $\overline{i}_1, ..., \overline{i}_s$ such that each $\overline{i}_k$ is drawn independently from $\overline{X}$, but it is not necessarily the case that $\overline{X} \equiv X_1 \times \cdots \times X_m$ for some distributions $X_1, ..., X_m \in [n]$.

In Section 4 we consider the general question of testing properties of (single) distributions and review general results. On the positive side, it turns out that any label-invariant property of distributions can be tested in complexity $\mathrm{poly}(1/\epsilon) \cdot n/\log n$, which means cutting off a logarithmic factor in comparison to the result obtained via the generic learning (mentioned at the end of Section 1.1, see also Exercise 3). On the negative side, it turns out that, for many natural properties, this is the best possible.

2 Testing equality to a fixed distribution

By testing equality to a fixed distribution D, we mean testing whether an unknown distribution over $[n]$ equals the distribution D. In other words, we refer to testing the property $\{D\}$, which is a property of single distributions. Recall that the analogous task is quite trivial in the context of testing properties of functions (i.e., testing whether an unknown function equals a fixed function can be performed by using $O(1/\epsilon)$ random samples). In contrast, ϵ-testing the property $\{D\}$ typically[9] requires $\Omega(\epsilon^{-2} \cdot \sqrt{n})$ samples, and this holds also in the case that D is uniform over $[n]$. It turns out that this bound can always be achieved; that is, for every distribution over $[n]$, testing the property $\{D\}$ can be performed in time $O(\epsilon^{-2} \cdot \sqrt{n})$.

We start by considering the special case in which D is the uniform distribution over $[n]$, denoted U_n. Testing the property $\{U_n\}$ will be reduced to estimating the collision probability of the tested distribution, where the collision probability of a distribution is the probability that two samples drawn independently from it collide (i.e., yield the same value). In Section 2.2 we shall reduce the task of testing the property $\{D\}$, for any D (over $[n]$), to the task of testing the property $\{U_n\}$.

2.1 The collision probability tester and its analysis

The collision probability of a distribution X is the probability that two samples drawn according to X are equal; that is, the collision probability of X is $\mathbf{Pr}_{i,j\sim X}[i=j]$, which equals $\sum_{i\in[n]} \mathbf{Pr}[X=i]^2$. For example, the collision probability of U_n is $1/n$. Letting $p(i) = \mathbf{Pr}[X=i]$, observe that

$$\sum_{i\in[n]} p(i)^2 = \frac{1}{n} + \sum_{i\in[n]} \left(p(i) - n^{-1}\right)^2, \tag{2}$$

which means that the collision probability of X equals the sum of the collision probability of U_n and the square of the $\mathcal{L}_2$-norm of $X - U_n$ (viewed as a vector, i.e., $\|X-U_n\|_2^2 = \sum_{i\in[n]} |p(i)-u(i)|^2$, where $u(i) = \mathbf{Pr}[U_n=i] = 1/n$).

The key observation is that, while the collision probability of U_n equals $1/n$, *the collision probability of any distribution that is ϵ-far from U_n is greater than $\frac{1}{n} + \frac{4\epsilon^2}{n}$*. To see the latter claim let p denote the corresponding probability function and note that if $\sum_{i\in[n]} |p(i) - n^{-1}| > 2\epsilon$, then

$$\sum_{i\in[n]} \left(p(i) - n^{-1}\right)^2 \;\geq\; \frac{1}{n} \cdot \left(\sum_{i\in[n]} \left|p(i) - n^{-1}\right|\right)^2$$

[9] Pathological examples do exist. For example, if D is concentrated on few elements, then the complexity depends on this number rather than on n. A general study of the complexity of ϵ-testing the property $\{D\}$ as a function of D (and ϵ) was carried out by Valiant and Valiant [24]. It turns out that this complexity depends on a (weird) pseudo-norm of D.

$$> \frac{(2\epsilon)^2}{n}$$

where the first inequality is due to Cauchy-Schwarz inequality.[10] Indeed, using Eq. (2), we get $\sum_{i\in[n]} p(i)^2 > \frac{1}{n} + \frac{(2\epsilon)^2}{n}$. This yields the following test.

Algorithm 3 (the collision probability tester): *On input* $(n, \epsilon; i_1, ..., i_s)$, *where* $s = O(\sqrt{n}/\epsilon^4)$, *compute* $c \leftarrow |\{j < k : i_j = i_k\}|$, *and accept if and only if* $\frac{c}{\binom{s}{2}} < \frac{1+2\epsilon^2}{n}$.

Note that Algorithm 3 approximates the collision probability of the distribution X from which the sample is drawn. The quality of this approximation is the key issue here. Recall that the collision probability of $X \in [n]$ is at least $1/n$, and so it stands to reason that a sample of size $O(\sqrt{n})$ can provide some approximation of it, since each pair in the sample provides an unbiased estimator[11] of the collision probability (i.e., for every $j < k$ it holds that $\mathbf{Pr}[i_j = i_k] = \sum_{i\in[n]} \mathbf{Pr}[X = i]^2$).

Lemma 4 (analysis of the collision probability estimation): *Suppose that* $i_1, ..., i_s$ *are drawn from a distribution* X *that has collision probability* μ. *Then,*

$$\mathbf{Pr}\left[\left|\frac{|\{j < k : i_j = i_k\}|}{\binom{s}{2}} - \mu\right| \geq \gamma \cdot \mu\right] < 1/3,$$

provided that $s = \Omega(\gamma^{-2} \cdot \mu^{-1/2})$.

Hence, if X is the uniform distribution (i.e., $\mu = 1/n$), then, with probability at least 2/3, Algorithm 3 accepts (since $\mathbf{Pr}[c/\binom{s}{2} \geq (1+\epsilon^2)/n] < 1/3$).[12] On the other hand, if $\mu > (1+4\epsilon^2)/n$, then (setting $\gamma = \epsilon^2$ again) it follows that $\mathbf{Pr}[c/\binom{s}{2} \leq (1-\epsilon^2) \cdot \mu] < 1/3$, whereas $(1-\epsilon^2) \cdot \mu > (1-\epsilon^2) \cdot (1+4\epsilon^2)/n > (1+2\epsilon^2)/n$. Hence, in this case, with probability at least 2/3, Algorithm 3 rejects.

Proof:[13] As noted before, each pair of samples provides an unbiased estimator of μ. If these pairs of samples would have been pairwise independent, then $O(\gamma^{-2}\mu^{-1})$ such pairs (of pairs) would have sufficed to obtain a $(1+\gamma)$ factor approximation of μ. But the pairs (of pairs) are not pairwise independent, although they are close to being so. Hence, the desired bound is obtained by going inside the standard analysis of pairwise independent sampling, and analyzing the effect of the few pairs (of pairs) that are not independent.

---

[10] That is, use $\sum_{i\in[n]} |p(i) - n^{-1}| \cdot 1 \leq \left(\sum_{i\in[n]} |p(i) - n^{-1}|^2\right)^{1/2} \cdot \left(\sum_{i\in[n]} 1^2\right)^{1/2}$.

[11] A random variable X (resp., an algorithm) is called an unbiased estimator of a quantity v if $\mathbb{E}[X] = v$ (resp., the expected value of its output equals v). Needless to say, the key question with respect to the usefulness of such an estimator is the magnitude of its variance. For example, for any NP-witness relation $R \subseteq \bigcup_{n\in\mathbb{N}}(\{0,1\}^n \times \{0,1\}^{p(n)})$, the (trivial) algorithm that on input x selects at random $y \in \{0,1\}^{p(|x|)}$ and outputs $2^{p(|x|)}$ if and only if $(x, y) \in R$, is an unbiased estimator of the number of witnesses for x, whereas counting the number of NP-witnesses is notoriously hard. The catch is, of course, that this estimation has a huge variance; letting $\rho(x) > 0$ denote the fraction of witnesses for x, this estimator has expected value $\rho \cdot 2^{2\cdot p(|x|)}$ whereas its variance is $(\rho(x) - \rho(x)^2) \cdot 2^{2\cdot p(|x|)}$, which is typically much larger than the expectation squared (i.e., when $0 < \rho(x) \ll 1/\mathrm{poly}(|x|)$).

[12] Indeed, here we use $\gamma = \epsilon^2$.

[13] The following proof is similar to the technical core of the analysis of the Bipartite tester in the bounded-degree graph model.

Specifically, we consider $m = \binom{s}{2}$ random variables $\zeta_{j,k}$ that represent the possible collision events; that is, for $j < k$, let $\zeta_{j,k} = 1$ if the j^{th} sample collides with the k^{th} sample (i.e., $i_j = i_k$) and $\zeta_{j,k} = 0$ otherwise. Then, $\mathbb{E}[\zeta_{j,k}] = \sum_{i\in[n]} \mathbf{Pr}[i_j = i_k = i] = \mu$ and $\mathbb{V}[\zeta_{j,k}] \leq \mathbb{E}[\zeta_{j,k}^2] = \mu$. Letting $\overline{\zeta}_{i,j} \stackrel{\text{def}}{=} \zeta_{i,j} - \mu$ and using Chebyshev's Inequality (while recalling that $\mathbb{V}[Z] = \mathbb{E}[(Z - \mathbb{E}[Z])^2]$), we get:

$$\begin{aligned}\mathbf{Pr}\left[\left|\sum_{j<k}\overline{\zeta}_{j,k}\right| > m\cdot\gamma\mu\right] &< \frac{\mathbb{E}\left[\left(\sum_{j<k}\overline{\zeta}_{j,k}\right)^2\right]}{(m\cdot\gamma\mu)^2}\\ &= \frac{1}{m^2\gamma^2\mu^2}\cdot\sum_{j_1<k_1,j_2<k_2}\mathbb{E}\left[\overline{\zeta}_{j_1,k_1}\overline{\zeta}_{j_2,k_2}\right]\end{aligned}$$

We partition the terms in the last sum according to the number of distinct indices that occur in them such that, for $t \in \{2,3,4\}$, we let $(j_1,k_1,j_2,k_2) \in S_t \subseteq [s]^4$ if and only if $|\{j_1,k_1,j_2,k_2\}| = t$ (and $j_1<k_1 \wedge j_2<k_2$). Hence,

$$\mathbf{Pr}\left[\left|\sum_{j<k}\overline{\zeta}_{j,k}\right| > m\cdot\gamma\mu\right] < \frac{1}{m^2\gamma^2\mu^2}\cdot\sum_{t\in\{2,3,4\}}\sum_{(j_1,k_1,j_2,k_2)\in S_t}\mathbb{E}\left[\overline{\zeta}_{j_1,k_1}\overline{\zeta}_{j_2,k_2}\right] \tag{3}$$

The contribution of each element in S_4 to the sum is zero, since the four samples are independent and so $\mathbb{E}[\overline{\zeta}_{j_1,k_1}\overline{\zeta}_{j_2,k_2}] = \mathbb{E}[\overline{\zeta}_{j_1,k_1}]\cdot\mathbb{E}[\overline{\zeta}_{j_2,k_2}] = 0$. Each element in S_2 (which necessarily satisfies $(j_1,k_1) = (j_2,k_2)$) contributes $\mathbb{E}[\overline{\zeta}_{j,k}^2] = \mathbb{V}[\zeta_{j_1,k_1}] \leq \mu$ to the sum, but there are only m such elements, and so their total contribution is at most $m\cdot\mu$. Turning to S_3, we note that each of its $O(ms)$ elements contributes

$$\begin{aligned}\mathbb{E}[\overline{\zeta}_{1,2}\overline{\zeta}_{2,3}] &\leq \mathbb{E}[\zeta_{1,2}\zeta_{2,3}]\\ &= \sum_{i\in[n]}\mathbf{Pr}[X=i]^3\\ &\leq \mu^{3/2}\end{aligned}$$

where the first inequality holds since the variables have non-negative expectation, and the second inequality holds since $\mathbf{Pr}[X=i] \leq \sqrt{\mu}$ (for each i).[14] Hence, the total contribution of the elements of S_3 is $O(ms)\cdot\mu^{3/2} = O(m\mu)^{3/2}$. Plugging all of this into Eq. (3), we get an upper bound of $\frac{m\mu+O(m\mu)^{3/2}}{m^2\mu^2\gamma^2} = O((m\mu\gamma^4)^{-1/2})$. Recalling that $m = \binom{s}{2} = \Omega(\gamma^{-4}\mu^{-1})$, the claim follows. ■

Reflection. When trying to test label-invariant properties of distributions, the only relevant information provided by the sample is the *collision statistics*, where the collision statistics of the sequence $(i_1,...,i_s)$ is the sequence $(c_1,...,c_t)$ such that c_j denotes the number of elements that occur j times in the sequence (i.e., $c_j = |\{i\in[n] : \#_i(i_1,...,i_s) = j\}|$, where $\#_i(i_1,...,i_s) = |\{k\in$

[14] Recall that X denotes the distribution from which the samples are drawn; hence, $\mathbb{E}[\zeta_{1,2}\zeta_{2,3}] = \sum_{i\in[n]}\mathbf{Pr}[i_1 = i_2 = i_3 = i]$ equals $\sum_{i\in[n]}\mathbf{Pr}[X=i]^3$. (Also, $\mathbf{Pr}[X=i]^2 \leq \mu$, for each i.) We mention that in the second inequality we used $\sum_{i\in[n]}\mathbf{Pr}[X=i]^3 \leq \sqrt{\mu}\cdot\sum_{i\in[n]}\mathbf{Pr}[X=i]^2$, and in the first inequality we used $\mathbb{E}[(Y-\mathbb{E}[Y])\cdot(Z-\mathbb{E}[Z])] = \mathbb{E}[YZ] - \mathbb{E}[Y]\cdot\mathbb{E}[Z]$.

$[s] : i_k = i\}|$). Indeed, by the label-invariance condition, the specific labels of the c_j elements that have each occurred j times do not matter for determining how likely it is that the sample was drawn from a distribution that has the property (or is at any given distance from the property). This is formally proved in Theorem 12. Intuitively, this is the case since, for every distribution $X \in [n]$ and every permutation $\pi : [n] \to [n]$, the sample $(i_1, ..., i_s)$ is as likely to be drawn from X as the sample $(\pi(i_1), ..., \pi(i_s))$ is to be drawn from $\pi(X)$.

