Testing random variables for independence and identity[*]

Tuğkan Batu[†] Eldar Fischer[‡] Lance Fortnow[§] Ravi Kumar[¶] Ronitt Rubinfeld[‡]
Patrick White[‖]

January 10, 2003

Abstract

Given access to independent samples of a distribution A over $[n] \times [m]$, we show how to test whether the distributions formed by projecting A to each coordinate are independent, i.e., whether A is ϵ-close in the L_1 norm to the product distribution $A_1 \times A_2$ for some distributions A_1 over $[n]$ and A_2 over $[m]$. The sample complexity of our test is $\tilde{O}(n^{2/3}m^{1/3}\text{poly}(\epsilon^{-1}))$, assuming without loss of generality that $m \leq n$. We also give a matching lower bound, up to $\text{poly}(\log n, \epsilon^{-1})$ factors.

Furthermore, given access to samples of a distribution X over $[n]$, we show how to test if X is ϵ-close in L_1 norm to an explicitly specified distribution Y. Our test uses $\tilde{O}(n^{1/2}\text{poly}(\epsilon^{-1}))$ samples, which nearly matches the known tight bounds for the case when Y is uniform.

[*] A preliminary version of this paper appeared in the 42nd Symposium on Foundations of Computer Science, 2002, Las Vegas, NV.

[†] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA. This work was partially supported by ONR N00014-97-1-0505. `batu@saul.cis.upenn.edu`.

[‡] NEC Research Institute, Princeton, NJ. Current Address: Computer Science Department, Technion, Technion City, Haifa, Israel. `eldar@cs.technion.ac.il`.

[§] NEC Research Institute, Princeton, NJ. {`fortnow, ronitt`}`@research.nj.nec.com`.

[¶] IBM Almaden Research Center, San Jose, CA. `ravi@almaden.ibm.com`.

[‖] Department of Computer Science, Cornell University, Ithaca, NY. This work was partially supported by ONR N00014-97-1-0505. `white@cs.cornell.edu`.

1 Introduction

Fred works at a national consumer affairs office, where each day he gets several consumer complaints. Because he has a hunch that there is some correlation between the zip code of the consumer and the zip code of the company, Fred wants to check whether these zip codes are dependent. However, since there are 10^{10} zip code pairs, he does not have enough samples for traditional statistical techniques. What can Fred do?

Suppose we are given a black box that generates independent samples of a distribution A over pairs (i, j) for $i \in [n]$ and $j \in [m]$ with $m \leq n$. We want to test whether the distribution over the first elements is independent of the distribution over the second elements, without making any additional assumptions on the structure of A.

Checking independence is a central question in statistics and there exist many different techniques for attacking it (see [5]). Classical tests such as the χ^2 test or the Kolmogorov-Smirnoff test work well when n and m are small, but for large n, m these tests require more than $n \cdot m$ samples, which may be huge. Can one develop a test that uses fewer than nm samples?

We also consider the problem of testing if a black-box distribution over $[n]$ is close to a known distribution. The χ^2 test is commonly used for this problem, but requires at least a linear number of samples. Can one develop a test that uses a sublinear number of samples?

Testing statistical properties of distributions has been studied in the context of property testing [7, 3] (see the survey by Ron [6]). Using the techniques of Goldreich and Ron [4], one can get (see [1]) an $\tilde{O}(\sqrt{r})$ algorithm to test if a black-box distribution over r elements is close in L_1 norm to uniform. Batu, Fortnow, Rubinfeld, Smith, and White [1] show how to test whether two black-box distributions over r elements are close in L_1 norm, using $\tilde{O}(r^{2/3})$ samples. In particular, this gives a test that answers the second question in the affirmative.

Our results. In this paper we develop a general algorithm (Section 3) for the independence testing problem with a sublinear sample complexity (in the size of $[n] \times [m]$). To our knowledge, this is the first sublinear time test which makes no assumptions about the structure of the distribution. Our test uses $O(n^{2/3}m^{1/3}\text{poly}(\log n, \epsilon^{-1}))$ samples, assuming without loss of generality that $n \geq m$, and distinguishes between the case that $A = A_1 \times A_2$, and the case that for all A_1 and A_2, $|A - A_1 \times A_2| \geq \epsilon$. Here, A_1 and A_2 are distributions over $[n]$ and $[m]$ respectively and $|A - B|$ represents the L_1 or statistical difference between two distributions. We also show that this bound is tight up to $\text{poly}(\log n, \epsilon^{-1})$ factors (Sectin 5).

We then give an algorithm (Section 4) to test if a black-box distribution over $[n]$ is close (in L_1 norm) to an explicitly specified distribution. Our algorithm uses $O(n^{1/2}\text{poly}(\log n, \epsilon^{-1}))$ samples – almost matching the upper and lower bounds of [4] for the uniform case.

Overview of our techniques. Our approach begins with presenting two different ways of testing independence of distributions. These two methods have different sample complexities and are desirable in different situations.

In the first method, we use the equivalence of testing independence to testing whether A is close to $\pi_1 A \times \pi_2 A$ where $\pi_i A$ is the distribution of A projected to the i-th coordinate. Since it is easy to generate samples of $\pi_1 A \times \pi_2 A$ given samples of A, we can apply the result of Batu et al. [1]. This immediately gives us a test for independence that uses $\tilde{O}(n^{2/3}m^{2/3})$ samples.

For the second method, first consider the case where $\pi_1 A$ and $\pi_2 A$ are uniform over $[n]$ and $[m]$, respectively. Testing the independence is equivalent to testing whether A is uniform in $[n] \times [m]$. We can use the test of Goldreich and Ron [4] for this using $\tilde{O}(\sqrt{nm}) = \tilde{O}(n)$ samples.

To reduce the general problem to that of $\pi_1 A$ and $\pi_2 A$ uniform we first use a *bucketing* technique (Section 2.3) to partition $[n]$ and $[m]$ into a polylogarithmic number of buckets of elements of similar probabilities given $\pi_1 A$ and $\pi_2 A$, respectively. To do this we must approximate the probabilities of each $\pi_1 A(i)$ and $\pi_2 A(j)$ which requires $\tilde{O}(\max(n,m)) = \tilde{O}(n)$ samples.

For all buckets $B_1 \subseteq [n]$ and $B_2 \subseteq [m]$ we could try to test independence of A restricted to $B_1 \times B_2$ since $\pi_1 A$ restricted to B_1 and $\pi_2 A$ restricted to B_2 are close to uniform. Unfortunately they are not close enough to uniform for our purposes. To overcome this we use a *sifting* technique (Section 2.4). We first collect many samples (i,j) with $i \in B_1$ and $j \in B_2$. We then create a virtual sampler that first chooses i uniformly from B_1 and then picks the first (i,j) we previously sampled. We then create a second virtual sampler that chooses j uniformly from B_2 and picks the first (i,j) from the first virtual sampler. We show that the second virtual sampler, which we call a *sieve*, preserves the dependencies of the original distribution, and has uniform projections to B_1 and B_2, so we can apply the uniform tester described above. We also show that this process only costs us a polylogarithmic factor in the number of samples we need, achieving a tester using $\tilde{O}(n)$ samples overall.

Then, we combine these two algorithms in an appropriate manner to exploit their different behavior. In particular, we partition the elements of $[n]$ to 'light' and 'heavy' elements based on $\pi_1 A$. We apply the first method to the light elements, and apply the second method to the heavy elements. This asymmetric approach helps us achieve an optimal trade-off in the sample complexities, resulting in the $\tilde{O}(n^{2/3}m^{1/3})$ bound.

2 Some preliminary tools

We use the $\tilde{O}$ notation to hide dependencies on the logarithm of any of the quantities in the expression, i.e., $f = \tilde{O}(g)$ if $f = O(g \cdot \mathrm{poly}(\log g))$. To simplify the exposition, we assume that all tests are repeated so that the confidence is sufficiently high. Since amplifying the confidence to $1-\delta$ can be achieved with $O(\log \frac{1}{\delta})$ trials, an additional multiplicative factor that is polylogarithmic in n or $n^{2/3}m^{1/3}$ (as the case may be) is all that we will require.

We use X, Y, Z to denote random variables over sets and A, B, C, D to denote random variables over pairs of sets. We often refer to the first coordinate of a sample from the latter type of distributions as the *prefix*.

For a set R, let U_R denote the uniform random variable over R. Let $X(i)$ denote the probability that $X = i$, and for a subset R' of R, let $X(R') \stackrel{\mathrm{def}}{=} \sum_{i \in R'} X(i)$. If A is a random variable over $S \times T$, let $\pi_i A$ denote the random variable obtained by projecting A into the i-th coordinate. Let $A(i,j)$ denote the probability that $A = (i,j)$.

Let $|\cdot|$ stand for the L_1 norm, $\|\cdot\|$ for the L_2 norm, and $\|\cdot\|_\infty$ for the L_∞ norm. If $|A - B| \le \epsilon$, we say that A is *ϵ-close* to B.