The most basic type of information that can be deduced from the collision statistics is an estimate to the collision probability of the original distribution. Given a sequence of samples $(i_1, ..., i_s)$, this estimate is computed as $|\{j < k : i_j = i_k\}|/\binom{s}{2}$. (Letting $(c_1, .., c_t)$ denote the collision statistics, this value equals $\sum_{j \geq 2} \binom{j}{2} \cdot c_j / \binom{s}{2}$.) In any case, this statistic is the basis of the test that is captured by Algorithm 3.

Testing uniformity. As stated right after Lemma 4, an immediate corollary of Lemma 4 is that the property of being the uniform distribution over $[n]$ can be tested in $O(\sqrt{n})$ time.

Corollary 5 (an upper bound on the complexity of testing uniformity): *Let U_n denote the uniform distribution over $[n]$. Then, the property $\{U_n\}$ can be ϵ-tested in sample and time complexity $O(\epsilon^{-4}\sqrt{n})$.*

We comment that an alternative analysis of this tester as well as some closely related tests yield an upper bound of $O(\epsilon^{-2}\sqrt{n})$, which is optimal.[15]

Approximating the $\mathcal{L}_2$ norm. Lemma 4 implies more than a tester for the property $\{U_n\}$. It actually asserts that the collision probability of a distribution can be approximated up to any desired multiplicative factor by using a number of samples that is inversely proportional to the square root of the collision probability. Viewing the collision probability of a distribution as the square of the $\mathcal{L}_2$-norm (i.e., $\|\cdot\|_2$) of the distribution (viewed as a vector), we get

Corollary 6 (approximating the $\mathcal{L}_2$-norm of a distribution):[16] *Given s samples from a unknown distribution p, Algorithm 3 yields an $(1+\gamma)$-factor approximation of $\|p\|_2$ with probability $1 - O(1/(\gamma^2\|p\|_2 \cdot s))$. Furthermore, this estimate equals $\sqrt{c/\binom{s}{2}}$, where c is as computed by Algorithm 3.*

We mention that, in a model that allows the algorithm to obtain samples on demand, the $\mathcal{L}_2$-norm of a distribution can be approximated within expected sample complexity that is inversely related to its norm (see Exercise 4).

Proof: Indeed, Lemma 4 only asserts that $\mathbf{Pr}\left[\left|(c/\binom{s}{2}) - \|p\|_2^2\right| \geq \gamma \cdot \|p\|_2^2\right] < 1/3$, provided that $s = \Omega(\gamma^{-2} \cdot \|p\|_2^{-1})$, but its proof actually establishes

$$\mathbf{Pr}\left[\left|\frac{c}{\binom{s}{2}} - \|p\|_2^2\right| \geq \gamma \cdot \|p\|_2^2\right] = O(1/(\gamma^2\|p\|_2 \cdot s)).$$

Hence, with probability $1 - O(1/(\gamma^2\|p\|_2 \cdot s))$, it holds that $c/\binom{s}{2}$ is $(1 \pm \gamma) \cdot \|p\|_2^2$, and the claim follows. ■

[15] Both the upper bound and the lower bound are due to [19]. Alternative proof of these bounds can be found in [7] (see also [11, Apdx.]) and [10, Sec. 3.1.1], respectively. The fact that $O(\sqrt{n}/\epsilon^2)$ samples actually suffice for the collision probability test (of Algorithm 3) was recently established by Diakonikolas *et al.* [12].

[16] Recall that $\|p\|_2 = \sqrt{\sum_{i \in [n]} p(i)^2}$, which is the square root of the collision probability of p.

2.2 The general case (treated by a reduction to testing uniformity)

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 $\epsilon/4$-close to an $O(n/\epsilon)$-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 7 (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 $q(i) = m_i/m$.*

Clearly, the uniform distribution over $[n]$ is n-grained. More generally, if a distribution D results from applying some function to the uniform distribution over $[m]$, then D is m-grained. On the other hand, any m-grained distribution must have support size at most m.

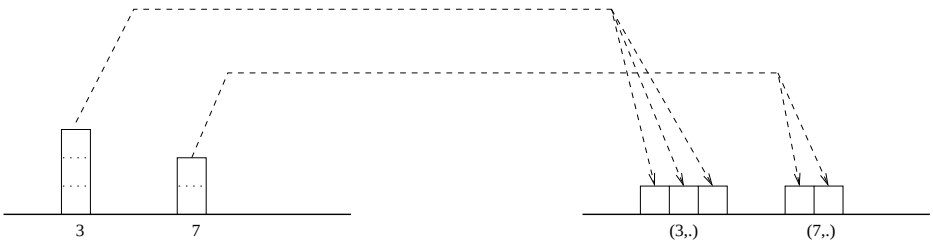

Figure 1: The grained-to-uniform filter (as applied to the fixed 5-grained distribution q that satisfies $q(3) = 3/5$ and $q(7) = 2/5$).

2.2.1 Testing equality to a fixed grained distribution

Fixing any m-grained distribution (represented by a probability 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 $\langle i, 0\rangle$. (An application of this filter is depicted in Figure 1.) 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

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

   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 distribution 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; see Exercise 5.)

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

Algorithm 8 (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 unknown 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 $[m_{i_k}]$, 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 $m+1$.*

(Alternatively, the reduction may just reject if any of the j_k equals 0.)[17]

The foregoing 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 Exercise 6). Hence, *the sample complexity of testing equality to m-grained distributions equals the sample complexity of testing uniformity over* $[m]$ (which is indeed a special case). Using any of the known uniformity tests that have sample complexity $O(\sqrt{n}/\epsilon^2)$,[18] we obtain –

Corollary 9 (testing equality to m-grained distributions): *For any fixed m-grained distribution D, the property $\{D\}$ can be ϵ-tested in sample complexity $O(\sqrt{m}/\epsilon^2)$.*

Note that the complexity of the said tester depends on the level of grainedness of D, which may be smaller than n (i.e., the *a prior* bound on the size of the support of the tested distribution). Hence, the foregoing *tester for equality to grained distributions* is of independent interest, which extends beyond its usage towards testing equality to arbitrary distributions.

2.2.2 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 $\frac{\gamma m_i/n}{q(i)} = 1$, because 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 $q(i) = 0$).[19] 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 as 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 preserving 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.

[17] The justification of this alternative is implicit in Exercise 6 (see Footnote 46). Another alternative is presented in Exercise 7.

[18] Recall that the alternatives include the tests of [19] and [7] or the collision probability test (of Algorithm 3), per its improved analysis in [12].

[19] 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 $q(i) > 0$.

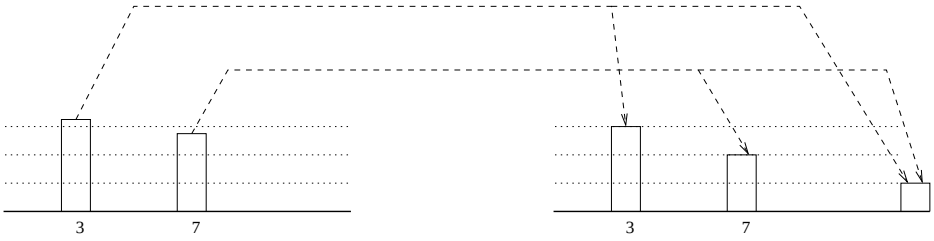

Figure 2: The general-to-grained filter (as applied to part of the fixed distribution q that satisfies $q(3) = 3.2/6n$ and $q(7) = 2.8/6n$). The dotted lines indicate multiples of γ/n.

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

$$p'(i) = \frac{1}{2} \cdot p(i) + \frac{1}{2} \cdot \frac{1}{n}. \tag{4}$$

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

   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 8. Letting $m_i = \lfloor q'(i) \cdot n/\gamma \rfloor$, on input $i \in [n]$, the filter outputs i with probability $\frac{m_i \gamma/n}{q'(i)}$, and outputs $n+1$ otherwise (i.e., with probability $1 - \frac{m_i \gamma/n}{q'(i)}$), where $\gamma > 0$ is a small constant (e.g., $\gamma = 1/6$ will do). (An application of this filter is depicted in Figure 2.)

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

   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

$$p''(i) = p'(i) \cdot \frac{m_i \cdot \gamma/n}{q'(i)} \tag{5}$$

   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 10 (reducing testing equality to a general distribution to testing equality to an $O(n)$-grained distribution): *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 unknown distribution p, the tester proceeds as follows:*

1. *It produces an 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. (4)&(5)*; 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 $[n+1]$.*

We stress that if $F''_{F'(q)} \circ F'$ 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'' that satisfies* Eq. (4)&(5).

Using the notations as in Eq. (4)&(5), 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 $q'' = F''_{q'}(q')$ 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, *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 8 and the one underlying Algorithm 10 is that the former preserves the exact distance between distributions, whereas the later only preserves 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. (4)&(5)) maps different i's to distributions of non-disjoint support. (Specifically, the filter of Eq. (4) maps every $i \in [n]$ to a distribution that assigns each $i' \in [n]$ probability at least $1/2n$, whereas the filter of Eq. (5) typically maps each $i \in [n]$ to a distribution with a support that contains the element $n+1$.)

2.2.3 From arbitrary distributions to the uniform one

Combining the reductions captured by Algorithms 10 and 8, we obtain:

Theorem 11 (testing equality to any fixed distribution): *For any fixed distribution D over $[n]$, the property $\{D\}$ can be ϵ-tested in sample complexity $O(\sqrt{n}/\epsilon^2)$.*

Indeed, this generalizes Corollary 5. We mention that $\Omega(\epsilon^{-2}\sqrt{n})$ is a lower bound for testing $\{D\}$ for many fixed distributions D over $[n]$, including the uniform one. Nevertheless, as indicated by Corollary 9, in some (natural) cases testing the property $\{D\}$ has lower complexity. We mention that the complexity of ϵ-testing the property $\{D\}$ as a function of D (and ϵ) is known [24]; it turns out that this complexity depends on a (weird) pseudo-norm of D.

Proof: We first reduce the problem of ϵ-testing equality to D to the problem of $\epsilon/3$-testing equality to a $O(n)$-grained distribution (by using Algorithm 10), and then reduce the latter task to testing equality over $[O(n)]$ (by using Algorithm 8). Finally, we use any of the known uniformity testers that have sample complexity $O(\sqrt{n}/\epsilon^2)$.[20] ■

2.3 A lower bound

We first establish the claim eluded to in the reflection that follows the proof of Lemma 4. We say that a distribution tester T is label-invariant if it ignores the labels of the samples and only considers their collision statistics. In other words, for every sequence $(i_1, ..., i_s)$ and every permutation $\pi : [n] \to [n]$, the verdict of T on input $(n, \epsilon; i_1, ..., i_s)$ is identical to its verdict on the input $(n, \epsilon; \pi(i_1), ..., \pi(i_s))$.

Theorem 12 (label-invariant algorithms suffice for testing label-invariant properties): *Let $\mathcal{D}$ be a label-invariant property of distributions that is testable with sample complexity s. Then, $\mathcal{D}$ has a label-invariant tester of sample complexity s.*

A similar statement holds for testing label-invariant properties of m-tuples of distributions.

Proof: Given a tester T of sample complexity s for $\mathcal{D}$, consider a tester T' that on input $(n, \epsilon; i_1, ..., i_s)$ selects uniformly a random permutation $\phi : [n] \to [n]$, invokes T on input $(n, \epsilon; \phi(i_1), ..., \phi(i_s))$, and rules accordingly. (Actually, it suffices to select random distinct values $\phi(i_j)$, for the distinct i_j's that appear in the sample.)

By construction, for every sequence $(i_1, ..., i_s)$ and every permutation $\pi : [n] \to [n]$, the verdict of T' on input $(n, \epsilon; i_1, ..., i_s)$ is identical to its verdict on the input $(n, \epsilon; \pi(i_1), ..., \pi(i_s))$. On the

[20] Recall that the alternatives include the tests of [19] and [7] or the collision probability test (of Algorithm 3), per its improved analysis in [12].

other hand, the verdict of T' on distribution X is identical to the output of T on the distribution Y obtained from X by selecting a random permutation ϕ and letting $Y \leftarrow \phi(X)$. Using the label-invariance feature of $\mathcal{D}$, it follows that T' is a valid tester (because, if X is in $\mathcal{D}$ then so is Y, and if X is ϵ-far from $\mathcal{D}$ then so is Y). ■

Corollary 13 (lower bound on the complexity of testing uniformity): *Let U_n denote the uniform distribution over $[n]$. Then, 0.99-testing the property $\{U_n\}$ requires $\Omega(\sqrt{n})$ samples.*

Note that this result does not say how the complexity of ϵ-testing the property $\{U_n\}$ depends on ϵ. Yet, the argument can be extended to show a lower bound of $\Omega(\min(n^{2/3}, \epsilon^{-2}\sqrt{n}))$ on the sample complexity of ϵ-testing $\{U_n\}$ (see Exercise 9). The latter lower bound is not tight either: Recall that it is known that ϵ-testing the property $\{U_n\}$ has sample (and time) complexity $\Theta(\epsilon^{-2}\sqrt{n})$ (cf. [19, 7]).

Proof: Using Theorem 12, it suffices to consider label-invariant testers. Note that, with probability at least $1 - (s^2/n)$, a sequence of s samples that are drawn from the uniform distribution on $[n]$ contains no collisions (i.e., the collision statistics is $c_1 = s$ and $c_j = 0$ for all $j > 1$).[21] But the same happens, with probability $1 - (s^2/(0.01n - 1))$, when the s samples are drawn the uniform distribution on $[0.01n - 1]$, which is 0.99-far from U_n. ■

3 Testing equality between two unknown distributions

Here we consider the problem of testing the property $\{(D_1, D_2) : D_1 = D_2\}$, where (D_1, D_2) denotes a generic pair of distributions (over $[n]$). We stress that this is a property of pairs of distributions, and accordingly the tester obtains samples from each of the two unknown distributions (whose equality is being tested).

The pivot of our presentation is a rather natural algorithm for estimating the $\mathcal{L}_2$-distance between two distributions. This algorithm *takes s samples from each of the distributions, and outputs*

$$\frac{\sqrt{\sum_{i\in[n]} ((x_i - y_i)^2 - (x_i + y_i))}}{s}, \tag{6}$$

where x_i (resp., y_i) *denotes the number of occurrences of i in the sample taken from the first* (resp., second) *distribution.*

To see why this makes sense, suppose first that the number of samples is huge (e.g., $s = \omega(n)$), which is not what we actually want (since we seek algorithms of sublinear complexity). Still, in this case x_i and y_i will reflect the actual probability of item i in each of the two distributions, and so $(\sum_{i\in[n]}(x_i - y_i)^2)^{1/2}/s$ is close to the $\mathcal{L}_2$-distance between the two distributions. Note that this is not exactly the quantity used in Eq. (6).

It turns out that Eq. (6) actually performs better. For starters, it ignores the contribution of items i that appears exactly once (i.e., $x_i + y_i = 1$). This is a good thing because, when $s = o(n)$, such a case indicates nothing and should not "count" towards asserting that the distance between the two distributions is large. In general, the statistic (x_i, y_i) contributes positively if $|x_i - y_i| > \sqrt{x_i + y_i}$, and contributes negatively if $|x_i - y_i| < \sqrt{x_i + y_i}$. This reflects the intuition

[21] Recall that c_j denotes the number of elements that occur j times in the sequence of samples $(i_1, ..., i_s)$; that is, $c_j = |\{i \in [n] : \#_i(i_1, ..., i_s) = j\}|$, where $\#_i(i_1, ..., i_s) = |\{k \in [s] : i_k = i\}|$.

that a deviation of less than a square root of the expectation actually indicates that i is as likely in both distributions. But the question, of course, is *how well does this algorithm approximate the $\mathcal{L}_2$-distance between two distributions?*

Answering this simple question (i.e., analyzing this simple algorithm) turns out non-obvious.[22] In particular, the analysis is simplified if the number of samples is not fixed (possibly as a function of other parameters), but is rather selected at random according to a Poisson distribution. Since this phenomenon is not unique to the current algorithm, but is rather very common within the study of testing properties of distributions, we start with a brief review of the Poisson distribution (and the reasons that it is useful in this study).

3.1 Detour: Poisson Distributions

When we take s samples from a distribution p, the number of occurrences of each value i behave as a binomial distribution with parameters s and $p(i)$; that is, the probability that i occurs t times is $\binom{s}{t} \cdot p(i)^t \cdot (1-p(i))^{s-t}$. But when we condition on the number of occurrences of $j \neq i$, this affects the distribution on the number of occurrences of i, and calculations that depend on the latter distribution become messy. In contrast, if we take a number of samples that is distributed as a Poisson distribution with parameter s (defined next), then the frequency of occurrence of i is independent of the frequency of occurrence of $j \neq i$. This fact is the reason for the popularity of taking a number of samples that is Poisson distributed rather than taking a fixed number of samples. The appeal of this practice is enhanced by the fact (shown in Proposition 15) that the number of samples under the Poisson distribution is well concentrated.

Definition 14 (Poisson distribution): *The* Poisson distribution with parameter $\lambda > 0$*, denoted* $\Psi(\lambda)$*, is a discrete distribution over non-negative integers such that the number k occurs with probability*

$$\frac{\lambda^k \cdot e^{-\lambda}}{k!} \tag{7}$$

where e is the natural base and $0! = 1$. (It is also convenient to fictitiously define the "Poisson distribution" for the parameter 0 (i.e., $\Psi(0)$) as the distribution that is identically 0.)[23]

We first observe that $\sum_{k \geq 0} \frac{\lambda^k \cdot e^{-\lambda}}{k!} = 1$: This follows from the fact that the Taylor expansion of e^x at 0 equals $\sum_{k \geq 0} \frac{e^0}{k!} \cdot (x-0)^k$, which implies that $e^\lambda = \sum_{k \geq 0} \frac{\lambda^k}{k!}$. We next establish the following facts regarding the Poisson distribution.

Proposition 15 (basic facts about the Poisson distribution): *Let $X \leftarrow \Phi(\lambda)$ be a random variable describing a number drawn from the Poisson distribution with parameter $\lambda > 0$. Then:*

1. *The expectation of X equals λ.*
2. *The variance of X equals λ.*

   In general, for every $t \in \mathbb{N}$, it holds that $\mathbb{E}[X^t] = \sum_{i=1}^{t} S(t,i) \cdot \lambda^i$, where $S(t,i) = \frac{1}{i!} \cdot \sum_{j=0}^{i} (-1)^{i-j} \cdot \binom{i}{j} \cdot j^t$ is the Stirling number of the second type.[24]

[22] Recall that this phenomenon is quite common also in the context of testing properties of functions.

[23] This is consistent with the common technical definitions of $0^0 = 0! = 1$.

[24] Recall that $S(t,i)$ is the number of i-partitions of $[t]$; that is, the number of ways to partition $[t]$ into i non-empty sets.

3. For every $\Delta > 0$, it holds that $\mathbf{Pr}[|X - \lambda| > \Delta] = \exp(-\Omega(\Delta^2/(\lambda + \Delta)))$.

We note, for perspective, that $\mathbf{Pr}[X = \lambda] = \Theta(\lambda)^{-1/2}$ for $\lambda > 0$.[25]

Teaching note: The proof of Proposition 15 consists of straightforward manipulations of the probability function of the Poisson distribution (as defined in Eq. (7)). Hence, the proof may be skipped, but the claims are important and should be communicated. The same applies to Proposition 16.

Proof: We first present a recursive formula for $\mathbb{E}[X^t]$. For every $t \geq 1$, we have

$$\begin{aligned}
\mathbb{E}[X^t] &= \sum_{k\geq 0} \frac{\lambda^k \cdot e^{-\lambda}}{k!} \cdot k^t \\
&= \lambda \cdot \sum_{k\geq 1} \frac{\lambda^{k-1} \cdot e^{-\lambda}}{(k-1)!} \cdot k^{t-1} \\
&= \lambda \cdot \sum_{k\geq 1} \frac{\lambda^{k-1} \cdot e^{-\lambda}}{(k-1)!} \cdot \sum_{i=0}^{t-1} \binom{t-1}{i} \cdot (k-1)^i \\
&= \lambda \cdot \sum_{i=0}^{t-1} \binom{t-1}{i} \cdot \sum_{k\geq 0} \frac{\lambda^k \cdot e^{-\lambda}}{k!} \cdot k^i.
\end{aligned}$$

Hence, we get

$$\mathbb{E}[X^t] = \lambda \cdot \sum_{i=0}^{t-1} \binom{t-1}{i} \cdot \mathbb{E}[X^i]. \tag{8}$$

Fact 1 follows from Eq. (8) (for $t = 1$) by using $\mathbb{E}[X^0] = 1$. Fact 2 follows from Eq. (8) (for $t = 2$) by using $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda \cdot (1 + \lambda) - \lambda^2$. The general formula for $\mathbb{E}[X^t]$ follows by induction on t (and using $S(0,0) = 1$ and $S(0,j) = S(j,0) = 0$ for $j \geq 1$):

$$\begin{aligned}
\mathbb{E}[X^t] &= \lambda \cdot \sum_{i=0}^{t-1} \binom{t-1}{i} \cdot \mathbb{E}[X^i] \\
&= \lambda \cdot \sum_{i=0}^{t-1} \binom{t-1}{i} \cdot \sum_{j=0}^{i} S(i,j) \cdot \lambda^j \\
&= \sum_{j=0}^{t-1} \sum_{i=j}^{t-1} \binom{t-1}{i} \cdot S(i,j) \cdot \lambda^{j+1} \\
&= \sum_{j=0}^{t-1} S(t,j+1) \cdot \lambda^{j+1}
\end{aligned}$$

where the last equality uses the combinatorial identity $S(t, j+1) = \sum_{i=j}^{t-1} \binom{t-1}{i} \cdot S(i,j)$.

[25] This holds since

$$\mathbf{Pr}[X = \lambda] = \frac{\lambda^\lambda \cdot e^{-\lambda}}{\lambda!} = \frac{\lambda^\lambda \cdot e^{-\lambda}}{\Theta(\lambda^{1/2}) \cdot (\lambda/e)^\lambda} = \Theta(\lambda)^{-1/2}.$$

Turning to Fact 3, for every $k > 0$, we have

$$\begin{aligned}
\mathbf{Pr}[X = \lambda - k] &= \frac{\lambda^{-k}}{(\lambda - k)!/(\lambda!)} \cdot \mathbf{Pr}[X = \lambda] \\
&< \lambda^{-k} \cdot \prod_{i=0}^{k-1} (\lambda - i) \\
&= \prod_{i=0}^{k-1} \left(1 - \frac{i}{\lambda}\right) \\
&< \left(1 - \frac{(k/2) - 1}{\lambda}\right)^{k/2} \\
&\approx \exp(-k^2/4\lambda),
\end{aligned}$$

where the approximation is up to constant factors. Similarly,

$$\begin{aligned}
\mathbf{Pr}[X = \lambda + k] &= \frac{\lambda^{k}}{(\lambda + k)!/(\lambda!)} \cdot \mathbf{Pr}[X = \lambda] \\
&< \lambda^{k} \cdot \prod_{i=1}^{k} (\lambda + i)^{-1} \\
&= \prod_{i=1}^{k} \left(1 - \frac{i}{\lambda + i}\right) \\
&< \left(1 - \frac{k/2}{\lambda + (k/2)}\right)^{k/2} \\
&\approx \exp(-k^2/(4\lambda + 2k)).
\end{aligned}$$

The claim follows. ■

The relevance to the study of sampling algorithms. We now turn to our original motivation for reviewing the Poisson distribution. Recall that $\Psi(s)$ denotes the Poisson distribution with parameter s.

Proposition 16 (Poisson sampling): *Let $p : [n] \to [0, 1]$ be a distribution and suppose that we select m according to $\Psi(s)$, and then select m samples from the distribution p. Then, the numbers of occurrences of the various values $i \in [n]$ are independently distributed such that the number of occurrences of the value i is distributed as $\Psi(s \cdot p(i))$.*

(The implies that if X_i's are selected independently such that X_i is a Poisson distribution with parameter λ_i, then $\sum_i X_i$ is a Poisson distribution with parameter $\sum_i \lambda_i$.)

Proof Sketch: We prove the claim for $n = 2$, but the proof generalizes easily.[26] Let X denote the number of occurrences of the value 1, and Y denote the number of occurrences of the value 2.

[26] Alternatively, the claim can be proved by induction on m.

Then, for every k and ℓ, it holds that

$$\begin{aligned}\mathbf{Pr}[X=k \wedge Y=\ell] &= \frac{s^{k+\ell}\cdot e^{-s}}{(k+\ell)!}\cdot\binom{k+\ell}{k}\cdot p(1)^k\cdot p(2)^\ell \\ &= \frac{(s\cdot p(1))^k\cdot(s\cdot p(2))^\ell\cdot e^{-s\cdot p(1)}\cdot e^{-s\cdot p(2)}}{k!\cdot\ell!} \\ &= \frac{(s\cdot p(1))^k\cdot e^{-s\cdot p(1)}}{k!}\cdot\frac{(s\cdot p(2))^\ell\cdot e^{-s\cdot p(2)}}{\ell!}\end{aligned}$$

which equals $\mathbf{Pr}[X=k]\cdot\mathbf{Pr}[Y=\ell]$. ■

3.2 The actual algorithm and its analysis

Having defined (and discussed) the Poisson distribution, we now present the actual algorithm that we shall analyze. This algorithm depends on a parameter s, which will determine the distribution of the number of samples obtained from two unknown distributions, denoted p and q.

Algorithm 17 (the basic $\mathcal{L}_2$-distance estimator): *On input parameters n and s, and access to $m' \leftarrow \Psi(s)$ samples from an unknown distribution p and to $m'' \leftarrow \Psi(s)$ samples from an unknown distribution q, the algorithm proceeds as follows.*

1. *For each $i \in [n]$, let x_i denote the number of occurrences of i in the sample taken from p, and y_i denote the number of occurrences of i in the sample taken from q.*
2. *Compute $z \leftarrow \sum_{i\in[n]}((x_i-y_i)^2-(x_i+y_i))$.*

   If $z<0$ output a special symbol, otherwise output $\sqrt{z}/s$.

Recall that by Item 3 of Proposition 15, it holds that $\mathbf{Pr}[|m-s|>s] = \exp(-\Omega(s))$. Hence, Algorithm 17 yields an algorithm that always uses $2s$ samples from each of the distributions. This algorithm selects $m' \leftarrow \Psi(s)$ and $m'' \leftarrow \Psi(s)$, aborts in the highly rare case that $\max(m',m'')>2s$, and otherwise invokes Algorithm 17 while providing it the first m' samples of p and the first m'' samples of q.

We now turn to the analysis of Algorithm 17. Let X_i (resp., Y_i) denote the number of occurrences of i when taking $\Psi(s)$ samples from distribution p (resp., q), and let $Z_i = (X_i-Y_i)^2-(X_i+Y_i)$. By Proposition 16, X_i (resp., Y_i) is a Poisson distribution with parameter $s\cdot p(i)$ (resp., $s\cdot q(i)$). The next (key) lemma implies that $\mathbb{E}[Z_i] = (s\cdot p(i)-s\cdot q(i))^2$, whereas $\mathbb{V}[Z_i]$ can be bounded by a specific degree three polynomial in $s\cdot p(i)$ and $s\cdot q(i)$. Actually, it is important to assert that the degree 3 term has the form $O(s^3)\cdot(p(i)+q(i))\cdot(p(i)-q(i))^2$.