We assume that a distribution X over R can be specified in one of two ways. We call X a *black-box* distribution if an algorithm can only get independent samples from X and otherwise has no knowledge about X. We call X an *explicit* distribution if it is represented by an oracle which on input $i \in R$ outputs the probability mass $X(i)$.

2.1 Independence and approximate independence

Let A be a distribution over $[n] \times [m]$. We say that A is *independent* if the induced distributions $\pi_1 A$ and $\pi_2 A$ are independent, i.e., that $A = (\pi_1 A) \times (\pi_2 A)$. Equivalently, A is independent if for all $i \in [n]$ and

$j \in [m]$, $A(i,j) = (\pi_1 A)(i) \cdot (\pi_2 A)(j)$.

We say that A is *ϵ-independent* if there is a distribution B that is independent and $|A - B| \leq \epsilon$. Otherwise, we say A is *not ϵ-independent* or is *ϵ-far from being independent*.

Now, closeness is preserved under independence:

Proposition 1 *Let A, B be distributions over $S \times T$. If $|A - B| \leq \epsilon/3$ and B is independent, then $|A - (\pi_1 A) \times (\pi_2 A)| \leq \epsilon$.*

Proposition 1 follows from the following lemmas.

Lemma 2 ([8]) *Let X_1, Y_1 be distributions over S and X_2, Y_2 be distributions over T. Then $|X_1 \times Y_1 - X_2 \times Y_2| \leq |X_1 - Y_1| + |X_2 - Y_2|$.*

Lemma 3 *Let A, B be distributions over $S \times T$. If $|A - B| \leq \epsilon$, then $|\pi_1 A - \pi_1 B| \leq \epsilon$ and $|\pi_2 A - \pi_2 B| \leq \epsilon$.*

PROOF OF PROPOSITION 1: Clearly, $B = (\pi_1 B) \times (\pi_2 B)$. Using the triangle inequality, Lemma 2 and Lemma 3, $|A - (\pi_1 A) \times (\pi_2 A)| \leq |A - B| + |B - (\pi_1 A) \times (\pi_2 A)| = |A - B| + |(\pi_1 B) \times (\pi_2 B) - (\pi_1 A) \times (\pi_2 A)| \leq \epsilon/3 + 2\epsilon/3 = \epsilon$. ■

2.2 Restriction and coarsening

We begin with the definitions.

Definition 4 *Given a random variable X over R, and $\emptyset \subset R' \subseteq R$, the* restriction *$X_{|R'}$ is the random variable over R' with distribution $X_{|R'}(i) = X(i)/X(R')$.*

Given a random variable X over R, and a partition $\mathcal{R} = \{R_1, \ldots, R_k\}$ of R, the coarsening *$X_{\langle \mathcal{R} \rangle}$ is the random variable over $[k]$ with distribution $X_{\langle \mathcal{R} \rangle}(i) = X(R_i)$.*

The definition of restriction resembles the definition of a conditional distribution, only a restriction is defined as a distribution over the subset R', while a conditional distribution is defined over the whole R by padding it with zeros.

In the light of the above definition, it follows that:

Observation 5 *If X is a random variable over R and $\mathcal{R} = \{R_1, \ldots, R_k\}$ is a partition of R, then for all i in $[k]$ and j in R_i, $X(j) = X_{\langle \mathcal{R} \rangle}(i) \cdot X_{|R_i}(j)$.*

In words, the probability of picking an element j belonging to the partition R_i according to X is equivalent to the probability of picking the partition R_i times the probability of picking j when restricted to the partition R_i. Using Observation 5, it follows that $A(i,j) = (\pi_1 A)(i) \cdot (\pi_2 A_{|\{i\} \times [m]})(j)$.

The following lemma (proof omitted) shows that two random variables are close if they are close with respect to restrictions and coarsening.

Lemma 6 *Let X, Y be random variables over R and let $\mathcal{R} = \{R_1, \ldots, R_k\}$ be a partition of R. If for all i in $[k]$, $|X_{|R_i} - Y_{|R_i}| \leq \epsilon_1$ and $|X_{\langle \mathcal{R} \rangle} - Y_{\langle \mathcal{R} \rangle}| \leq \epsilon_2$, then $|X - Y| \leq \epsilon_1 + \epsilon_2$.*

Note that if $(1-\epsilon)X(R_i) \leq Y(R_i) \leq (1+\epsilon)X(R_i)$ for every $i \in [k]$, then $|X_{\langle \mathcal{R} \rangle} - Y_{\langle \mathcal{R} \rangle}| \leq \epsilon$.

The following lemma (proof omitted) shows a partial converse: if X and Y are close, then they are close when restricted to sufficiently 'heavy' partitions of the domain.

Lemma 7 *Let X, Y be distributions over R and let $R' \subseteq R$. Then $|X_{|R'} - Y_{|R'}| \leq 2|X - Y|/X(R')$.*

2.3 Bucketing

Bucketing is a general tool that decomposes an arbitrary explicitly given distribution into a collection of distributions that are almost uniform.

Given an explicit distribution X over R, we define *Bucket*(X, R, ϵ) as a partition $\{R_0, R_1, \ldots, R_k\}$ of R with $k = (2/\log(1+\epsilon)) \cdot \log|R|$, $R_0 = \{i \mid X(i) < 1/(|R|\log|R|)\}$, and for all i in $[k]$,

$$R_i = \left\{ j \;\middle|\; \frac{(1+\epsilon)^{i-1}}{|R|\log|R|} \le X(j) < \frac{(1+\epsilon)^i}{|R|\log|R|} \right\}.$$

The following lemma shows that if one considers the restriction of X to any of the buckets R_i, then the distribution is close to uniform.

Lemma 8 *Let X be an explicit distribution over R and let $\{R_0, \ldots, R_k\} =$ Bucket(X, R, ϵ). For $i \in [k]$ we have $|X_{|R_i} - U_{R_i}| \le \epsilon$, $\|X_{|R_i} - U_{R_i}\|^2 \le \epsilon^2/|R_i|$, and $X(R_0) \le 1/\log|R|$.*

PROOF: Clearly, $X(R_0) \le 1/\log|R|$. For $i \ge 1$, consider an arbitrary (non-empty) subset R_i and without loss of generality, assume $R_i = \{1, \ldots, \ell\}$ with $X(1) \le \cdots \le X(\ell)$. Let $Y = X_{|R_i}$. Then, $Y(\ell)/Y(1) < 1+\epsilon$. Also, by averaging, $Y(1) \le 1/\ell \le Y(\ell)$. Hence $Y(\ell) \le (1+\epsilon)Y(1) \le (1+\epsilon)/\ell$. Similarly it can be shown that $Y(1) \ge 1/(\ell(1+\epsilon)) > (1-\epsilon)/\ell$. Thus, it follows that $|Y(j) - 1/\ell| \le \epsilon/\ell$ for all $j = 1, \ldots, \ell$ and therefore, $\sum_{j \in R_i} |Y(j) - U_{R_i}| \le \epsilon$ and $\sum_{j \in R_i} (Y(j) - U_{R_i})^2 \le \epsilon^2/\ell$. ■

Given an "approximation" $\tilde{X}$ of X, the bucketing of $\tilde{X}$ has similar properties as the bucketing of X.

Corollary 9 *Let $X, \tilde{X}$ be distributions over R such that $\tilde{X}$ approximates X i.e., $\forall i \in R, (1-\epsilon)X(i) \le \tilde{X}(i) \le (1+\epsilon)X(i)$ for some $\epsilon > 0$. Then,* Bucket$(\tilde{X}, R, \epsilon)$ *is a partition $\{R_0, \ldots, R_k\}$ of R with $k = O(\epsilon^{-1}\log|R|)$ such that for all $i \in [k]$, $|X_{|R_i} - U_{R_i}| \le 3\epsilon$ and $X(R_0) \le (1+\epsilon)/\log|R|$.*

In our applications of bucketing, we usually ignore the 0-th bucket R_0 since the probability mass on this bucket will be negligible for our purposes.

2.4 Sieves

In our construction, we will have a distribution A whose projections are nearly uniform. By using techniques from Goldreich and Ron [4], we can quickly test independence of distributions whose projections are truly uniform. In this section, we show how to reduce the nearly uniform case to the uniform case. We achieve this reduction by constructing a *sieve* that collects samples from A and sifts out samples in a way that achieves the desired uniformity.

We use the sieve in batch mode, i.e., given an input parameter t, we output t samples according to some B, based on samples we take of the original A. An (A, B)-sieve is specified in terms of the relationship between the properties of the output distribution B and those of input distribution A; in our case of an A whose projections are nearly uniform the sieve will produce a B that is close to A, while uniformizing its projections, and preserving its independence if such existed. The other important parameter is the sample complexity of the sieve, which is the total number of samples of A it uses for a given t and the domain of A.