Lemma 18 (the expectation and variance of the Z_i's): *Suppose that $X \leftarrow \Psi(a)$ and $Y \leftarrow \Psi(b)$ are independent Poisson distributions, and let $Z=(X-Y)^2-(X+Y)$. Then, $\mathbb{E}[Z]=(a-b)^2$ and $\mathbb{V}[Z] \le B(a,b)$ for some universal bivariate polynomial B of degree three. Furthermore, $B(a,b) = O((a-b)^2\cdot(a+b)+(a+b)^2)$.*

Proof Sketch: For the expectation of Z, using Proposition 15, we have

$$\begin{aligned}\mathbb{E}[(X-Y)^2-(X+Y)] &= \mathbb{E}[X^2-2XY+Y^2]-(a+b) \\ &= \mathbb{E}[X^2]-2\cdot\mathbb{E}[X]\cdot\mathbb{E}[Y]+\mathbb{E}[Y^2]-(a+b) \\ &= (a^2+a)-2ab+(b^2+b)-(a+b)\end{aligned}$$

which equals $(a-b)^2$ as asserted. Turning to the variance of Z, we only provide a proof of the main part. By Part 2 of Proposition 15, for every $t \in \mathbb{N}$, there exists a degree t polynomial P_t such that $\mathbb{E}[\Psi(\lambda)^t] = P_t(\lambda)$; furthermore, $P_t(z) = z^t + P'_{t-1}(z)$, where P'_{t-1} has degree $t-1$ (and free term that equals zero). Using this fact, it follows that

$$\begin{aligned}\mathbb{V}[(X-Y)^2-(X+Y)] &= \mathbb{E}[((X-Y)^2-(X+Y))^2] - \mathbb{E}[(X-Y)^2-(X+Y)]^2 \\ &= \mathbb{E}[(X-Y)^4] - 2\cdot\mathbb{E}[(X-Y)^2\cdot(X+Y)] + \mathbb{E}[(X+Y)^2] - ((a-b)^2)^2\end{aligned}$$

which is a bivariate polynomial B of total degree four in a and b, since $\mathbb{E}[X^iY^j] = \mathbb{E}[X^i]\cdot\mathbb{E}[Y^j] = P_i(a)\cdot P_j(b)$ for every $i,j \in \mathbb{N}$. Furthermore, using the aforementioned form of P_t (i.e., $P_t(z) = z^t + P'_{t-1}(z)$), it follows that B is of degree three, since the degree-four terms of $\mathbb{E}[(X-Y)^4]$ are cancelled by $(a-b)^4$. This establishes the main claim. A very tedious calculation shows that $B(a,b) = 4\cdot(a-b)^2\cdot(a+b) + 2\cdot(a+b)^2$. (Needless to say, an insightful or at least a non-painful proof of the fact that $B(a,b) = O((a-b)^2\cdot(a+b)+(a+b)^2)$ would be most welcome.) ■

Teaching note: The proofs of the next four results are rather technical. In our applications (see Section 3.3), we shall only use Corollary 22, and the reader may just take this result on faith. The proof of Corollary 19 illustrates the benefit of Poisson sampling, by relying on the fact that the X_i's (resp., Y_i's) are independent. The proofs of Theorem 20 and Corollaries 21 and 22 are rather tedious, and reading them can serve as an exercise.

Corollary 19 (the expectation and variance of the square of the output of Algorithm 17): *Let X_i (resp., Y_i) denote the number of occurrences of i when taking $\Psi(s)$ samples from distribution p (resp., q), and let $Z_i = (X_i-Y_i)^2-(X_i+Y_i)$ and $Z = \sum_{i]\in[n]} Z_i$. Then, $\mathbb{E}[Z] = s^2\cdot\|p-q\|_2^2$ and $\mathbb{V}[Z] = O(s^3\cdot\|p-q\|_2^2\cdot\beta + s^2\beta^2)$, where $\beta = \max(\|p\|_2,\|q\|_2) \geq 1/\sqrt{n}$.*

Hence, Z/s^2 is an unbiased estimator of $\mu \stackrel{\text{def}}{=} \|p-q\|_2^2$, whereas $\mathbb{V}[Z/s^2] = O(\mu\cdot\beta/s) + O(\beta^2/s^2)$. It follows that the probability that Z/s^2 deviates from μ by more than ϵ is

$$\frac{O(\mu\beta)}{s\cdot\epsilon^2} + \frac{O(\beta^2)}{s^2\cdot\epsilon^2}. \tag{9}$$

For $\epsilon = \Omega(\mu)$, Eq. (9) simplifies to $O(\beta/s\epsilon) + O(\beta/s\epsilon)^2$, which means that setting $s = \Omega(\beta/\epsilon)$ will do. Before exploring this direction, let us prove Corollary 19.

Proof: Invoking Lemma 18, we have

$$\begin{aligned}\mathbb{E}[Z] &= \sum_{i\in[n]}\mathbb{E}[Z_i] \\ &= \sum_{i\in[n]}(s\cdot p(i)-s\cdot q(i))^2\end{aligned}$$

which equals $s^2\cdot\|p-q\|_2^2$.

We now turn to the analysis of $\mathbb{V}[Z]$. The key fact here is that the Z_i's are (pairwise) independent. This follows by the independence of the X_i's (resp., Y_i's), where the independence of the X_i's (resp., Y_i's) follows by Proposition 16, whereas the X_i's are independent of the Y_i's by definition.

(Indeed, this is the reason that m' and m'' were generated independently of one another.) Now, invoking Lemma 18, we have

$$\begin{aligned}\mathbb{V}[Z] &= \sum_{i\in[n]} \mathbb{V}[Z_i] \\ &= \sum_{i\in[n]} B(s\cdot p(i), s\cdot q(i)),\end{aligned}$$

where $B(a,b) = O((a-b)^2\cdot(a+b)+(a+b)^2)$. Applying Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\sum_{i\in[n]} (p(i)-q(i))^2\cdot(p(i)+q(i)) &\leq \left(\sum_{i\in[n]}(p(i)-q(i))^4\right)^{1/2}\cdot\left(\sum_{i\in[n]}(p(i)+q(i))^2\right)^{1/2} \\ &\leq \|p-q\|_4^2\cdot\|p+q\|_2 \\ &\leq \|p-q\|_2^2\cdot\|p+q\|_2.\end{aligned}$$

Finally, using

$$\begin{aligned}\sum_{i\in[n]} B(s\cdot p(i), s\cdot q(i)) &= O(s^3)\cdot\sum_{i\in[n]}(p(i)-q(i))^2\cdot(p(i)+q(i)) + O(s^2)\cdot\sum_{i\in[n]}(p(i)+q(i))^2 \\ &\leq O(s^3)\cdot\|p-q\|_2^2\cdot\|p+q\|_2 + O(s^2)\cdot\|p+q\|_2^2,\end{aligned}$$

the claim follows. ■

Algorithm 17 as an approximator of $\mathcal{L}_2$ and $\mathcal{L}_1$ distances. Recall that Algorithm 17 was presented as an $\mathcal{L}_2$-distance approximator. We now establish this feature of Algorithm 17.[27]

Theorem 20 (Algorithm 17 as a $\mathcal{L}_2$-distance approximator): *Suppose that* $\max(\|p\|_2, \|q\|_2) \leq \beta$.

1. *Let* $\gamma \in (0, 0.1)$. *If* $\delta = \|p-q\|_2 > 0$ *and* $s = \Omega(\beta/\gamma^2\delta^2)$, *then, with probability at least* $2/3$, *Algorithm 17 outputs a value in* $(1\pm\gamma)\cdot\delta$.
2. *Let* $\epsilon \in (0,1)$. *If* $s = \Omega(\beta/\epsilon^2)$, *then, with probability at least* $2/3$, *Algorithm 17 distinguishes between the case that* $\|p-q\|_2 \geq \epsilon$ *and the case that* $\|p-q\|_2 \leq \epsilon/2$.

Note that Part 2 is meaningful only for $\epsilon \leq 2\beta$, since $\|p-q\|_2 \leq \|p\|_2 + \|q\|_2 \leq 2\beta$ always holds.

Proof: Recall that Corollary 19 means that $\mathbb{E}[Z/s^2] = \delta^2$ and $\mathbb{V}[Z/s^2] = O(\delta^2\cdot(\beta/s) + (\beta/s)^2)$. Starting with Part 1, we have

$$\begin{aligned}\mathbf{Pr}\left[\left|\frac{Z}{s^2} - \delta^2\right| > \gamma\cdot\delta^2\right] &\leq \frac{\mathbb{V}[Z/s^2]}{(\gamma\delta^2)^2} \\ &\leq \frac{O(\delta^2\cdot\beta)}{s\cdot\gamma^2\delta^4} + \frac{O(\beta^2)}{s^2\cdot\gamma^2\delta^4} \\ &= \frac{O(\beta)}{s\cdot\gamma^2\delta^2} + \gamma^2\cdot\left(\frac{O(\beta)}{s\cdot\gamma^2\delta^2}\right)^2.\end{aligned}$$

[27] Unfortunately, establishing this feature seems to require the sharper analysis of the variance of Z that is provided in the furthermore part of Lemma 18. Recall that this part of Lemma 18 establishes $\mathbb{V}[Z_i] \leq B(p(i),q(i))$, where $B(a,b) = O((a-b)^2\cdot(a+b)+(a+b)^2)$, which implies $\mathbb{V}[Z] = O(s^3\cdot\|p-q\|_2^2\cdot\beta + s^2\beta^2)$, where $\beta = \max(\|p\|_2, \|q\|_2)$ (see Corollary 19). As noted in the proof of Lemma 18, it seems easier to only prove that $\mathbb{V}[Z_i]$ is a degree three polynomial in $\max(p(i),q(i))$, and $\mathbb{V}[Z] = O(s^3\beta^3 + s^2\beta^2)$ will follow (but does not suffice for the following proof).

Using $s = \Omega(\beta/\gamma^2\delta^2)$, we get $\mathbf{Pr}[Z/s^2 = (1 \pm \gamma) \cdot \delta^2] \geq 2/3$, and Part 1 follows (since $\mathbf{Pr}[\sqrt{Z}/s = (1 \pm \gamma)^{1/2} \cdot \delta] \geq 2/3$ and $(1 \pm \gamma)^{1/2} \approx 1 \pm (\gamma/2)$).

Turning to Part 2, we note that by Part 1, if $\|p - q\|_2 \geq \epsilon$ and $s = \Omega(\beta/\epsilon^2)$, then $\mathbf{Pr}[\sqrt{Z}/s < 0.9\epsilon] \leq 1/3$. On the other hand, if $\delta = \|p - q\|_2 \leq \epsilon/2$ and $s = \Omega(\beta/\epsilon^2)$, then (as shown next) $\mathbf{Pr}[\sqrt{Z}/s > 0.6\epsilon] \leq 1/3$. The point is that, in this case, $\mathbb{V}[Z/s^2] = O(\epsilon^2 \cdot (\beta/s) + (\beta/s)^2)$, and we can perform a calculation as in Part 1. Specifically, we get

$$\begin{aligned}
\mathbf{Pr}\left[\frac{\sqrt{Z}}{s} > 0.6\epsilon\right] &\leq \mathbf{Pr}\left[\left|\frac{Z}{s^2} - \delta^2\right| > (0.6^2 - 0.5^2) \cdot \epsilon^2\right] \\
&\leq \frac{\mathbb{V}[Z/s^2]}{\Omega(\epsilon^4)} \\
&\leq \frac{O(\epsilon^2 \cdot \beta)}{s \cdot \epsilon^4} + \frac{O(\beta^2)}{s^2 \cdot \epsilon^4} \\
&= \frac{O(\beta)}{s \cdot \epsilon^2} + \left(\frac{O(\beta)}{s \cdot \epsilon^2}\right)^2
\end{aligned}$$

where the first inequality is due to the fact that $x > v > u$ and $y \leq u$ implies $x^2 - y^2 > v^2 - u^2$. Recalling that $s = \Omega(\beta/\epsilon^2)$, we get $\mathbf{Pr}[\sqrt{Z}/s > 0.6\epsilon] \leq 1/3$, and Part 2 follows. ∎

Corollary 21 (Algorithm 17 as a crude $\mathcal{L}_1$-distance approximator): *Suppose that* $\max(\|p\|_2, \|q\|_2) \leq \beta$ *and let* $\epsilon \in (0, 1)$. *If* $s = \Omega(\beta n/\epsilon^2)$, *then, with probability at least* 2/3, *Algorithm 17 distinguishes between the case that* $p = q$ *and the case that* $\|p - q\|_1 \geq \epsilon$. *In other words, Algorithm 17 yields an* ϵ*-tester of sample complexity* $O(\beta n/\epsilon^2)$ *for equality between two given distributions* (i.e., the property $\{(p, q) : p = q\}$).

In the case that $\beta = O(1/\sqrt{n})$, the claimed tester has sample complexity $O(\sqrt{n}/\epsilon^2)$, which is optimal, but for very large β (e.g., $\beta = \Omega(1)$) this tester is not optimal. Nevertheless, as shown in Section 3.3, Corollary 21 (or rather its revision provided as Corollary 22), can be used towards obtaining optimal testers for the general case (i.e., for arbitrary β).

Proof: Clearly, $p = q$ implies $\|p - q\|_2 = 0$. On the other hand, if $\|p - q\|_1 \geq \epsilon$, then

$$\begin{aligned}
\|p - q\|_2 &= \left(\sum_{i \in [n]} (p(i) - q(i))^2\right)^{1/2} \\
&\geq \sum_{i \in [n]} |p(i) - q(i)| \cdot 1/\sqrt{n} \\
&\geq \epsilon/\sqrt{n}
\end{aligned}$$

where the first inequality is due to Cauchy-Schwarz inequality.[28] By Part 2 of Theorem 20, if $s = \Omega(\beta/(\epsilon/\sqrt{n})^2) = \Omega(\beta n/\epsilon^2)$, then, with probability at least 2/3, Algorithm 17 distinguishes between the case that $\|p - q\|_2 \geq \epsilon/\sqrt{n}$ and the case that $\|p - q\|_2 = 0$, and the claim follows. ∎

[28] That is, use $\sum_{i \in [n]} |p(i) - n^{-1}| \cdot 1 \leq \left(\sum_{i \in [n]} |p(i) - n^{-1}|^2\right)^{1/2} \cdot \left(\sum_{i \in [n]} 1^2\right)^{1/2}$.

From $\max(\|p\|_2, \|q\|_2)$ to $\min(\|p\|_2, \|q\|_2)$. Theorem 20 and Corollary 21 rely on an upper bound on the $\mathcal{L}_2$-norm of *both* distributions. It turns out that (in two of the three cases)[29] it suffices to upper bound the $\mathcal{L}_2$-norm of *one* of the two distributions. This is the case because $\|p-q\| \geq \|p\| - \|q\|$, for any norm $\|\cdot\|$, since $\|q + (p-q)\| \leq \|q\| + \|p-q\|$. Hence, we can first check whether $\|p\|_2 \approx \|q\|_2$, reject if the answer is negative and invoke the algorithm that refers to $\max(\|p\|_2, \|q\|_2)$ otherwise.

Corollary 22 (Part 2 of Theorem 20 and Corollary 21, revised): *Suppose that* $\min(\|p\|_2, \|q\|_2) \leq \beta$.

1. *If $s = \Omega(\beta/\epsilon^2)$ and $\epsilon \in (0, \beta]$, then there exists an algorithm that uses s samples and distinguishes between the case that $\|p-q\|_2 \geq \epsilon$ and the case that $\|p-q\|_2 \leq \epsilon/2$.*
2. *If $s = \Omega(\beta n/\epsilon^2)$ and $\epsilon \in (0,1)$, then there exists an ϵ-tester of sample complexity $O(\beta n/\epsilon^2)$ for equality between two given distributions.*

This result is non-vacuous for $\beta \geq n^{-1/2}$, whereas when $\beta = O(n^{-1/2})$ we can use $s = O(\sqrt{n}/\epsilon^2)$.

Proof: We first approximate $\|p\|_2$ and $\|q\|_2$ by invoking the $\mathcal{L}_2$-approximation algorithm of Corollary 6 with $s = \Omega(1/\beta)$. This allows us to distinguish the case that $\|p\|_2 \leq 2\beta$ from the case that $\|p\|_2 \geq 3\beta$, and ditto for $\|q\|_2$. If one of the two distributions is judged to have norm greater than $2.5 \cdot \beta$ (whereas the other is smaller than β by the hypothesis), then we can safely announce that the distributions are far apart (hereafter referred to as an early verdict). Otherwise, we assume that $\max(\|p\|_2, \|q\|_2) \leq 3\beta$, in which case we can afford to invoke Algorithm 17, where in Part 1 we use $s = O(\beta/\epsilon^2)$ and in Part 2 we use $s = O(\beta n/\epsilon^2)$.

In analyzing this algorithm we assume that the approximation provided by the algorithm of Corollary 6 is within a factor of 1 ± 0.1 of the true value. Hence, if $\max(\|p\|_2, \|q\|_2) > 3\beta$, then (with high probability) this is reflected by the early verdict, since in this case (w.h.p.) the approximate value of $\max(\|p\|_2, \|q\|_2)$ is greater than $2.5 \cdot \beta$. On the other hand, if $\max(\|p\|_2, \|q\|_2) \leq 2\beta$, then (with high probability) the approximate value of $\max(\|p\|_2, \|q\|_2)$ is smaller than $2.5 \cdot \beta$, and we invoke Algorithm 17. (In the latter case, the output of Algorithm 17 is as desired: For Part 1 we use Part 2 of Theorem 20, whereas for Part 2 we use Corollary 21.)

We now show that, when made, the early verdict is rarely wrong. Hence, we assume that $\max(\|p\|_2, \|q\|_2) > 2\beta$, and show that in this case it is justified to assert that p and q are sufficiently far apart. For Part 1 this is justified because $\|p-q\|_2 \geq |\|p\|_2 - \|q\|_2| > 2\beta - \beta \geq \epsilon$, where we use the hypothesis $\epsilon \leq \beta$. In Part 2, we just observe that $\|p\|_2 \neq \|q\|_2$ implies $p \neq q$, which justifies rejection.

It is left to upper bound the sample complexity of the full algorithm. In Part 1 the overall sample complexity is $O(1/\beta) + O(\beta/\epsilon^2) = O(\beta/\epsilon^2)$, where the inequality is due to the hypothesis $\epsilon \leq \beta$. In Part 2 the overall sample complexity is $O(1/\beta) + O(\beta n/\epsilon^2) = O(\beta n/\epsilon^2)$, where the inequality is due to the fact $\beta \geq 1/\sqrt{n}$ (and the hypothesis $\epsilon \leq 1$). ■

3.3 Applications: Reduction to the case of small norms

As noted upfront, Corollary 21 (resp., Corollary 22) is interesting only when the probability distributions have very small $\mathcal{L}_2$-norm (resp., when at least one of the probability distributions has very small $\mathcal{L}_2$-norm). This deficiency is addressed by the following transformation that preserves

[29] Specifically, for Part 2 of Theorem 20 and for Corollary 21.

$\mathcal{L}_1$-distances between distributions, while mapping a target distribution into one of small max-norm (and, hence, small $\mathcal{L}_2$-norm). In other words, the transformation flattens the target distribution (according to max-norm and thus also according to $\mathcal{L}_2$-norm), while preserving $\mathcal{L}_1$-distances between distributions. Hence, the transformation offers a unified way of deriving many testing results by a reduction to the case of small norms. We shall illustrate this phenomenon by presenting two reductions (in Sections 3.3.2 and 3.3.3, respectively).

3.3.1 Flattening distributions

The core of the aforementioned reductions is a (randomized) filter, tailored for a given distribution $q : [n] \to [0,1]$ and a parameter m. This filter maps q to a distribution $q' : [n+m] \to [0,1]$ of max-norm at most $1/m$, which implies that $\|q'\|_2 \leq 1/\sqrt{m}$, while preserving the variation distances between distributions. Setting $m = n$, we obtain a distribution q' with extremely small $\mathcal{L}_2$-norm, since in this case $\|q'\|_2 = O(1/\sqrt{2n})$, where $1/\sqrt{2n}$ is the minimum $\mathcal{L}_2$-norm of any distribution over $[2n]$. But, as we shall see in Section 3.3.3, other settings of m are also beneficial. In any case, it seems fair to say that q' is *flat*, and view the filter as intended to flatten q.

The aforementioned filter is closely related to the filter underlying Algorithm 8. Specifically, for any probability function $q : [n] \to [0,1]$ and a parameter m (e.g., $m = n$), we consider a randomized filter, denoted $F_{q,m}$, that maps $[n]$ to $S = \{\langle i,j\rangle : i \in [n] \wedge j \in [m_i]\}$, where $m_i = \lfloor m \cdot q(i) \rfloor + 1$, such that $F_{q,m}(i)$ is uniformly distributed in $\{\langle i,j\rangle : j \in [m_i]\}$. Hence, if i is distributed according to the probability function p, then each $\langle i,j\rangle \in S$ occurs as output with probability $p(i)/m_i$; that is, if X is distributed according to p, then

$$\mathbf{Pr}[F_{q,m}(X) = \langle i,j\rangle] = p(i) \cdot \frac{1}{m_i}\,. \tag{10}$$

The key observations about this filter are:

1. *The filter $F_{q,m}$ maps q to a distribution with small max-norm*: If Y is distributed according to q, then, for every $\langle i,j\rangle \in S$, it holds that

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

   which is upper-bounded by $1/m$. Hence, the $\mathcal{L}_2$-norm of $F_{q,m}(q)$ is at most $\sqrt{m \cdot (1/m)^2} = \sqrt{1/m} \leq \frac{1+(n/m)}{\sqrt{|S|}}$, where the inequality is due to $|S| = \sum_{i\in[n]} m_i \leq \sum_{i\in[n]}(m \cdot q(i)+1) = m+n$. In case, $m = n$, we get $\|F_{q,m}(q)\|_2 \leq 2/\sqrt{|S|}$.

2. *The filter preserves the variation distance between distributions*: The total variation distance between $F_{q,m}(X)$ and $F_{q,m}(X')$ equals the total variation distance between X and X'. 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 distribution that only depends on i, and it was already proved in the corresponding item of Section 2.2.1 (see also Exercise 5).

In short, the filter $F_{q,m}$ flattens q while preserving the total variation distance between q and any other distribution p. We also stress that knowledge of q (and m) allows to implement $F_{q,m}$ as well as to map S to $[m']$, where $m' = |S|$.

3.3.2 Testing equality to a fixed distribution

The foregoing observations regarding the filter $F_{q,n}$ (when using the setting $m = n$), lead to the following reduction of *testing equality to a fixed distribution* D to the task captured by Part 2 of Corollary 22. (Indedd, this yields an alternative proof of Theorem 11.)

Algorithm 23 (reducing testing equality to an arbitrary distribution to testing equality for distributions of small $\mathcal{L}_2$-norm): *Let D be an arbitrary distribution with probability function $q : [n] \to [0,1]$, and T be an ϵ'-tester of sample complexity $s(m', \beta, \epsilon')$ for equality between distribution pairs over $[m']$ such that at least one of the two distributions has $\mathcal{L}_2$ norm at most β. On input $(n, \epsilon; i_1, ..., i_s)$, where $i_1, ..., i_s \in [n]$ are $s = s(2n, n^{-1/2}, \epsilon)$ samples drawn according to an* unknown *distribution p, the tester proceeds as follows:*

1. *It produces a s-long sequence $(i'_1, ..., i'_s)$ by sampling each i'_k from the* known *distribution D.*
2. *It produces a s-long sequence $(e'_1, ..., e'_s)$ by applying $F_{q,n}$ to $(i'_1, ..., i'_s)$, where $F_{q,n}$ is as in* Eq. (10)*; that is, for every $k \in [s]$, it produces $e'_k \leftarrow F_{q,n}(i'_k)$.*

   (Recall that each e'_k is in S, and that the $\mathcal{L}_2$-norm of $F_{q,n}(q)$ is at most $1/\sqrt{n} \leq 2/\sqrt{|S|}$.)
3. *It produces a s-long sequence $(e_1, ..., e_s)$ by applying $F_{q,n}$ to $(i_1, ..., i_s)$; that is, for every $k \in [s]$, it produces $e_k \leftarrow F_{q,n}(i_k)$.*
4. *It invokes the ϵ-tester T for equality providing it with the sequence sequence $(e_1, ..., e_s, e'_1, ..., e'_s)$. Note that this is a sequence over S, but it can be translated to a sequence over $[m']$ such that $m' = |S|$* (by mapping each element of S to its rank in S).

We stress that if $F_{q,n}$ is invoked t times on the same i, then the t outcomes are (identically and) *independently distributed.*

Hence, *the complexity of testing equality to a general distribution D over $[n]$ is upper-bounded by the complexity of testing equality between two unknown distributions over $[2n]$ such that one of them has $\mathcal{L}_2$-norm at most $1/\sqrt{n}$.* Using Part 2 of Corollary 22, we re-establish Theorem 11.[30]

Digest. We solved a testing task regarding a single unknown distribution by reducing it to a testing task regarding two unknown distributions. This was done (in Step 1 of Algorithm 23) by generating samples from the fixed distribution D, and presenting these samples as samples of a second (supposedly unknown) distribution. Obviously, there is nothing wrong with doing so (i.e., such a reduction is valid), except that it feels weird to reduce a seemingly easier problem to a seemingly harder one. Note, however, that the two problems are not really comparable, since the problem of testing two distributions refers to a special case in which one of these distributions is flat. Indeed, the core of the reduction is the use of the flattening filter, which mapped the fixed distribution to a flat one, and by doing so allows to apply the two-distribution tester (which requires one of the distributions to be flat).

In Section 3.3.3, we shall see a reduction that uses the flattening filter in order to reduce one testing problem regarding two distributions to another problem testing problem regarding two distributions (of which one is flat).

[30] By Part 2 of Corollary 22, the tester T, used in the foregoing reduction, can be implemented within complexity $O(\sqrt{n}/\epsilon^2)$.

3.3.3 Testing equality between two unknown distributions

The filter $F_{q,m}$ captured in Eq. (10) can be applied also to testing properties of tuples of distributions. Actually, this is a more interesting application, since reducing a problem regarding a single unknown distribution to a problem regarding two unknown distributions seems an over-kill. On the other hand, the reader may wonder how one can apply this filter (i.e., the filter $F_{q,m}$) when the distribution (i.e., q) is not known. The answer is that we shall use one part of the sample of q in order to obtain some statistics of q, denoted $\widetilde{q}$, and then use a filter tailored to this statistics (i.e., $F_{\widetilde{q},\widetilde{m}}$). Of course, the larger the sample we take of q, the better statistics $\widetilde{q}$ we derive, which in turn offers lower norm of $F_{\widetilde{q},\widetilde{m}}(q)$. This leads to the following reduction, where m is a parameter that governs the size of the aforementioned sample.

Algorithm 24 (reducing testing equality between pairs of arbitrary distribution to testing equality between pairs of distributions such that at least one of them has a small $\mathcal{L}_2$-norm): *Let T be an ϵ-tester of sample complexity $s(m',\beta,\epsilon)$ for equality between distribution pairs over $[m']$ such that at least one of the two distributions has $\mathcal{L}_2$ norm at most β. On input $(n,\epsilon;i_1,...,i_{s+2m};i'_1,....,i'_s)$, where $i'_1,...,i'_s \in [n]$ are $s = s(n+2m, O(m^{-1/2}),\epsilon)$ samples drawn according to an unknown distribution p and $i_1,...,i_{s+2m} \in [n]$ are $s+2m$ samples drawn according to an unknown distribution q, the tester proceeds as follows:*

1. *Generates $\widetilde{m} \leftarrow \Psi(m)$, and halts and accepts if $\widetilde{m} > 2m$.*

   Let $\widetilde{q} : [n] \to [0,1]$ be the distribution function that corresponds to the sample $(i_{s+1},...,i_{s+\widetilde{m}})$; that is, $\widetilde{q}(i) = |\{k \in [\widetilde{m}] : i_{s+k} = i\}|/\widetilde{m}$.