We first show that there is a sieve that takes a distribution A over $S \times T$ for which the first coordinate is close to uniform, and produces a new distribution which is close to the original one, and for which the first coordinate is uniform; moreover, this sieve preserves independence.

Lemma 10 *There exists an (A, B)-sieve for random variables over $S \times T$ such that for any t, with high probability, (1) if $A = (\pi_1 A) \times (\pi_2 A)$ then $B = U_S \times (\pi_2 A)$, and (2) if $|\pi_1 A - U_S| \le \epsilon/4$ then $|A - B| \le \epsilon$. The sample complexity of the sieve is $O(\max\{|S|, t\} \log^3 \max\{|S|, t\})$.*

PROOF: First, we describe the construction of the sieve. Let t be given and let $\ell = O(t/|S| \log |S| \log t)$. The sieve maintains a data structure which for every $i \in S$, contains a list L_i of ℓ elements of T. Each list starts out empty and is filled according to the following steps:

(1) Obtain $O(\max\{|S|, t\} \log^3 \max\{|S|, t\})$ samples from A and for each sample (i, j) from A, add j to L_i if $|L_i| \le \ell$.

(2) For each $i' \in S$, if $|L_{i'}| < \ell$, then discard $L_{i'}$. In this case, obtain ℓ more samples from A and for each sample (i, j) from A, add j to $L_{i'}$.

For $i \in S$, let B_i be a random variable with the distribution $\pi_2 A_{|\{i\} \times T}$ if L_i was not discarded in step (2) and with the distribution $\pi_2 A$ otherwise. Thus, L_i contains ℓ independent samples of B_i.

Next, we describe the operation of the sieve. Upon a sample request, the sieve generates a uniformly random $i \in_R S$. If $|L_i| > 0$, then the sieve picks the first element j in L_i, outputs (i, j), and deletes the first element in L_i. If $|L_i| = 0$, then the sieve gets a sample (i', j') from A and outputs (i, j').

First, notice that with high probability (via a Chernoff bound), no L_i becomes empty in any of the t requests for samples. Also, it is clear that the output of the sieve is the random variable B defined by generating a uniform $i \in_R S$ and then simulating the corresponding B_i. The exact distribution of B may depend on the outcome of the preprocessing stage of the sieve, but we show that with high probability B satisfies the assertions of the lemma.

For the first assertion, note that if $A = (\pi_1 A) \times (\pi_2 A)$, then the second coordinate is independent of the first coordinate. So, $B_i = \pi_2 A$ for every i (regardless of whether L_i was filled by step (1) or (2)). Thus, $B = U_S \times (\pi_2 A)$.

To show the second assertion, let $I = \{i \mid \pi_1 A(i) \ge 1/(2|S|)\}$. Another application of the Chernoff bound shows that with high probability, for every $i \in I$, B_i is distributed as $\pi_2(A_{|\{i\} \times T})$ (since L_i would not be discarded in step (2)). Thus, for every $i \in I$, L_i contains ℓ independent samples of $B_i = \pi_2 A_{|\{i\} \times T}$. Also, since $|\pi_1(A) - U_S| \le \epsilon/4$, we have $|S \backslash I| \le \epsilon |S|/2$. We get

$$\begin{aligned}
|A - B| &= \textstyle\sum_{i \in I} \sum_{j \in T} |A(i,j) - B(i,j)| \\
&\quad + \textstyle\sum_{i \in S \backslash I} \sum_{j \in T} |A(i,j) - B(i,j)| \\
&\le \textstyle\sum_{i \in I} \sum_{j \in T} |A(i,j) - B(i,j)| \\
&\quad + \textstyle\sum_{i \in S \backslash I} \sum_{j \in T} (A(i,j) + B(i,j)) \\
&= \textstyle\sum_{i \in I} \sum_{j \in T} \pi_2 A_{|\{i\} \times T}(j) \cdot \left| \pi_1 A(i) - \frac{1}{|S|} \right| \\
&\quad + \textstyle\sum_{i \in S \backslash I} (\pi_1 A(i) + \pi_1 B(i)) \\
&\le \textstyle\sum_{i \in I} \left| \pi_1 A(i) - \frac{1}{|S|} \right| + \sum_{i \in S \backslash I} \pi_1 A(i) + \frac{|S \backslash I|}{|S|} \\
&\le \tfrac{1}{4}\epsilon + \tfrac{1}{4}\epsilon + \tfrac{1}{2}\epsilon = \epsilon
\end{aligned}$$

■

Sieves can be composed, i.e., an (A, C)-sieve can be combined with a (C, B)-sieve to give an (A, B)-sieve. If the sample complexity of the (A, C)-sieve is given by the function $f(t)$, and that of the (C, B)-sieve is given by $g(t)$, then the sample complexity of the combined (A, B)-sieve will be given by $h(t) = f(g(t))$.

Corollary 11 *There exists an (A, B)-sieve for random variables over $S \times T$ such that if $|\pi_1 A - U_S| \le \epsilon/25$, and $|\pi_2 A - U_T| \le \epsilon/25$, then with high probability, (1) $|B - A| \le (24/25)\epsilon$; (2) if $A = (\pi_1 A) \times (\pi_2 A)$ then $B = U_{S \times T}$; and (3) if A is not ϵ-independent, then $|B - U_{S \times T}| \ge (1/25)\epsilon$. The sample complexity of the sieve is $O(\max\{|S| + |T|, t\} \log^6 \max\{|S|, |T|, t\})$.*

PROOF: We apply the (A, C)-sieve from Lemma 10 on the first coordinate. Using this sieve we obtain a random variable C (with high probability) such that $|C - A| \leq 4\epsilon/25$, $\pi_1 C = U_S$, and such that C is independent if A is independent. Now, using Lemma 3, $|\pi_2 C - \pi_2 A| \leq 4\epsilon/25$ and since by our hypothesis, $|\pi_2 A - U_T| \leq \epsilon/25$, we get $|\pi_2 C - U_T| \leq \epsilon/5$.

We now construct a (C, B)-sieve from Lemma 10, only this time switching coordinates and sifting on the second coordinate. Using this sieve, we obtain a random variable B (with high probability) such that $|B - C| \leq 20\epsilon/25$ and $\pi_2 B = U_T$.

Moreover, according to Lemma 10 if A is independent (and thus so are C and B) then $\pi_1 B$ has the same distribution as $\pi_1 C = U_S$. Since $\pi_1 B = U_S, \pi_2 B = U_T$ and they are independent, we get that B is uniform on $S \times T$.

Clearly, $|B - A| \leq |B - C| + |C - A| \leq (24/25)\epsilon$. This implies that if A is not ϵ-independent, then B is $(1/25)\epsilon$-far from any independent distribution on $S \times T$, and in particular from $U_{S\times T}$. ■

2.5 Tools from earlier work

We use the following results from earlier work. The first theorem states that the L_2 norm of a distribution can be approximated in sublinear time. This can be proved using techniques from Goldreich and Ron [4].

Theorem 12 (based on [4]) *Given a black-box distribution X over R, there is a test using $O(\sqrt{|R|}\epsilon^{-2}\log(1/\delta))$ queries that estimates $\|X\|^2$ to within a factor of $(1 \pm \epsilon)$, with probability at least $1 - \delta$.*

The next theorem states that there is a sublinear time test for L_2 closeness of two black-box distributions.

Theorem 13 ([1]) *Given two black-box distributions X, Y over R, there is a test requiring $O(\epsilon^{-4}\log(1/\delta))$ samples which (1) if $\|X - Y\| \leq \epsilon/2$ it outputs PASS with probability at least $1 - \delta$ and (2) if $\|X - Y\| \geq \epsilon$ it outputs FAIL with probability at least $1 - \delta$.*

The next theorem states that there is a sublinear time test for L_1 closeness of two black-box distributions.

Theorem 14 ([1]) *Given two black-box distributions X and Y over R, there exists a test that requires $O(|R|^{2/3}\epsilon^{-4}\log|R|\log(1/\delta))$ samples which (1) if $|X - Y| \leq \max(\epsilon^2/(32\sqrt[3]{|R|}), \epsilon/(4\sqrt{|R|}))$, it outputs PASS with probability at least $1 - \delta$ and (2) if $|X - Y| > \epsilon$, it outputs FAIL with probability at least $1 - \delta$.*

The next theorem improves this result in the case that the L_∞ norms of the distributions are sufficiently small.

Theorem 15 ([1]) *Given two black-box distributions X, Y over R, with $\|X\|_\infty \leq \|Y\|_\infty$, there is a test requiring $O((|R|^2\|X\|_\infty\|Y\|_\infty\epsilon^{-4} + \sqrt{|R|}\|X\|_\infty\epsilon^{-2})\log(1/\delta))$ samples that (1) if $|X - Y| \leq \frac{\epsilon^2}{\sqrt[3]{|R|}}$, it outputs PASS with probability at least $1 - \delta$ and (2) if $|X - Y| > \epsilon$, it outputs FAIL with probability at least $1 - \delta$.*

The following theorem states that all sufficiently large entries of a probability vector can be estimated efficiently.

Theorem 16 *Given a black-box distribution X over R, a threshold t and an accuracy $\epsilon > 0$, there is an algorithm that requires $O(t^{-1}\epsilon^{-2}\log|R|\log(1/\delta))$ samples and outputs an estimate $\tilde{X}$ such that with probability at least $1 - \delta$, for every $i \in R$ with $X(i) \geq t$ we have $(1 - \epsilon)X(i) \leq \tilde{X}(i) \leq (1 + \epsilon)X(i)$; the algorithm also outputs a set $R' \subseteq R$ that includes $\{i \in R \mid X(i) \geq t\}$ and on which the above approximation is guaranteed.*

The proof (omitted) of the above theorem is a simple application of a Chernoff bound to several independent samples from X. Finally, by similar methods to Theorem 15 (in conjunction with those of [4]), we can show the following (proof omitted):

Theorem 17 *Given a black-box distribution X over R, there is a test that takes $O(\epsilon^{-4}\sqrt{|R|}\log(|R|)\log(1/\delta))$ samples, outputs PASS with probability at least $1-\delta$ if $X = U_R$, and outputs FAIL with probability at least $1-\delta$ if $|X - U_R| > \epsilon$.*

3 Testing independence

In this section we give a test for independence of a distribution A over $[n] \times [m]$. Without loss of generality, let $n \geq m$. The basic steps are the following. We partition $[n]$ into "heavy" and "light" prefixes while estimating the probabilities of the heavy prefixes explicitly. We then apply different approaches for each of these two classes: For the distribution restricted to the heavy prefixes, we use bucketing and sifting to transform the distribution into one that is easier to test for independence (Section 3.1). For the light prefixes, we use a different bucketing and previous results that allow one to test that such distributions are close in the L_1 distance (Section 3.2). Finally we ensure that the distributions restricted to the the different prefixes are consistent.

Let ϵ be given and let $m = n^{\beta}$. Let $0 < \alpha < 1$ be a parameter to be determined later. Let S' denote the set of indices in the first coordinate with probability mass at least $n^{-\alpha}$, which we will also refer to as the heavy prefixes. Formally, let $S' = \{i \in [n] \mid (\pi_1 A)(i) \geq n^{-\alpha}\}$. Similarly, we also define: $S'' = \{i \in [n] \mid (\pi_1 A)(i) \geq \frac{1}{2}n^{-\alpha}\}$. Using a total of $O(n^{\alpha}\epsilon^{-2}\log n)$ samples, we can estimate $(\pi_1 A)(i), i \in S''$ by $\tilde{A}_1(i)$ to within an $\epsilon/75$ factor using Theorem 16. Let $\tilde{S}$ be the set of all i for which $\tilde{A}_1(i) \geq \frac{2}{3}n^{-\alpha}$. Then, $\tilde{S} \supset S'$, and moreover $\tilde{S}$ does not contain any i for which $(\pi_1 A)(i) \leq n^{-\alpha}/2$.

Our main idea is to first test that A is independent conditioned on the set of heavy prefixes and then to test that A is independent conditioned on the set of light prefixes. To create these conditionings, we first distinguish (using $\tilde{O}(\epsilon^{-1})$ samples) between $(\pi_1 A)(\tilde{S}) \geq \epsilon$ and $(\pi_1 A)(\tilde{S}) \leq \epsilon/2$. If the latter case occurs, then the distribution conditioned on the heavy prefixes cannot contribute more than $\epsilon/2$ to A's distance from independence. Otherwise, if we are guaranteed that the second case does not occur, we can simulate the distribution for $A_{|\tilde{S}\times[m]}$ easily—we sample from A until we find a member of $\tilde{S} \times [m]$ which we output; this takes $O(\epsilon^{-1}\log(nm))$ queries with a high enough success probability. We then apply an independence test that works well for heavy prefixes to $A_{|\tilde{S}\times[m]}$.

Next we distinguish between $(\pi_1 A)([n]\backslash\tilde{S}) \geq \epsilon$ and $(\pi_1 A)([n]\backslash\tilde{S}) \leq \epsilon/2$. Again if the latter occurs, then the distribution conditioned on light elements can contribute at most $\epsilon/2$ to the distance from independence. Otherwise, if the latter does not occur, as before we simulate the distribution $A_{|([n]\backslash\tilde{S})\times[m]}$, and use it with a test that works well for distributions restricted to light prefixes (they will still remain light enough provided that $(\pi_1 A)([n]\backslash\tilde{S}) \geq \epsilon/2$).

Finally, we obtain a test for independence (Section 3.3) by merging the testing over light and heavy prefixes and then applying Theorem 14 to ensure the consistency of the distributions.

3.1 The heavy prefixes

We show that using sieves, the heavy prefixes can be tested for independence using roughly $\tilde{O}((n^{\alpha} + m)\text{poly}(\epsilon^{-1}))$ samples. In fact, the following theorem yields a general algorithm for testing independence; it is just that the sample complexity is particularly appealing in the heavy prefix case. Note that in this case $|S| = O(n^{\alpha})$.

Theorem 18 *There is an algorithm that given a black-box distribution A over $S \times T$: (1) if A is independent, it outputs PASS with high probability and (2) if A is not 3ϵ-independent, it outputs FAIL with high probability. The algorithm uses $\tilde{O}((|S| + |T|)\text{poly}(\epsilon^{-1}))$ samples.*

PROOF: Let $\tilde{A}_1$ be an explicit distribution which approximates $\pi_1 A$. Consider the following independence test:

```
Algorithm TestHeavyIndependence$(A, \tilde{A}_1, \epsilon)$
    (1) $\mathcal{S} \overset{\text{def}}{=} \{S_0, S_1, \ldots, S_k\} = $ Bucket $(\tilde{A}_1, S, \epsilon/75)$.
    (2) Obtain an approximation $\tilde{A}_2$ of $\pi_2 A$ to within an
        $\epsilon/75$ factor, on a $\tilde{T}$ that includes all $j \in [m]$ which
        have probability at least $(m \log m)^{-1}$.
    (3) $\mathcal{T} \overset{\text{def}}{=} \{T_0, T_1, \ldots, T_\ell\} =$Bucket $(\tilde{A}_2, \tilde{T}, \epsilon)$; add
        $T \backslash \tilde{T}$ to $T_0$.
    (4) For $(S_i, T_j), i \in [k], j \in [\ell]$ do
    (5)     If $A(S_i \times T_j)$ is not small, then
    (6)         If $\pi_1 A_{|S_i \times T_j}$ or $\pi_2 A_{|S_i \times T_j}$ are not both $\epsilon/25$-uniform or if $A_{|S_i \times T_j}$ is not $\epsilon$-independent, then FAIL.
    (7) If $A_{\langle \mathcal{S} \times \mathcal{T} \rangle}$ is not $\epsilon/2$-independent, then FAIL.
    (8) PASS.
```

Note that, if needed, $\tilde{A}_1$ can be obtained using $\tilde{O}|S|\text{poly}(\epsilon^{-1})$ samples. After step (2), S_0 can be ignored (as usual). By Theorem 17, the uniformity test in step (6) can be done using $O(\epsilon^{-4}\sqrt{n})$ samples of $A_{|S_i \times T_j}$. The independence test in step (7) can be done by brute force, for instance, since the alphabet is only logarithmic in $|S|$ and $|T|$. Also, by bucketing, we know that $|\pi_1 A - U_{S_i}| \leq \epsilon/25, \forall i \in [k]$ and $|\pi_2 A - U_{T_j}| \leq \epsilon/25, \forall j \in [\ell]$. For deciding in step (5) whether to execute step (6), we distinguish between $A(S_i \times T_j) \geq \epsilon/(k\ell)$ and $A(S_i \times T_j) \leq \epsilon/(2k\ell)$, by taking $\tilde{O}(k\ell/\epsilon)$ many samples of A and counting how many of them are in $S_i \times T_j$. Step (6) requires sampling of $A_{|S_i \times T_j}$; this is done by repeatedly sampling A until a member of $S_i \times T_j$ is found. As we are assured in step (6) that $A(S_i \times T_j) > \epsilon/(2k\ell)$, it suffices to take $O(\epsilon^{-1} \log^3(nm))$ samples of A in order to generate a single sample of $A_{|S_i \times T_j}$ (remember that k and ℓ are logarithmic in n and m).

We now present the independence test in step (6) which is used for each pair of buckets from $\mathcal{S}$ and $\mathcal{T}$.