2. *Produces a s-long sequence $(e'_1,...,e'_s)$ by applying $F_{\widetilde{q},\widetilde{m}}$ to $(i'_1,...,i'_s)$, where $F_{\widetilde{q},\widetilde{m}}$ is as in* Eq. (10)*; that is, for every $k \in [s]$, it produces $e'_k \leftarrow F_{\widetilde{q},\widetilde{m}}(i'_k)$.*

   (Recall that each e'_k is in $S = \{\langle i,j\rangle : i \in [n] \wedge j \in [m_i]\}$, where $m_i = \lfloor \widetilde{m} \cdot \widetilde{q}(i) \rfloor + 1 = \widetilde{m} \cdot \widetilde{q}(i) + 1$. Hence, $|S| = \widetilde{m} + n$.)[31]

   (We shall show that, with high probability, the $\mathcal{L}_2$-norm of $F_{\widetilde{q},\widetilde{m}}(q)$ is at most $O(\sqrt{1/m})$.)

3. *Produces a s-long sequence $(e_1,...,e_s)$ by applying $F_{\widetilde{q},\widetilde{m}}$ to $(i_1,...,i_s)$; that is, for every $k \in [s]$, it produces $e_k \leftarrow F_{\widetilde{q},\widetilde{m}}(i_k)$.*

4. *Invokes the tester T for equality providing it with the input $(n+2m,\epsilon;e_1,...,e_s;e'_1,...,e'_s)$. Note that $(e_1,...,e_s,e'_1,...,e'_s)$ is a sequence over S, but it can be translated to a sequence over $[n+2m]$* (by mapping each element of S to its rank in S).

We stress that if $F_{\widetilde{q},\widetilde{m}}$ is invoked t times on the same i, then the t outcomes are (identically and) *independently distributed.*

[31] **Advanced comment:** We stress that we use the filter $F_{\widetilde{q},\widetilde{m}}$ rather than the filter $F_{\widetilde{q},n}$; in other words, we used $m_i = \lfloor \widetilde{m} \cdot \widetilde{q}(i) \rfloor + 1$ rather than $m_i = \lfloor n \cdot \widetilde{q}(i) \rfloor + 1$. Intuitively, this choice represents a lower "penalty" on i's such that $q(i) \ll 1/m$, since under our choice the ratio between the m_i's reflects better the ratio between the $q(i)$'s. Specifically, suppose that $q(i) = 1/cm$ and $q(j) = c/m$, for some large constant $c > 1$. Then, under our choice $m_i \geq 1$, whereas $m_j = O(1)$, with high probability. In contrast, under the alternative choice, both $m_i = 1$ and $m_j = \Omega(n/m)$ hold, with high probability. The formally inclined reader may trace the effect of this difference in the proof of Lemma 25.

Recall that, for every $\widetilde{q}$ (and $\widetilde{m}$), the total variation distance between $F_{\widetilde{q},\widetilde{m}}(p)$ and $F_{\widetilde{q},\widetilde{m}}(q)$ equals the total variation distance between p and q. Hence, the analysis of Algorithm 24 reduces to proving that, with high probability, it holds that the $\mathcal{L}_2$-norm of $F_{\widetilde{q},\widetilde{m}}(q)$ is at most $O(\sqrt{1/m})$.

Lemma 25 (the $\mathcal{L}_2$-norm of $F_{\widetilde{q},\widetilde{m}}(q)$): *Let $\widetilde{m}$ and $\widetilde{q}$ be as in Algorithm 24. Then, for every t, the probability that $\|F_{\widetilde{q},\widetilde{m}}(q)\|_2$ exceeds $t \cdot m^{-1/2}$ is lower than t^{-2}.*

We stress that this lemma refers to a probability space that includes the event that $\widetilde{m} > 2m$, but this event occurs with probability $\exp(-m)$ and it can be ignored (in the analysis of Algorithm 24).

Proof: We first bound the expected square of the $\mathcal{L}_2$-norm of $F_{\widetilde{q},\widetilde{m}}(q)$, where the expectation is taken over the sample of q that defines $\widetilde{q}$ (and over the choice of $\widetilde{m} \leftarrow \Psi(m)$). Let ζ_i be a random variable representing the distribution of m_i; that is, $\zeta_i - 1$ equals $|\{k \in [\widetilde{m}] : i_k = i\}|$, which indeed equals $\widetilde{m} \cdot \widetilde{q}(i)$. Then, for fixed $\widetilde{m}$ and $(i_{s+1}, ..., i_{s+2m}$ (which determines $\widetilde{q}$ and S), the square of the $\mathcal{L}_2$-norm of $q' = F_{\widetilde{q},\widetilde{m}}(q)$ equals

$$\sum_{\langle i,j\rangle \in S} q'(\langle i,j\rangle)^2 = \sum_{i\in[n]} \sum_{j\in[\zeta_i]} (q(i)/\zeta_i)^2 = \sum_{i\in[n]} q(i)^2/\zeta_i.$$

Hence, our task is to upper-bound $\mathbb{E}[1/\zeta_i]$, while assuming $q(i) > 0$ (as otherwise $\zeta_i \equiv 1$). Recalling that (by Proposition 16) the random variable $\zeta'_i = \zeta_i - 1$ is distributed as $\Psi(m \cdot q(i))$, we have[32]

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{1+\zeta'_i}\right] &= \mathbb{E}\left[\int_0^1 x^{\zeta'_i} dx\right] \\
&= \int_0^1 \mathbb{E}\left[x^{\zeta'_i}\right] dx \\
&= \int_0^1 e^{(x-1)\cdot m\cdot q(i)} dx \\
&= \frac{1 - e^{-m\cdot q(i)}}{m \cdot q(i)}
\end{aligned}$$

which is at most $1/(m \cdot q(i))$. Hence, the expected value of the $\|F_{\widetilde{q},\widetilde{m}}(q)\|_2^2$ equals

$$\begin{aligned}
\mathbb{E}\left[\sum_{i\in[n]} q(i)^2/\zeta_i\right] &= \sum_{i\in[n]} q(i)^2 \cdot \mathbb{E}[1/\zeta_i] \\
&\leq \sum_{i\in[n]} \frac{q(i)^2}{m \cdot q(i)}
\end{aligned}$$

which equals $1/m$. Using Markov's inequality, we have $\mathbf{Pr}[\|F_{\widetilde{q},\widetilde{m}}(q)\|_2^2 > t^2/m] < 1/t^2$. ■

[32] The first equality is due to the fact that for every $c \in \mathbb{N} \cup \{0\}$ it holds that $\int_0^1 x^c dx = (1-0)/(c+1)$. The third equality is due to the fact that for every $r \in [0,1]$ it holds that $\mathbb{E}[r^{\Psi(\lambda)}] = e^{(r-1)\lambda}$, which can be proved by straightforward manipulations of the probability function of the Poisson distribution (as defined in Eq. (7)).

Setting the parameter m. Algorithm 24 works under any choice of the parameter $m = \Omega(1)$, and combined with Part 2 of Corollary 22 it yields an ϵ-tester of sample complexity $O(m + (n + 2m) \cdot m^{-1/2}/\epsilon^2)$. Needless to say, we set m such as to minimize this expression, which means using $m = \min(n^{2/3}/\epsilon^{4/3}, n)$. Hence, we get

Theorem 26 (testing equality of two unknown distributions): *The property consisting of pairs of identical distributions over $[n]$* (i.e., $\{(D, D) : D \in [n]\}$) *can be ϵ-tested in sample and time complexity $O(\max(n^{2/3}/\epsilon^{4/3}, \sqrt{n}/\epsilon^2))$.*

We mention that this result is tight; that is, ϵ-testing equality of two unknown distributions over $[n]$ requires $\Omega(\max(n^{2/3}/\epsilon^{4/3}, \sqrt{n}/\epsilon^2))$ samples [26] (see also [7]).

4 On the complexity of testing properties of distributions

As noted at the end of Section 1.1, any distribution $p : [n] \to [0, 1]$ can be learned up to accuracy of ϵ by a $O(n/\epsilon^2)$-time algorithm (Exercise 3). Thus, our focus is on testers that outperform this bound. We have already seen such testers in Sections 2 and 3, but here we address the question of testing proprties of distributions in full generality.

A very general positive answer is provided via "learning distributions up to relabelling" (where the notion of "relabelling" is implicit in Section 1.3). Specifically, we call the distribution $q : [n] \to [0, 1]$ a relabelling of the distribution $p : [n] \to [0, 1]$ if there exists a permutation $\pi : [n] \to [n]$ such that $q(i) = p(\pi(i))$ for every $i \in [n]$. Equivalently, we may consider the task of learning the histogram of an unknown distribution $p : [n] \to [0, 1]$, where the histogram of p is defined as the set of pairs $\{(v, m) : m = |\{i \in [n] : p(i) = v\}| > 0\}$.[33] The following result of Valiant and Valiant [23] asserts that *the histogram of an unknown distribution can be learned faster* (and using less samples) *than the distribution itself, where the saving is of a logarithmic factor.*

Theorem 27 (learning the histogram):[34] *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 $1 - \exp(-n^{\Omega(1)})$, a histogram of a distribution that is ϵ-close to p.*

(The error probability is stated here, since error reduction to such a (lower than usual) level would have increased the time and sample complexities by more than a $O(\log n)$ factor.) The implication of this result on testing any label-invariant property of distributions is immediate.

Corollary 28 (testing label-invariant properties of single distributions): *Let $\mathcal{D}$ be a label-invariant property of distributions over $[n]$. Then, $\mathcal{D}$ has a tester of sample complexity $s(n, \epsilon) = O(\epsilon^{-2} \cdot n/\log n)$.*

The tester consists of employing the algorithm of Theorem 27 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 $\mathcal{D}$. Using

[33] Note that this is one of the two equivalent definitions of a histogram that were presented in Section 1.3. We prefer this definition here since it yields a more succinct representation.

[34] Valiant and Valiant [23] 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 [25, Fact 1] (using $\tau = 0$), and a detailed proof appears in [18].

the same idea, we get algorithms for estimating the distance of an unknown distribution to any label-invariant property of distributions. Actually, obtaining such an estimation may be viewed as a special case of Corollary 28, by considering, for any property $\mathcal{D}$ and any distance parameter $\delta > 0$, the set of all distributions that are δ-close to $\mathcal{D}$.

On the negative side, it turns out that, for many natural properties, the foregoing tester is the best possible (up to a factor of $1/\epsilon$). This fact is stated in Corollary 30, which is proved based on Theorem 29.

Theorem 29 (optimality of Theorem 27):[35] *For every sufficiently small $\eta > 0$, there exist two distributions $p_1, p_2 : [n] \to [0,1]$ that are indistinguishable by $O(\eta n/\log n)$ samples although p_1 is η-close to the uniform distribution over $[n]$ and p_2 is η-close to the uniform distribution over $[n/2]$.*[36]

Hence, learning the histograms of distributions in the sense stated in Theorem 27 (even with proximity parameter $\epsilon = 1/5$) requires $\Omega(n/\log n)$ samples.[37] Furthermore, as detailed in Claim 30.1, any property that contains all distributions that are close to the uniform distribution over $[n]$ but is far from the uniform distribution over $[n/2]$ cannot be tested by $o(n/\log n)$ samples. Ditto for a property that contains all distributions that are close to the uniform distribution over $[n/2]$ but is far from the uniform distribution over $[n]$. In particular:

Corollary 30 (optimality of Corollary 28): *For all sufficiently small constant $\delta > 0$, testing each of the following* (label-invariant) *properties of distributions over $[n]$ requires $\Omega(n/\log n)$ samples.*

1. *The set of distributions that are δ-close to the uniform distribution over $[n]$.*
2. *The set of distributions that are δ-close to having support size $n/2$ and not having any element in the support that has probability less than $1/n$.*
3. *The set of distributions that are δ-close to being m-grained, for any $m \in [\Omega(n), O(n)]$.*

Here, testing means ϵ-testing for a sufficiently small constant $\epsilon > 0$. Furthermore, the bound holds for any $\delta \in (0, \Omega(\eta_0))$ and any $\epsilon \in (0, 0.5 - 2\delta)$, where $\eta_0 \in (0, 0.25)$ is the constant implicit in Theorem 29 (i.e., in the phrase "for all sufficiently small $\eta > 0$").[38]

Note that the lower bound does not necessarily hold for the "base property" (i.e., the case of $\delta = 0$): Item 1 provides a striking example, since (as we saw) the uniform distribution over $[n]$ is testable by

[35] Like in Footnote 34, we note that Valiant and Valiant [23] 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 [25, Fact 1] (using $\tau = 0$), and a detailed proof appears in [18].

[36] Here indistinguishability means that the distinguishing gap of such potential algorithms is $o(1)$. Note that the statement is non-trivial only for $\eta < 1/4$, since the uniform distribution over $[n]$ is 0.5-close to the uniform distribution over $[n/2]$.

[37] This is the case because otherwise, given $o(n/\log n)$ samples of p_1 (resp., p_2), w.h.p., the algorithm outputs a histogram of a distribution that is ϵ-close to p_1 (resp., p_2), which in turn is η-close to the uniform distribution over $[n]$ (resp., over $[n/2]$). But by Theorem 29 the output in these two cases is distributed almost identically, which implies that (w.h.p.) this output describes a distribution that is $(\epsilon + \eta)$-close both to the uniform distribution over $[n]$ and to the uniform distribution over $[n/2]$, which is impossible since these two distributions are at distance $1/2$ apart (whereas we can have $\epsilon = 1/5$ and $\eta < 1/20$).

[38] The case of small $\delta > 0$, which may depend on ϵ, is typically called "tolerant testing" (for the "base property"); see Parnas, Ron, and Rubinfeld [20].

$O(\sqrt{n})$ samples. (The restriction on m in Item 3 is inherent; for example, note that any distribution over $[n]$ is ϵ-close to being n/ϵ-grained.)

Proof: We first detail the general observation that underlies all results, while letting U_m denote the uniform distribuition over $[m]$.