Lemma 19 *There is an algorithm that given a black-box distribution A over $S \times T$ such that $|\pi_1 A - U_S| \leq \epsilon/25, |\pi_2 A - U_T| \leq \epsilon/25$: (1) if A is independent, it outputs PASS with high probability and (2) if A is not ϵ-close to $U_{S \times T}$, it outputs FAIL with high probability (in particular, only one of these cases can occur for a distribution satisfying the above conditions). The algorithm uses $\tilde{O}((|S| + |T|)\text{poly}(\epsilon^{-1}))$ samples.*

PROOF: We apply the (A, B)-sieve from Corollary 11. By its properties, if A is independent then $B = U_{S \times T}$, and if A is not ϵ-close to $U_{S \times T}$, then $|B - U_{S \times T}| \geq \epsilon/25$ (because $|A - B| \leq \frac{24}{25}\epsilon$). We can distinguish between these cases using Theorem 17, with $\tilde{O}(\epsilon^{-1}\sqrt{|S \times T|})$ samples from the sieve, which in itself takes less than a total of $\tilde{O}(\epsilon^{-4}(|S| + |T|)\log^6(\epsilon^{-1}(|S| + |T|)))$ samples from A. ■

Note that in the application of Lemma 19, its sampling estimate should be further multiplied by $O(\epsilon^{-1} \log^3(nm))$ to get the total number of samples made from A, because it is applied separately to the restriction of A to each $S_i \times T_j$.

We now return to the proof of the theorem. If A is independent, then for all $i \in [k], j \in [\ell]$, the restriction $A_{|S_i \times T_j}$ is independent so steps (4)–(6) pass (remember that Lemma 19 ensures that independent distributions pass step (6)). In the above case, also $A_{\langle\mathcal{S}\times\mathcal{T}\rangle}$ is independent, so step (7) and thus the entire algorithm passes as well.

Conversely, if for each $i \in [k]$ and $j \in [\ell]$ for which step (6) was performed $\pi_1 A_{|S_i\times T_j}$ and $\pi_2 A_{|S_i\times T_j}$ are both $\epsilon/25$-uniform, $|A_{|S_i\times T_j} - U_{S_i\times T_j}| \leq \epsilon$ (this step will not pass otherwise by Lemma 19), and $|A_{\langle\mathcal{S}\times\mathcal{T}\rangle} - D| \leq \frac{1}{2}\epsilon$ where D over $[k]\times[\ell]$ is an independent distribution, then we show that A is 3ϵ-independent. First note that $A(T_0) \leq (1-\epsilon)/\log n$. Now, we define a new random variable B over $S\times T$ which is defined by first generating an $(i,j)\in[k]\times[\ell]$ according to D, and then generating $(i',j')\in S_i\times T_j$ according to $U_{S_i\times T_j}$. It is easy to see that B is independent. Finally, by Lemma 6, $|A-B| \leq (3/2)\epsilon + \epsilon + (1-\epsilon)/\log n \leq 3\epsilon$, where the second term comes for possibly ignoring pairs i,j for which $A(i,j) < \epsilon/(k\ell)$ and the third term comes from ignoring $A(T_0)$.

The sample complexity of this algorithm is dominated by the complexity for each pair of buckets going through the test of Lemma 19. It brings us to a total sample complexity of $\tilde{O}((|S|+|T|)\text{poly}(\epsilon^{-1}))$ samples. ■

3.2 The light prefixes

We show that using the test for L_1 distance between distributions, the light prefixes can be tested for independence using roughly $\tilde{O}((n^{2-2\alpha}m + n^{2/3})\text{poly}(\epsilon^{-1}))$ samples. Formally, we prove:

Theorem 20 *There is an algorithm that given a black-box distribution A over $S\times T$ with $\|\pi_1 A\|_\infty \leq 2\epsilon^{-1}|S|^\alpha$ such that: (1) if A is independent, it outputs PASS with high probability and (2) if A is not 3ϵ-independent, it outputs FAIL with high probability. The algorithm uses $\tilde{O}((|S|^{2-2\alpha}|T| + |S|^{2/3})\text{poly}(\epsilon^{-1}))$ samples.*

PROOF: The following is the outline of the algorithm. Note that $\{\{x\} \mid x\in S\}$ is the partition of S into singletons.

```
Algorithm TestLightIndependence($A, \epsilon$)
    (1) Obtain an approximation $\tilde{A}_2$ of $\pi_2 A$ within an $\epsilon/75$
        factor, on a $\tilde{T}$ which includes all $j \in [m]$ which
        have probability at least $(m\log m)^{-1}$.
    (2) $\mathcal{T} \stackrel{\text{def}}{=} \{T_0, T_1, \ldots, T_\ell\}$ =Bucket ($\tilde{A}_2, \tilde{T}, \epsilon$); add
        $T\backslash\tilde{T}$ to $T_0$.
    (3) For $j = 1, \ldots, \ell$ do
    (4)     If $A(S\times T_j)$ is not small, then
    (5)         If $|A_{|S\times T_j} - (\pi_1 A_{|S\times T_j}) \times (\pi_2 A_{|S\times T_j})| \geq \epsilon$,
                then FAIL.
    (6) Let $j'$ be such that $A(S\times T_{j'}) > \epsilon/(4\ell)$.
    (7) For $j = 1, \ldots, \ell$ do
    (8)     If $A(S\times T_j)$ is not small, then
    (9)         If $|A_{|S\times T_{j'}} - A_{|S\times T_j}| \geq \epsilon$, then FAIL.
   (10) PASS.
```

The decisions in step (4) and step (8) are done in a similar manner to what was done in Theorem 18. We distinguish between $A(S\times T_j) \geq \epsilon/(2\ell)$ and $A(S\times T_j) \leq \epsilon/(4\ell)$ by taking $\tilde{O}(\ell/\epsilon)$ samples of A. This guarantees that we need to take $O(\text{poly}(\log(nm))\ell/\epsilon)$ samples of A for every sample of $A_{|S\times T_j}$ required in step (5) and step (9), by re-sampling A until we obtain a member of the required set (similarly step (6) guarantees this for sampling $A_{|S\times T_{j'}}$).

The projections appearing in step (5) are sampled by sampling the respective distribution and ignoring a coordinate. Obtaining the j' in step (6) can be done for example using a brute-force approximation of $A_{\langle\{S\}\times\mathcal{T}\rangle}$.

The test for the distribution difference in step (5) is done by using Theorem 15 with parameter ϵ and the distributions $A_{|S\times T_j}$ and $(\pi_1 A_{|S\times T_j})\times(\pi_2 A_{|S\times T_j})$; the bound on the L_∞ norm of the distributions involved will be given below. The test for the difference in step (9) is done similarly, but this time using Theorem 14 with parameter ϵ.

Notice that $\|A_{|S\times T_j}\|_\infty \le 2|S|^{-\alpha}/\epsilon$ for every T_j (because of the bound on $\|\pi_1 A\|_\infty$), and that $\|\pi_2 A_{|S\times T_j}\|_\infty \le (1+3\epsilon)|T_j|^{-1}$.

The total sample complexity for steps (3)–(5) is given by $\log|T|$ times the sample complexity for iteration j. The sample complexity of the latter is given by Theorem 15, which is $\tilde{O}((1+3\epsilon)\cdot(|S||T_j|)^2\cdot|S|^{-\alpha}\cdot|S|^{-\alpha}|T_j|^{-1}\cdot\epsilon^{-5})$, times the $\tilde{O}(\ell/\epsilon)$ for sampling from the restrictions to the buckets. This clearly dominates the sample complexity for step (6), and the sample complexity for steps (7)–(9), which is $\tilde{O}(|S|^{2/3}\epsilon^{-5})$ by multiplying the estimate of Theorem 14, the sample complexity of the restricted distributions, and the number of iterations.

As for correctness, if A is independent then it readily follows that the algorithm accepts, while on the other hand it is not hard to see that if the distribution pairs compared in step (5) and step (9) are indeed all ϵ-close, then A is 3ϵ-independent. ■

3.3 Putting them together

We now give the algorithm for the general case.

Theorem 21 *For $n \ge m$, there is an algorithm that given a distribution A over $[n]\times[m]$ and an $\epsilon > 0$: (1) if A is independent, it outputs PASS with high probability and (2) if A is not 7ϵ-independent, it outputs FAIL with high probability. The algorithm uses $\tilde{O}(n^{2/3}m^{1/3}\mathrm{poly}(\epsilon^{-1}))$ samples.*

PROOF: The following is the outline of the algorithm.

```
Algorithm TestIndependence($A, n, m, \epsilon$)
    (1) Let $\beta$ be such that $m = n^\beta$, and set $\alpha = (2+\beta)/3$.
    (2) Obtain an approximation $\tilde{A}_1$ of $\pi_1 A$ to within an
          $\epsilon/75$ factor, on an $\tilde{S}$ which includes all $i \in [n]$ which
          have probability at least $n^{-\alpha}$ and no $i \in [n]$ which
          has probability at most $n^{-\alpha}/2$.
    (3) If $(\pi_1 A)(\tilde{S})$ is not small then
    (4)      If TestHeavyIndependence($A_{|\tilde{S}\times[m]}, \tilde{A}_{1|\tilde{S}\times[m]}, \epsilon$)
               fails then FAIL.
    (5) If $(\pi_1 A)([n]\backslash\tilde{S})$ is not small then
    (6)      If TestLightIndependence($A_{|([n]\backslash\tilde{S})\times[m]}, \epsilon$) fails
               then FAIL.
    (7) If both $(\pi_1 A)(\tilde{S})$ and $(\pi_1 A)([n]\backslash\tilde{S})$ are not small
          then
    (8)      If $\pi_2 A_{|\tilde{S}\times[m]}$ and $\pi_2 A_{|([n]\backslash\tilde{S})\times[m]}$ are not $\epsilon$-close,
        then FAIL.
    (9) PASS.
```

In the above algorithm, steps (3), (5) and (7) use sampling to distinguish between the cases where the respective quantities are at least ϵ and the cases where they are at most $\epsilon/2$. Step (4) (if required) is done by using Theorem 18, and step (6) is done by using Theorem 20; by the choice of α in step (1), the number of queries in both is $\tilde{O}(n^{2/3}m^{1/3}\text{poly}(\epsilon^{-1}))$ times the $O(\epsilon^{-1}\log(nm))$ queries required for sifting the restricted distributions (a factor which does not change the above estimate).

In step (8) the two distributions are fed into the algorithm of Theorem 14, parametrized to guarantee failure if these distributions are more than ϵ-apart; this uses a number of queries that is dominated by the terms in the rest of the algorithm.

It is clear that if A is independent, then the test will accept with high probability. We now prove that if the test accepts, then A is at least 7ϵ-independent.

If steps (4), (6) and (8) are performed and none of the above tests fails, then by a final application of Lemma 6, where $\mathcal{R} = \{\tilde{S} \times [m], ([n]\backslash\tilde{S}) \times [m]\}$, we get that our distribution is at least 7ϵ-independent (because step (8) guarantees that the coarsening is not more than ϵ-far from being independent). If steps (4) and (8) are not performed, then $A(\tilde{S} \times [m]) < \epsilon$, so it contributes no more than ϵ to the farness of A from being independent, and so step (6) is sufficient to guarantee 4ϵ-independence. Similarly 4ϵ-independence holds if steps (6) and (8) are not performed since in this case $A(([n]\backslash\tilde{S}) \times [m])$ is small. This covers all possible cases and concludes the proof. ■

4 Testing against a known distribution

In this section we assume that the distributions X and Y are over $[n]$, where X is a black-box distribution and Y is explicitly given. The task is to determine if $|X - Y| < \epsilon$ using as few samples (from X) as possible. We show that this can be done using roughly $\tilde{O}(\sqrt{n}\text{poly}(\epsilon^{-1}))$ samples.

The main technical idea is to use bucketing (Section 2.3) to reduce this problem to that of testing that each of several distributions is approximately uniform. We first bucket the given distribution Y; recall that bucketing gives a partition $\{R_0, \ldots, R_k\}$ of the domain so that the distribution is close to uniform in each of the partitions R_i (Lemma 8). For each partition R_i, we sample X and test if $X_{|R_i}$ is close to uniform on R_i. This can be accomplished using Theorem 12.

First, we need an additional step to interpret L_2 results in terms of the L_1 norm.

Lemma 22 *For any distribution X over R, $\|X\|^2 - \|U_R\|^2 = \|X - U_R\|^2$.*

Lemma 23 *Let X, Y be distributions over $[n]$ and let $(R_0, \ldots, R_k) = \mathit{Bucket}(Y, [n], \epsilon)$. For each i in $[k]$, if $\|X_{|R_i}\|^2 \leq (1 + \epsilon^2)/|R_i|$ then $|X_{|R_i} - U_{R_i}| \leq \epsilon$ and $|X_{|R_i} - Y_{|R_i}| \leq 2\epsilon$.*

PROOF: By Cauchy-Schwartz $|X_{|R_i} - U_{|R_i}| \leq \sqrt{|R_i|}\, \|X_{|R_i} - U_{|R_i}\|$ which by Lemma 22, equals $\sqrt{|R_i|}(\|X_{|R_i}\|^2 - \|U_{|R_i}\|^2)^{1/2} = \sqrt{|R_i|}((1 + \epsilon^2)/|R_i| - 1/|R_i|)^{1/2} = \epsilon$. As for the second statement, using Lemma 8 and triangle inequality, $|X_{|R_i} - Y_{|R_i}| < |X_{|R_i} - U_{|R_i}| + |U_{|R_i} - Y_{|R_i}| \leq 2\epsilon$.

■

Now, we give the complete algorithm to test if a black-box distribution X is close to an explicitly specified distribution Y.

Algorithm *TestIdentity*(X, Y, n, ϵ)

```
(1) $\mathcal{R} \overset{\text{def}}{=} \{R_0, \ldots, R_k\} = \textit{Bucket}(Y, n, \epsilon/\sqrt{2})$.
(2) Let $M$ be a set of $O(\sqrt{n}\epsilon^{-2}\log n)$ samples from $X$.
(3) For each partition $R_i$ do
(4)     Let $M_i = M \cap R_i$ (preserving repetitions);
        let $\ell_i = |M_i|$ (counting also repetitions).
(5)     If $Y(R_i) \geq \epsilon/k$ then
(6)         If $\ell_i < O(\sqrt{n}\epsilon^{-2})$ then FAIL.
(7)         Estimate $\|X_{|R_i}\|^2$ using $M_i$. (Thm. 12)
(8)         If $\|X_{|R_i}\|^2 > (1+\epsilon^2)/|R_i|$ then FAIL.
(9) If $|X_{\langle\mathcal{R}\rangle} - Y_{\langle\mathcal{R}\rangle}| > \epsilon$ then FAIL.
(10) PASS.
```

Theorem 24 *Algorithm TestIdentity(X, Y, n, ϵ) is such that: (1) if $|X - Y| \leq \frac{\epsilon^3}{4\sqrt{n}\log n}$, it outputs PASS with high probability and (2) if $|X - Y| > 6\epsilon$, it outputs FAIL with constant probability. The algorithm uses $\tilde{O}(\sqrt{n}\text{poly}(\epsilon^{-1}))$ samples.*

PROOF: Step (9) can be done by using brute force to distinguish between $|X_{\langle\mathcal{R}\rangle} - Y_{\langle\mathcal{R}\rangle}| > \epsilon$ and $|X_{\langle\mathcal{R}\rangle} - Y_{\langle\mathcal{R}\rangle}| < \frac{1}{2}\epsilon$. This does not take a significant number of additional samples, as k is logarithmic in n.

Note that by Chernoff bounds, the probability of failing in step (6) can be made sufficiently small, unless there is a large difference between $X(R_i)$ and $Y(R_i)$ for some i. Suppose that the algorithm outputs PASS. This implies that for each partition R_i for which steps (6)–(8) were performed (which are those for which $Y(R_i) \geq \epsilon/k$), we have $\|X_{|R_i}\|^2 \leq (1+\epsilon^2)/|R_i|$. From Lemma 23 we get that for each of these R_i, $|X_{|R_i} - Y_{|R_i}| \leq 2\epsilon$.

We also have that the sum of $Y(R_i)$ over all R_i for which steps (6)–(8) were skipped is at most ϵ. Also, $|X_{\langle\mathcal{R}\rangle} - Y_{\langle\mathcal{R}\rangle}| \leq \epsilon$ by step (9); so the total difference between X and Y over these partitions sums up to no more than 3ϵ. Adding this to the 3ϵ difference over the partitions that were not skipped in steps (6)–(8) (given by applying Lemma 6 with $|X_{|R_i} - Y_{|R_i}| \leq 2\epsilon$ and $|X_{\langle\mathcal{R}\rangle} - Y_{\langle\mathcal{R}\rangle}| \leq \epsilon$), we get that $|X - Y| \leq 6\epsilon$.

On the other hand, suppose $|X - Y| < \frac{\epsilon^3}{4\sqrt{n}\log n}$. From the definition of the bucketing algorithm, step (1) will return a partition with $k = (2/\log(1+\epsilon/\sqrt{2})) \cdot \log n < (2\sqrt{2}/\epsilon) \cdot \log n$ elements. Using Lemma 7 for all partitions R_i with $Y(R_i) \geq \epsilon/k > \epsilon^2/(2\sqrt{2}\log n)$, we have $|X_{|R_i} - Y_{|R_i}| < \epsilon/(\sqrt{2n})$. In terms of $\|\cdot\|$, this implies $\|X_{|R_i} - Y_{|R_i}\|^2 < \epsilon^2/(2n) < \epsilon^2/(2|R_i|)$. Since from Lemma 8, $\|Y_{|R_i} - U_{R_i}\|^2 < \epsilon^2/(2|R_i|)$, then by the triangle inequality, $\|X_{|R_i} - U_{R_i}\|^2 \leq \|X_{|R_i} - Y_{|R_i}\|^2 + \|Y_{|R_i} - U_{R_i}\|^2 \leq \epsilon^2/|R_i|$. So by Lemma 22, $\|X_{|R_i}\|^2 = \|X_{|R_i} - U_{R_i}\|^2 + \|U_{R_i}\|^2 \leq (1+\epsilon^2)/|R_i|$. Therefore the algorithm will pass with high probability on all such partitions; it is also not hard to see that the algorithm will pass step (9) as well.