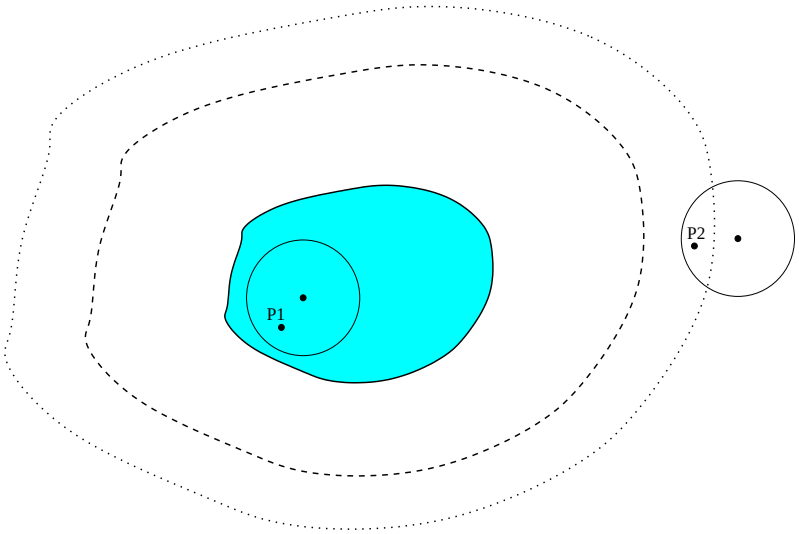

Figure 3: The proof of Claim 30.1. The shaded region represents $\mathcal{D}$, and the dashed (resp., dotted) line represents distance ϵ (resp., $\epsilon+\eta$) from $\mathcal{D}$. The left (resp., right) circle represent the set of distributions that are η-close to U_n (resp., to $U_{n/2}$), which contains p_1 (resp., p_2).

Claim 30.1 (the general observation): *Let $\eta \in (0, \eta_0]$ and suppose that $\mathcal{D}$ is a property of distributions over $[n]$ such that all distributions that are η-close to U_n are in $\mathcal{D}$ and $U_{n/2}$ is $(\epsilon+\eta)$-far from $\mathcal{D}$. Then, ϵ-testing $\mathcal{D}$ requires $\Omega(n/\log n)$ samples. Ditto when all distributions that are η-close to $U_{n/2}$ are in $\mathcal{D}$ and U_n is $(\epsilon+\eta)$-far from $\mathcal{D}$.*

Proof: We focus on the primary claim. Invoking Theorem 29, observe that p_1 is in $\mathcal{D}$ (since p_1 is η-close to U_n), whereas p_2 is ϵ-far from $\mathcal{D}$ (since p_2 is η-close to $U_{n/2}$, which is $(\epsilon+\eta)$-far from $\mathcal{D}$). (See Figure 3.) The main claim follows since Theorem 29 asserts that p_1 and p_2 are indistinguishable by $o(n/\log n)$ samples, wheraes ϵ-testing $\mathcal{D}$ requires distinguishing them. The secondary claim follows by reversing the roles of p_1 and p_2 (i.e., noting that in this case p_2 is in $\mathcal{D}$ whereas p_1 is ϵ-far from $\mathcal{D}$). ■

Using Claim 30.1, we establish the various items of the current corollary. Specifically, denoting by $\mathcal{D}_i$ the set of distributions defined in Item i, we proceed as follows. For $\epsilon+2\delta<1/2$, we recall that $\mathcal{D}_1$ equals the set of all distributions that are δ-close to U_n, and observe that $U_{n/2}$ is $(\epsilon+\delta)$-far from $\mathcal{D}_1$ (since otherwise $U_{n/2}$ is $((\epsilon+\delta)+\delta)$-close to U_n, which contradicts $\epsilon+2\delta<1/2$). Item 1 follows by applying the primary claim (with $\eta=\min(\delta,\eta_0)$).

Turning to Item 2, for $\epsilon+2\delta<1/2$, observe that $\mathcal{D}_2$ contains all distributions that are δ-close to $U_{n/2}$ whereas U_n is $(\epsilon+\delta)$-far from $\mathcal{D}_2$ (since otherwise U_n is $((\epsilon+\delta)+\delta)$-close to a distribution with support size $n/2$, which contradicts $\epsilon+2\delta<1/2$). Item 2 follows by applying the secondary

claim (with $\eta = \min(\delta, \eta_0)$). The same holds for Item 3 when $m = n/2$, but we have to handle the other cases too. For $m < n/2$ we proceed as in the case of $m = n/2$, 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 $m > n/2$ (satisfying $m = O(n)$), we provide a lower bound by reducing the case of $m \in (0.25n, 0.5n]$ to the case of $m = O(n)$.

Claim 30.2 (a reduction for Item 3): *Let $\mathcal{G}_{n,m,\delta}$ denote the set of distributions over $[n]$ that are δ-close to being m-grained. Then, for every $\epsilon > 0$ and $t \in \mathbb{N}$, the task of ϵ-testing $\mathcal{G}_{n,m,\delta}$ is reducible to the task of ϵ/t-testing $\mathcal{G}_{n+1,tm,\delta/t}$, while preseving the number of samples.*

Proof Sketch: Consider a randomized filter, denoted $F_{n,t}$, that with probability $1/t$ maps $i \in [n]$ to itself, and otherwise maps it to $n+1$. This filter maps m-grained distributions over $[n]$ to tm-grained distributions over $[n+1]$. Furthermore, a distribution $p : [n] \to [0,1]$ that is at distance d from being m-grained is mapped by $F_{n,t}$ to a distribution that is at distance d/t from being tm-grained. (This follows by considering the distribution of $F_{n,t}(p)$ conditioned on obtaining a value in $[n]$, and noting that the condition holds with probability $1/t$.) ■

Item 3 follows by using an adequate $t = O(1)$. Specifically, wishing to establish the claim for $m > n/2$, pick $t = \lceil 2m/n \rceil$ and reduce from ϵ-testing $\mathcal{G}_{n,\lfloor m/t \rfloor,\delta}$, which yields a lower bound for ϵ/t-testing $\mathcal{G}_{n,m',\delta/t}$ such that $m' = t \cdot \lfloor m/t \rfloor \in (m-t, m]$. (See Exercise 11 for a reduction to the case of $m' = m$.) ■

5 Final notes

As stated at the very beginning of this lecture, testing properties of distributions, also known as *distribution testing*, is fundamentally different from testing properties of functions (as discussed in the rest of this course). Nevertheless, as observed in [21, Sec. 2.1] and detailed in [17, Sec. 6.3], testing properties of distribution is closely related to testing *symmetric* properties of functions (i.e., properties that are invariant under all permutations of the domain). Articulating this relation requires stating the complexity of testers of (symmetric) properties of functions in terms of the size of the range of the function (rather than in terms of the size of its domain).[39]

The key observation is that, when testing symmetric properties of functions over a domain that is significantly larger than the range, we may confine our attention to the frequency in which the various range elements appear as values of the function. Hence, a function $f : S \to R$ is identified with the distribution generated by selecting uniformly at random $s \in S$ and outputting $f(s)$. In such a case, we may restrict the tester to obtaining the value of the function at uniformly distributed arguments, *while ignoring the identity of the argument*.[40]

While the foregoing perspective attempts to link distribution testing (i.e., testing properties of distributions) to the rest of property testing, the following perspective that advocates the study of distributions testing of super-linear (sample) complexity goes in the opposite direction. Recall that any property of functions can be tested by querying all arguments of the function (i.e., locations in the object), and that the aim of property testing is to obtain sub-linear time (or at least query)

[39] Of course, one may use a statement that refers to the sizes of both the domain and the range.

[40] The hypothesis that the domain is sufficiently large justifies ignoring the probability that the same $s \in S$ was selected twice.

complexity. In contrast, distribution testing does not trivialized when one obtains $O(n)$ samples from a distribution over $[n]$. In particular, learning such a distribution up to a deviation of ϵ requires $\Omega(n/\epsilon^2)$ samples. So the question is whether one can do better than this yardstick. While the study of property testing typically focuses on the dependence on n, as noted by Ilias Diakonikolas, in some settings of distribution testing, one may care about the dependence on ϵ more than about the dependence on n.

5.1 History and credits

The study of testing properties of distributions was initiated by Batu, Fortnow, Rubinfeld, Smith, and White [4].[41] Generalizing a test of uniformity, which was implicit in the work of Goldreich and Ron [16], Batu *et al.* [4, 3] presented testers for the property consisting of pairs of identical distributions as well as for all properties consisting of any single distribution.[42] Both results are presented in this text, but the presentation follows an approach proposed recently by Diakonikolas and Kane [10].

Our presentation focused on two classes of properties of distributions – the class of single-distribution properties that are singletons (i.e., testing equality to a *known* distribution) and the class of pairs of distributions that are equal or close according to some norm. We start with the tester for the property of being the uniform distribution over $[n]$, which is implicit in [16], where it is applied to test the distribution of the endpoint of a relatively short random walk on a bounded-degree graph. (As noted in the text, the analysis that we present yields optimal sample complexity in terms of n, but not in terms of ϵ; a recent result of Diakonikolas *et al.* [12] establishes the optimality of this tester over both n and ϵ.)[43]

Next, we apply the approach that underlies [10] in order to reduce testing any property consisting of a single distribution (i.e., testing equality to a *known* distribution) to testing the uniform distribution; this reduction appeared in [14].

Turning to the task of testing equality between a pair of *unknown* distributions, we start with a (sample optimal) tester for the case that the distributions have small $\mathcal{L}_2$-norm, which is provided in [7], and then apply the reduction presented in [10].

The results surveyed in Section 4 are due to Valiant and Valiant [23]. We stress that the current chapter covers ony few of the "testing of distributions" problems that were studied in the last decade and a half (see, e.g., [2, 22, 23, 26]). The interested reader is referred to Canonne's survey [5] (which also reviews alternative models such as the model of conditional sampling [6]).

Lastly, we mention the work of Daskalakis, Diakonikolas, and Servedio, which crucially uses testing as a tool for learning [8]. Indeed, the use of testing towards learning is in line with the one of the generic motivations for testing, but this work demonstrates the potential in a very concrete manner.

[41] As an anecdote, we mention that, in course of their research, Goldreich, Goldwasser, and Ron considered the feasibility of testing properties of distributions, but being in the mindset that focused on complexity that is polylogarithmic in the size of the object (see discussion in the history section of the first lecture), they found no appealing example and did not report of these thoughts in their paper [15].

[42] The original results obtained an optimal dependence on n but not on ϵ. Specifically, in these results the complexity is proportional to $\mathrm{poly}(1/\epsilon)$ rather than to $O(1/\epsilon^2)$. Optimal results were obtained in [19, 7, 24].

[43] Recall that the optimal $O(\sqrt{n}/\epsilon^2)$ upper bound was first established by Paninski [19] (for $\epsilon = \Omega(n^{-1/4})$) and then extended in [7] for all $\epsilon > 0$, where both bounds are based on the analysis of a slightly different test. The optimality of this upper bound (i.e., a matching lower bound) was first established in [19] (see alternative proof in [10, Sec. 3.1.1]).

5.2 Exercise

Some of the following exercises are quite educational. We call the reader's attention to Exercise 12, which was not referred to in the main text, that shows that distribution testers can be made deterministic at a minor cost.

Exercise 1 (one-sided testers for properties of distributions): *Suppose that $\mathcal{D}$ is a property of distributions over $[n]$ such that for some* monotone collection of non-empty sets[44] *$\mathcal{C}$ it holds that the distribution X is in $\mathcal{D}$ if and only if the support of X is in $\mathcal{C}$. Prove that $\mathcal{D}$ has a one-sided error tester of sample complexity $O(n/\epsilon)$.* (Observe that if the condition regarding $\mathcal{D}$ does not hold, then there exist a distribution X in $\mathcal{D}$ and a distribution Y not in $\mathcal{D}$ such that the support of Y is a subset of the support of X.)

Guideline: The tester rejects a distribution Y if and only if the multi-set of samples that it sees corresponds to a set that is not in $\mathcal{C}$. Note that if Y is ϵ-far from having a support that equals S, then a sample of $O(n/\epsilon)$ elements drawn from Y will hit a point outside of S with probability at least $1 - \exp(-n)$.

Exercise 2 (on the optimality of the sample complexity asserted in Exercise 1): *Show that there exists a property of distributions $\mathcal{D}$ as in Exercise 1 such that the sample complexity of ϵ-testing $\mathcal{D}$ with one-sided error is $\Omega(n/\epsilon)$.*

Guideline: Consider the set $\mathcal{D}$ of all distributions over $[n]$ such that each distribution in $\mathcal{D}$ has support of size smaller than $n/2$. Note that a one-sided tester may reject a distribution only when it sees at least $n/2$ different elements of $[n]$ in the sample. On the other hand, the distribution p that satisfies $p(n) = 1 - 3\epsilon$ and $p(i) = 3\epsilon/(n-1)$ for all $i \in [n-1]$ is ϵ-far from $\mathcal{D}$.

Exercise 3 (learning via the empirical distribution):[45] *Let $p : [n] \to [0,1]$ be a probability function. Consider an algorithm that on input $m = O(n/\epsilon^2)$ samples, $i_1, ..., i_m \in [n]$, that are drawn according to p, outputs the* empirical distribution *$\widetilde{p}$ defined by letting $\widetilde{p}(i) = |\{j \in [m] : i_j = i\}|/m$ for every $i \in [n]$; that is, $\widetilde{p}$ represents the relative frequency of each of the values $i \in [n]$ in the sample $i_1, ..., i_m$. Using the following steps, prove that, with high probability, $\widetilde{p}$ is ϵ-close to p.*

1. *For every $i \in [n]$, let X_i denote the distribution of the fraction of the number of occurrences of i in the sample. Then, $\mathbb{E}[X_i] = p(i)$ and $\mathbb{V}[X_i] \leq p(i)/m$.*
2. *Show that $\mathbb{E}[|X_i - p(i)|] \leq \mathbb{V}[X_i]^{1/2}$.*
3. *Show that $\mathbb{E}\left[\sum_{i\in[n]} |X_i - p(i)|\right] \leq \sqrt{n/m}$.*

Setting $m = 9n/\epsilon^2$, we get $\mathbb{E}\left[\sum_{i\in[n]} |X_i - p(i)|\right] \leq \epsilon/3$, which implies that $\mathbf{Pr}\left[\sum_{i\in[n]} |X_i - p(i)| > \epsilon\right] < 1/3$.

Guideline: In Step 1, use $\mathbb{V}[m \cdot X_i] = m \cdot p(i) \cdot (1 - p(i))$, since $m \cdot X_i$ is the sum of m independent Bernoulli trials, each having success probability $p(i)$. In Step 2, use $\mathbb{E}[|X_i - p(i)|] \leq \mathbb{E}[|X_i -$

[44] A collection of non-empty subsets $\mathcal{C}$ is called monotone if $S \in \mathcal{C}$ implies that each non-empty subset of S is in $\mathcal{C}$.

[45] This seems to be based on folklore, which was communicated to the author by Ilias Diakonikolas.

$p(i)|^2]^{1/2} = \mathbb{V}[X_i]^{1/2}$, where the inequality is due to $\mathbb{V}[Y] = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \geq 0$ (for any $Y \geq 0$) and the equality uses $p(i) = \mathbb{E}[X_i]$. In Step 3, use

$$\mathbb{E}\left[\sum_{i\in[n]} |X_i - p(i)|\right] = \sum_{i\in[n]} \mathbb{E}[|X_i - p(i)|] \leq \sum_{i\in[n]} \sqrt{p(i)/m},$$

where the last inequality is due to Steps 1-2, and finally justify $\sum_{i\in[n]} \sqrt{p(i)/m} < \sqrt{n/m}$.

Exercise 4 (approximating the $\mathcal{L}_2$-norm of a distribution): *Consider a model in which the algorithm obtains samples on demand; that is, the algorithm is only presented with the parameters n and ϵ, and it obtains an additional sample when asking for it. Hence, the number of samples used by such an algorithm is a random variable, and we consider the distribution of that random variable. Now, for any $\gamma > 0$, using Corollary 6, present an algorithm that when obtaining samples from an unknown distribution p, outputs, with probability at least $2/3$, an $(1+\gamma)$-factor approximation of $\|p\|_2$ while using at most $\widetilde{O}(1/\gamma^2\|p\|_2)$ samples. Furthermore, show that the expected number of samples used by the algorithm is $\widetilde{O}(1/\|p\|_2)$.*

Guideline: The basic idea is to proceed in iterations such that in the i^{th} iteration we check the hypothesis that $\|p\|_2 \approx 2^{-i}$. Hence, in the i^{th} iteration we apply Corollary 6, using $O(2^i)$ samples, and note that the probability that we halt before iteration $\log_2(1/\|p\|_2) - t$ (resp., after iteration $\log_2(1/\|p\|_2) + t$) is $O(2^{-t})$. An alternative solution is implied by an exercise in the previous lecture notes (see "getting rid of a rough estimate").

Exercise 5 (filters that perfectly preserve distances between distributions): *Let $F : [n] \to S$ be a randomized process such that the supports of the different $F(i)$'s are disjoint. Prove that for every two distributions X and X' over $[n]$, the total variation distance between $F(X)$ and $F(X')$ equals the total variation distance between X and X'. Note that distances may not be preserved if the supports of some $F(i)$'s are not disjoint, and that the level of preservation is related to the relation between the distributions of the various $F(i)$'s.*

Guideline: Letting S_i denote the support of $F(i)$, use $\sum_{j\in S} \mathbf{Pr}[F(X) = j] = \sum_{i\in[n]} \sum_{j\in S_i} \mathbf{Pr}[X = i] \cdot \mathbf{Pr}[F(i) = j]$.

Exercise 6 (testing the uniform distribution over $[n]$, extended): *By definition, a tester for the uniform distribution over $[n]$ is supposed to satisfy the conditions of Definition 1 when given an arbitrary distribution over $[n]$; in particular, when given the parameters n and ϵ, the tester is required to reject any distribution over $[n]$ that is ϵ-far from U_n* (the uniform distribution over $[n]$). *Show that any such tester T can be easily adapted to satisfy the rejection requirement also when given an arbitrary distribution, which may have a support that is not a subset of $[n]$.*

Guideline: The adapted tester rejects if the sample contains any element not in $[n]$ and otherwise invokes T on the sample. Provide a rigorous analysis of this tester.[46]

[46] Compare the execution of T, on any input (i.e., n, ϵ and a sequences of samples), to that of the adapted tester, denoted T', and observe that whenever T rejects so does T'.

Exercise 7 (testing uniform distributions, yet another look): *In continuation to Exercise 6, present a 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 $[n]$.*

Guideline: 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 $F(U_m) \equiv U_{2m}$, while

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

Exercise 8 (optimizing the reduction that underlies the proof of Theorem 11):[47] *Optimize the choice of γ in Algorithm 10 so to obtain "optimal" sample complexity in that reduction. Note that the filter of* Eq. (4) *can also be generalized by using a suitable parameter, which can then be optimized.* (Recall that n/γ must be an integer.)

Guideline: Start by generalizing the filter of Eq. (4) by introducing a parameter $\beta \in (0,1)$ and letting $p'(i) = (1-\beta) \cdot p(i) + \beta/n$. Present the complexity of the resulting tester as a function of β and γ (in addition to its dependence on n and ϵ), and minimize this function (by first optimizing the choice of γ for any fixed β).

Exercise 9 (extending the lower bound of Corollary 13): *Show that ϵ-testing the property $\{U_n\}$ requires $\Omega(\min(n^{2/3}, \epsilon^{-2}\sqrt{n}))$ samples.*

Guideline: Note that, with probability $1-(s^3/n^2)$, a sequence of s samples that are drawn from the uniform distribution on $[n]$ contains no three-way collisions (i.e., $c_j = 0$ for all $j > 2$).[48] But this happens, with similar probability, also when the distribution assigns probability either $(1-2\epsilon)/n$ or $(1+2\epsilon)/n$ to each element. Hence, if $\sqrt{n}/\epsilon^2 < n^{2/3}$, then $\Omega(\sqrt{n}/\epsilon^2)$ samples are required in order to tell the two distributions apart.[49]

Exercise 10 (upper bounds on the length of approximate histogram): *Recall that Theorem 27 implies that every distribution $p : [n] \to [0,1]$ is ϵ-close to a distribution that has a histogram of length $O(\epsilon^{-2} \cdot n/\log n)$. Provide a direct proof of this fact by proving that p is ϵ-close to a distribution that has a histogram of length $O(\epsilon^{-1} \cdot \log(n/\epsilon))$.*

[47] We do not consider such an optimization important, but it may serve as a good exercise.

[48] Recall that c_j denotes the number of elements that occur j times in the sequence of samples $(i_1, ..., i_s)$; that is, $c_j = |\{i \in [n] : \#_i(i_1, ..., i_s) = j\}|$, where $\#_i(i_1, ..., i_s) = |\{k \in [s] : i_k = i\}|$.

[49] Note that the collision probability of the second distribution equals $(1 + \Theta(\epsilon^2))/n$.

Guideline: First, modify p into p' such that p' is uniform on all i's that have probability at most $\epsilon/2n$ in p (i.e., $p'(i) = p(i)$ if $p(i) > \epsilon/2n$ and $p'(i_i) = p'(i_2)$ for every i_1, i_2 that satisfy $p(i_1), p(i_2) \leq \epsilon/2n$). Next, partition the remaining i's into buckets B_j's such that $B_j = \{i : (1+0.5\epsilon)^{j-1} \cdot \epsilon/2n < p(i) \leq (1+0.5\epsilon)^j \cdot \epsilon/2n\}$, and modify p' such that it is uniform on the i's in each B_j.

Exercise 11 (reduction among testing grained properties): *For every $m_1 < m_2$, present a reduction of the task of estimating the distance to m_1-grained distributions over $[n]$ to estimating the distance to m_2-grained distributions over $[n]$. Specifically, present a filter that maps m_1-grained distributions to m_2-grained distributions such that the filter preserved the distance between distributions up to a fixed scaling* (of m_1/m_2).

Guideline: For starters, consider the filter F'_{m_1,m_2} that maps $i \in [n]$ with itself with probability m_1/m_2 and maps it to $n+1$ otherwise. Then, consider the filter F_{m_1,m_2} that maps the excessive probability mass (of $(m_2 - m_1)/m_2$) to n (rather than to $n+1$).

Exercise 12 (distribution testers can be made deterministic at a minor cost): *Let $\mathcal{D}$ be a property of distributions over $[n]$. Show that if $\mathcal{D}$ can be tested in sample complexity $s(n,\epsilon)$, then it can be tested by a deterministic machine of sample complexity $3s(n,\epsilon) + O(\epsilon^{-1} \cdot (\log s(n,\epsilon) + \log\log n))$.* (The factor of 3 increase in the sample complexity is due to the desire to maintain the same error bound, and it can be avoided if one is willing to increase the error probability from 1/3 to, say, 0.35.)

Guideline: First reduce the randomness complexity of the randomized tester by using ideas as in an exercise in the first lecture, obtaining a tester of randomness complexity $\log s(n,\epsilon) + \log\log n$ that has error probability at most 0.34 (rather than at most 1/3). This is done by considering all n^s possible s-long sequences of samples, and picking a set of $s \cdot \log n + O(1)$ random pads that approximate the behavior of the tester (on all possible sample sequences). Next, present a deterministic tester that emulates the execution of a randomized tester that uses s samples and r random coins, by using $O(r/\epsilon)$ additional samples. The idea is to partition these additional samples into pairs and try to extract a random bit from each pair (x,y) such that the bit is 1 (resp., 0) if $x < y$ (resp., if $x > y$), where in case $x = y$ no bit is extracted. For some suitable constant c, suppose that we have $c \cdot r/\epsilon$ such pairs, and consider the following three (somewhat overlapping) cases.

1. The typical case is that at least r random bits were extracted. In this case, we just emulate the randomized tester.

2. A pathological case, which arises when the tested distribution X is concentrated on one value $i \in [n]$ (i.e., $\mathbf{Pr}[X=i] > 1-\epsilon/2$), is that a majority of the pairs equal (i,i) for some $i \in [n]$. In this case we accept if and only if the distribution X' that is identically i is $\epsilon/2$-close to $\mathcal{D}$.

3. An extremely rare case is that less than r bits were extracted but no pair (i,i) appears in majority. This case is extremely unlikely, and it does not matter what we do.

The analysis refers to two overlapping cases regarding X. On the one hand, if the tested distribution X satisfies $\mathbf{Pr}[X=i] < 1-\epsilon/4$ for all $i \in [n]$, then Case 1 occurs with very high probability. On the other hand, if there exists $i \in [n]$ such that $\mathbf{Pr}[X=i] > 1-\epsilon/2$, then with very high probability either Case 1 or Case 2 occurs; in this case (where X is $\epsilon/2$-close to X'), if X is in $\mathcal{D}$ (resp., X is ϵ-far from $\mathcal{D}$), then X' is $\epsilon/2$-close to $\mathcal{D}$ (resp., $\epsilon/2$-far from $\mathcal{D}$).

Exercise 13 (on the algebra of distribution testing):[50] *Let $\mathcal{D}'$ and $\mathcal{D}''$ be properties of distributions over $[n]$ that are each testable within sample complexity s.*

1. *Show that $\mathcal{D}' \cup \mathcal{D}''$ is testable within sample complexity $O(s)$.*
2. *Show that the sample-complexity of testing $\mathcal{D}' \cap \mathcal{D}''$ may be $\Omega(n/\log n)$ even if $s = O(1/\epsilon)$ and each of the properties is label-invariant* (and contains only distributions in which each element in their support appears with probability at least $1/n$).

Guideline: Part 1 can be proven as the corresponding result in the first lecture. To prove Part 2, start with any of the properties $\mathcal{D}$ of Corollary 30, and consider the class $\mathcal{C}$ of all distributions over $[n]$ that assign each element in their support probability at least $1/n$. Let $\mathcal{D}'$ (resp., $\mathcal{D}''$) consist of $\mathcal{D}$ as well as of all distributions in $\mathcal{C}$ that have a support of even (resp., odd) size. Then, each distribution over $[n]$ is $1/n$-close to $\mathcal{D}'$ (resp., $\mathcal{D}''$), whereas $\mathcal{D}' \cap \mathcal{D}'' = \mathcal{D}$.

References

[1] J. Acharya, C. Daskalakis, and G. Kamath. Optimal Testing for Properties of Distributions. `arXiv:1507.05952 [cs.DS]`, 2015.

[2] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, Vol. 35 (1), pages 132–150, 2005

[3] 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.

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

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

[6] C.L. Canonne, D. Ron, and R. Servedio. Testing probability distributions using conditional samples. *SIAM Journal on Computing*, Vol. 44 (3), pages 540–616, 2015.

[7] 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.

[8] C. Daskalakis, I. Diakonikolas, and R.A. Servedio. Learning k-Modal Distributions via Testing. *Theory of Computing*, Vol. 10, pages 535–570, 2014.

[9] L. Devroye. The Equivalence of Weak, Strong and Complete Convergence in L1 for Kernel Density Estimates. *Annals of Statistics*, Vol. 11 (3), pages 896–904, 1983.

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

[50] See analogous section on the algebra of testing properties of functions (in the first lecture).

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

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

[13] O. Goldreich (ed.). *Property Testing: Current Research and Surveys*. Springer, LNCS, Vol. 6390, 2010.

[14] O. Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. *ECCC*, TR16-015, February 2016.

[15] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of the ACM*, pages 653–750, July 1998. Extended abstract in *37th FOCS*, 1996.

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

[17] O. Goldreich and D. Ron. On Sample-Based Testers. In *6th Innovations in Theoretical Computer Science*, pages 337–345, 2015.

[18] 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`

[19] 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.

[20] M. Parnas, D. Ron, and R. Rubinfeld. Tolerant property testing and distance approximation. *Journal of Computer and System Science*, Vol. 72(6), pages 1012–1042, 2006.

[21] M. Sudan. Invariances in Property Testing. In [13].

[22] G.J. Valiant. Algorithmic Approaches to Statistical Questions. PhD Thesis, University of California at Berkeley, 2012.

[23] 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. Full version in *ECCC*, TR10-179 and TR10-180, 2010.

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

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

[26] P. Valiant. Testing Symmetric Properties of Distributions, PhD Thesis, MIT, 2012.