The sample complexity is $\tilde{O}(\sqrt{n}\epsilon^{-2})$ from step (2), which dominates the sample complexity of step (9) (no other samples are taken throughout the algorithm). ■

5 Lower bound for testing independence

Theorem 25 *For any algorithm $\mathcal{A}$ using $o(n^{2/3}m^{1/3})$ samples whenever $n \geq m$, there exist two joint distributions over $[n] \times [m]$ for any sufficiently large $n \geq m$, with one of them being independent and the other not being $(1/6)$-independent, such that $\mathcal{A}$ cannot distinguish between these two joint distributions with probability greater than 2/3.*

PROOF: Fix an algorithm $\mathcal{A}$ using $o(n^{2/3}m^{1/3})$ samples. We first define two joint distributions A_0 and B_0 over $[n] \times [m]$. Let $\beta = \log_n m$ and $\alpha = (2+\beta)/3$.

$$\Pr\left[A_0 = (i,j)\right] = \begin{cases} \frac{1}{2n^\alpha m} & \text{if } 1 \le i \le n^\alpha \\ \frac{1}{mn} & n/2 < i \le n \\ 0 & \text{otherwise} \end{cases}$$

$$\Pr\left[B_0 = (i,j)\right] = \begin{cases} \frac{1}{2n^\alpha m} & \text{if } 1 \le i \le n^\alpha \\ \frac{2}{mn} & \text{if } \begin{array}{l} n/2 < i \le n \text{ and} \\ j \in [1,\ldots,m/2] \end{array} \\ 0 & \text{otherwise} \end{cases}$$

We now define two joint distributions A and B such that A, B modify A_0 and B_0 by randomly relabeling each element in $[n]$ and $[m]$. First choose random permutations σ_0 of $[n]$ and $\sigma_1, \ldots, \sigma_n$ of $[m]$. Define A to be the distribution such that

$$\Pr\left[A = (\sigma_0(i), \sigma_i(j))\right] = \Pr\left[A_0 = (i,j)\right].$$

Likewise define B to be the distribution such that

$$\Pr\left[B = (\sigma_0(i), \sigma_i(j))\right] = \Pr\left[B_0 = (i,j)\right].$$

Note that A and B are actually families of distributions (indexed by the permutations). Throughout the rest of the proof, we will refer to A and B, with an abuse of notation, as individual distributions in these families. Since we fixed the algorithm $\mathcal{A}$, we could choose the permutations $\sigma_0, \ldots, \sigma_n$ to obtain the members of these families that maximizes the error probability of the algorithm $\mathcal{A}$.

The distribution A is independent whereas the distribution B is $\frac{1}{6}$-far from independent. This follows from B being $\frac{1}{2}$-far from $\pi_1 B \times \pi_2 B$ and Proposition 1. The distributions $\pi_1 A$ and $\pi_1 B$ are identical, and they give half the weight to a small number, namely n^α, of the elements, and distribute the remaining weight to half of the elements. The distribution $\pi_2 A$ is uniform over its domain independent of the value of $\pi_1 A$. The distribution $\pi_2 B$, however, is uniform over its domain only when $\pi_1 B$ outputs an element with the higher weight, otherwise, conditioned on the event that $\pi_1 B$ takes on a value with the lower probability, $\pi_2 B$ is uniform only on a subset of its domain that is half the size. The choice of σ_i's makes the distribution $\pi_2 B$ uniform on its domain.

Definition 26 *For a pair $(i,j) \in [n] \times [m]$, i is the* prefix. *An element $(i,j) \in [n] \times [m]$ such that $\Pr\left[A \text{ (or } B\text{) takes on value } (i,j)\right] = \frac{1}{2n^\alpha m}$ is called a* heavy element. *The prefix i of a heavy element (i,j) is called a* heavy prefix. *Elements and prefixes with non-zero probabilities that are not heavy are called* light.

When restricted to the heavy prefixes, both joint distributions are identical. The only difference between A and B comes from the light prefixes, and the crux of the proof will be to show that this difference will not change the relevant statistics in a statistically significant way. We do this by showing that the only really relevant statistic is the number of prefixes that occur exactly twice and each time with different suffix. We then show that this statistic has a very similar distribution when generated by A and B because the expected number of such prefixes that are light is much less than the standard deviation of the number of such prefixes that are heavy.

Next, we describe an aggregate representation of the samples that $\mathcal{A}$ takes. We then prove that we can assume without loss of generality that $\mathcal{A}$ is given this representation of the samples as input instead of the samples themselves. Then, we conclude the proof by showing that distributions on the fingerprint when the samples are taken from A or B are indistinguishable.

Definition 27 *Fix a set of samples $S = \{(x_1, y_1), \ldots, (x_s, y_s)\}$ from distribution A over $[n] \times [m]$. Say the* pattern *of prefix x_i is $\vec{c}$ where c_j is the number of y's such that (x_i, y) appears exactly j times in S. Define the function $d_S(\vec{c})$ to be the number of prefixes x for which the pattern of x is $\vec{c}$. We refer to d_S as the* fingerprint *of S. We will just use $d(\vec{c})$ when S is clear from context.*

The next claim shows that the fingerprint of the sample is just as useful as the samples themselves to distinguish between A and B.

Claim 28 *Given algorithm $\mathcal{A}$ which for joint distributions chosen from the family A or B, correctly distinguishes whether the distribution is independent or ϵ-far from independent, there exists algorithm $\mathcal{A}'$ which gets as input only the fingerprint of the generated sample and has the same correctness probability as $\mathcal{A}$.*

PROOF: Note that one can view a sample of size s chosen from the distribution A (respectively B) as first picking s samples from A_0 (respectively, B_0), then picking a set of random permutations of the element labels and outputting the random relabeling of the samples. Thus the randomness used to generate the sample can be divided into two parts: the first set of coins $\phi = (\phi_1, \ldots, \phi_u)$ are the coins used to generate the sample from A_0 (B_0) and the second set of coins $\psi = (\psi_1, \ldots, \psi_v)$ are the coins used to generate the random permutations of the element labels.

The main idea behind the proof is that given the fingerprint of a sample from A_0 (respectively B_0), the algorithm $\mathcal{A}'$ can generate a labeled sample with the same distribution as A (respectively, B) without knowing which part of the fingerprint is due to heavy or light elements or whether the sample is from A or B. In particular, given the fingerprint, assign $d(\vec{b})$ distinct labels from $[n]$ to each pattern $\vec{b}$. Suppose that $x_{\vec{b}}$ is assigned to pattern $\vec{b}$. Then create a sample which includes i copies of $(x_{\vec{b}}, y_j)$ for each nonzero b_i and distinct y_j for $1 \leq j \leq b_i$. Then choose random permutations $\sigma_0, \sigma_1, \ldots, \sigma_n$ of $[n]$ and $[m]$ and use them to relabel the prefixes and suffixes of the sample accordingly.

Thus, $\mathcal{A}'$ generates a sample from the fingerprint and feeds it to $\mathcal{A}$ as input. For each choice of the sample from A_0 according to random coins ϕ, we have that $\Pr_{\psi}[\mathcal{A}' \text{ correct}] = \Pr_{\psi}[\mathcal{A} \text{ correct}]$. Therefore, $\Pr_{\phi,\psi}[\mathcal{A}' \text{ correct}] = \Pr_{\phi,\psi}[\mathcal{A} \text{ correct}]$. ■

The following lemma shows that it is only the heavy prefixes, which have identical distributions in both A and B, that contribute to most of the entries in the fingerprint.

Lemma 29 *The expected number of light prefixes that occur at least three times in the sample such that at least two of them are the same element is $o(1)$ for both A and B.*

PROOF: For a fixed light prefix, the probability that at least three samples will land in this prefix and two of these samples will collide is $o(n^{-1})$. Since there are $n/2$ light prefixes, by the linearity of expectation, the expected number of such light prefixes in the sample is $o(1)$. ■

We would like to have the pattern of each prefix be independent of the patterns of the other prefixes. To achieve this we assume that algorithm $\mathcal{A}$ first chooses an integer s_1 from the Poisson distribution with the parameter $\lambda = s = o(n^{2/3}m^{1/3})$. The Poisson distribution with the positive parameter λ has the probability mass function $p(k) = \exp(-\lambda)\lambda^k/k!$. Then, after taking s_1 samples from the input distribution, $\mathcal{A}$ decides whether to accept or reject the distribution. In the following, we show that $\mathcal{A}$ cannot distinguish A from B with success probability at least $2/3$. Since s_1 will have a value larger than $s/2$ with probability at least $1 - o(1)$ and we will show an upper bound on the statistical distance of the distributions of two random variables (i.e., the distributions on the fingerprints), it will follow that no symmetric algorithm with sample complexity $s/2$ can distinguish A from B.

Let F_{ij} be the random variable that corresponds to the number of times that the element (i, j) appears in the sample. It is well known that F_{ij} is distributed identically to the Poisson distribution with parameter $\lambda = sr_{ij}$, where r_{ij} is the probability of element (i, j) (cf., Feller [2], p. 216). Furthermore, it can also be shown that all F_{ij}'s are mutually independent. The random variable $F_i \overset{\text{def}}{=} \sum_j F_{ij}$ is distributed identically to the Poisson distribution with parameter $\lambda = s \sum_j r_{ij}$.

Let D_A and D_B be the distributions on all possible fingerprints when samples are taken from A and B, respectively. The rest of the proof proceeds as follows. We first construct two processes P_A and P_B that generate distributions on fingerprints such that P_A is statistically close to D_A and P_B is statistically close to D_B. Then, we prove that the distributions P_A and P_B are statistically close. Hence, the theorem follows by the indistinguishability of D_A and D_B.

Each process has two phases. The first phase is the same in both processes. They randomly generate the prefixes of a set of samples using the random variables F_i defined above. The processes know which prefixes are heavy and which prefixes are light, although any distinguishing algorithm does not. For each heavy prefix, the distribution on the patterns is identical in A and B and is determined by choosing samples according to the uniform distribution on elements with that prefix. The processes P_A and P_B use the same distribution to generate the patterns for each heavy prefix. For each each light prefix i that appears k times for $k \neq 2$, both P_A and P_B will determine the pattern of the prefix to be $(k, \vec{0})$. This concludes the first phase of the processes.

In the second phase, P_A and P_B determine the entries of the patterns for the light prefixes that appear exactly twice. These entries are distributed differently in P_A and P_B. There are only two patterns to which these remaining prefixes can contribute: $(2, \vec{0})$ and $(0, 1, \vec{0})$. For each light prefix that appears exactly twice, P_A sets the pattern to be $(2, \vec{0})$ with probability $1 - (1/m)$ and $(0, 1, \vec{0})$ otherwise. For such light prefixes, P_B sets the pattern to be $(2, \vec{0})$ with probability $1 - (2/m)$ and $(0, 1, \vec{0})$ otherwise.

Since the patterns for all prefixes are determined at this point, both process output the fingerprint of the sample they have generated. We show:

Lemma 30 *The output of P_A, viewed as a distribution, has L_1 distance $o(1)$ to D_A. The output of P_B, viewed as a distribution, has L_1 distance $o(1)$ to D_B.*

PROOF: The distribution that P_A generates is the distribution D_A conditioned on the event that all light prefixes has one of the following patterns: $(k, \vec{0})$ for $k \geq 0$ or $(0, 1, \vec{0})$. Since this conditioning holds true with probability at least $1 - o(1)$ by Lemma 29, $|P_A - D_A| \leq o(1)$. The same argument applies to P_B and D_B. ■

Finally, we show (omitted) that the component of the fingerprint that creates the difference between P_A and P_B is normally distributed in both cases. Moreover, the expectations of these two distributions are close enough (relative to their standard deviations) so that they are indistinguishable. Using this, it can be shown (omitted):

Lemma 31 $|P_A - P_B| \leq 1/6$.

PROOF: Given the number of times a prefix appears in the sample, the pattern of that prefix is independent of the patterns of all the other prefixes. By the generation process, the L_1 distance between P_A and P_B can only arise from the second phase. We show that the second phases of the processes do not generate an L_1 distance larger than $1/6$.

Let G (respectively, H) be the random variable that corresponds to the values $d(2,\vec{0})$ when the input distribution is A (respectively, B). Let d' be the part of the fingerprint excluding entries $d(2,\vec{0})$ and $d(0,1,\vec{0})$. We will use the fact that for any d', $\Pr[P_A \text{ gen. } d'] = \Pr[P_B \text{ gen. } d']$ in the following calculation.

$$\begin{aligned}
|P_A - P_B| &= \sum_d |\Pr[P_A \text{ gen. } d] - \Pr[P_B \text{ gen. } d]\,| \\
&= \sum_{d'} \Pr\left[P_A \text{ gen. } d'\right] \sum_{k\geq 0} \\
&\quad |\Pr\left[P_A \text{ gen. } d(2,\vec{0}) = k|d'\right] - \\
&\quad \Pr\left[P_B \text{ gen. } d(2,\vec{0}) = k|d'\right]\,| \\
&= \sum_{C\geq 0} \Pr[P_A \text{ gen. } C \text{ prefixes twice}] \sum_{0\leq k\leq C} \\
&\quad |\Pr\left[P_A \text{ gen. } d(2,\vec{0}) = k|C\right] - \\
&\quad \Pr\left[P_B \text{ gen. } d(2,\vec{0}) = k|C\right]\,| \\
&= |G - H|
\end{aligned}$$

Consider the composition of G and H in terms of heavy and light prefixes. In the case of A, let G_h be the number of heavy prefixes that contribute to $d(2,\vec{0})$ and G_l be the number of such light prefixes. Hence, $G = G_h + G_l$. Define H_h, H_l analogously. Then, G_h and H_h are distributed identically. In the rest of the proof, we show that the fluctuations in G_h dominate the magnitude of G_l.

Let ξ_i be the indicator random variable that takes value 1 when prefix i has the pattern $(2,\vec{0})$. Then, $G_h = \sum_{\text{heavy } i} \xi_i$. By the assumption about the way samples are generated, the ξ_i's are independent. Therefore, G_h is distributed identically to the binomial distribution on the sum of n^α Bernoulli trials with success probability $\Pr[\xi_i = 1] = \exp(-s/2n^\alpha)(s^2/8n^{2\alpha})(1-(1/m))$. An analogous argument shows that G_l is distributed identically to the binomial distribution with parameters $n/2$ and $\exp(-s/n)(s^2/2n^2)(1-(1/m))$. Similarly, H_l is distributed identically to the binomial distribution with parameters $n/2$ and $\exp(-s/n)(s^2/2n^2)(1-(2/m))$.

As n and m grow large enough, both G_h and G_l can be approximated well by normal distributions. Therefore, by the independence of G_h and G_l, G is also approximated well by a normal distribution. Similarly, H is approximated well by a normal distribution. That is,

$$\Pr[G = t] \to \frac{1}{\sqrt{2\pi}\mathrm{StDev}[G]} \exp(-(t - \mathrm{E}[G])^2/2\mathrm{Var}[G])$$

as $n \to \infty$.

Thus, $\Pr[G = t] = \Omega(1/\mathrm{StDev}[G])$ over an interval I_1 of length $\Omega(\mathrm{StDev}[G])$ centered at $\mathrm{E}[G]$. Similarly, $\Pr[H = t] = \Omega(1/\mathrm{StDev}[H])$ over an interval I_2 of length $\Omega(\mathrm{StDev}[H])$ centered at $\mathrm{E}[H]$.

Since $\mathrm{E}[G] - \mathrm{E}[H] = \mathrm{E}[G_l] - \mathrm{E}[H_l] = \exp(-s/n)(s^2/4n)(1/m) = o(\mathrm{StDev}[G])$, $I_1 \cap I_2$ is an interval of length $\Omega(\mathrm{StDev}[G_h])$. Therefore,

$$\sum_{t \in I_1 \cap I_2} |\Pr[G = t] - \Pr[H = t]| \leq o(1)$$

because for $t \in I_1 \cap I_2$, $|\Pr[G = t] - \Pr[H = t]| = o(1/\mathrm{StDev}[G])$. We can conclude that $\sum_t |\Pr[G = t] - \Pr[H = t]|$ is less than 1/6 after accounting for the probability mass of G and H outside $I_1 \cap I_2$. ■

The theorem follows by Lemma 30 and Lemma 31.

■

References

[1] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. *Proc. 41st FOCS*, pp. 259–269, 2000.

[2] W. Feller. *An Introduction to Probability Theory and Applications (Vol. I).* John Wiley & Sons Publishers, 1968.

[3] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *J. ACM*, 45(4):653–750, 1998.

[4] O. Goldreich and D. Ron. On testing expansion in bounded-degree graphs. *ECCC*, TR00-020, 2000.

[5] E. L. Lehmann. *Testing Statistical Hypotheses*. Wadsworth and Brooks/Cole, 1986.

[6] D. Ron. *Property Testing (A Tutorial).* In *Handbook on Randomized Computing (Vol. II)*, Kluwer Academic Publishers, 2001.

[7] R. Rubinfeld and M. Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM J. Comput.*, 25(2):252–271, 1996.

[8] A. Sahai and S. Vadhan. Manipulating statistical difference. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 43:251–270, 1999.