Electronic Colloquium on Computational Complexity, Report No. 201 (2025)

Interactive proof systems for FARNESS

Oded Goldreich Tal Herman Guy N. Rothblum

December 4, 2025

Abstract

We consider interactive proofs for the promise problem, called ϵ-FARNESS, in which the yes-instances are pairs of distributions over $[n]$ that are ϵ-far from one another, and the no-instances are pairs of identical distributions. For any $t \leq n^{2/3}$, we obtain an interactive proof in which the verifier has sample complexity $O(t/\epsilon^2)$ and the (honest) prover has sample complexity $\mathrm{poly}(1/\epsilon) \cdot (n/\sqrt{t})$. For $t = n^{2/3}$ this result is the best possible, because (as proved by Batu and Canonne (FOCS 2017)) the corresponding decision procedure has sample complexity $\Omega(n^{2/3})$.

We also obtain interactive proofs for the promise problem in which the yes-instances are distributions over $[n]$ that are ϵ-far from the uniform distribution, and the no-instance is the uniform distribution. For any $t \leq \sqrt{n}$, we obtain an interactive proof in which the verifier has sample complexity $O(t/\epsilon^2)$ and the (honest) prover has sample complexity $\mathrm{poly}(1/\epsilon) \cdot \widetilde{O}(n/t)$. This stand in contrast to the fact (proved by Chiesa and Gur (ITCS 2018)) that the verifier in any interactive proof for the complement promise problem must have sample complexity $\Omega(\sqrt{n})$.

Contents

ISSN 1433-8092

1 Introduction

The standard formulation of distribution testing problems specifies a set of distributions $\mathcal{D}$ and considers inputs of the form (X, ϵ) such that X is a distribution and $\epsilon > 0$ is a proximity parameter. Specifically, the testing problem is defined as a promise problem in which instances are pairs of the form (X, ϵ), YES-instances are pairs (X, ϵ) such that X is in $\mathcal{D}$, and NO-instances are pairs (X, ϵ) such that X is ϵ-far from $\mathcal{D}$.[1]

In contrast, an alternative formulation, typically used in lower bounds, is obtained by fixing the value of the proximity parameter (i.e., ϵ). In this case, the testing problem is a promise problem in which instances are distributions, YES-instances are distributions X that are in $\mathcal{D}$, and NO-instances are distributions X that are ϵ-far from $\mathcal{D}$. For simplicity, we shall use this formulation, but our presentation can be easily adapted to the first formulation.

Turning to general promise problems (regarding distributions), we note that, *from a technical perspective*, in the context of decision procedures, flipping the role of YES-instances and NO-instances has no effect (assuming that one refers to two-sided error probability deciders).[2] In both cases, the task is distinguishing between distributions in one set and distributions in another set. In contrast, the symmetry breaks in the context of interactive proofs, because different requirements apply to YES-instances and to NO-instances; that is, completeness (i.e., *existence* of a successful prover strategy) applies to YES-instances, whereas soundness (i.e., failure of *every* possible prover strategy) applies to NO-instances. Hence, a problem may have a more efficient proof system than its complement.

We note that, *from a conceptual perspective*, there is no symmetry between the set $\mathcal{D}$ and the set of distributions that are ϵ-far from $\mathcal{D}$. The former set is the primary, whereas the latter set is secondary, as evident from the fact that it is defined with reference to the primary. Still, the secondary set may be natural *per se*.

(We comment that an analogous discussion applies also in the context of testing properties of functions (except that in that context two-sided error testers are not necessarily the norm). We note that this discussion is related but different from the study of testing dual problems put forward by Tell [19].)[3]

1.1 The specific problems we consider

The first promise problem we consider corresponds to one of the most popular distribution testing problems, called UNIFORMITY.

Definition 1.1 (UNIFORMITY): *For a fixed n and $\epsilon > 0$, the instances are distributions over $[n]$, the* YES*-instances are distributions X such that $X \equiv U_n$, and the* NO*-instances are distributions X that are ϵ-far from U_n, where U_n denotes the uniform distribution over $[n]$ and X is ϵ-*far from *Y if the total variation distance between X and Y exceeds ϵ.*

[1] Recall that a distribution X is said to be ϵ-far from a set of distribution $\mathcal{D}$ if, for every distribution Y in $\mathcal{D}$, the total variation distance between X and Y is greater than ϵ.

[2] Recall that two-sided error probability is the common convention for testing distributions; this convention is justified in [9, Sec. 11.1.1].

[3] In the context of "testing dual problems" (as in [19]), one defines a dual problem by fixing a set S and a distance parameter δ, and considering the set of all functions that are δ-far from S, denoted $F = \Gamma_\delta(S)$. But once this dual set F is fixed, one considers testing it, which means distinguishing, for every $\epsilon > 0$, between functions in F and functions that are ϵ-far from F. Furthermore, as shown by Tell [19], the set of functions that are δ-far from F may be a proper superset of S.

The complement promise problem, denoted ϵ-FARfromUNI, consists of switching the YES and NO-instances. That is, the YES-instances of ϵ-FARfromUNI are distributions that are ϵ-far from U_n, whereas the NO-instances are distributions X such that $X \equiv U_n$.

These promise problems are a special case of a class of prblems in which we are asked to distinguish between distributions that equal a fixed distribution D_n over $[n]$ and distributions that are ϵ-far from D_n. Indeed, in Definition 1.1 we used $D_n = U_n$.

Going a step farther, we consider the problem of distinguishing between pairs of identical distributions and pairs of distributions that are far away from one another. That is, in these problems, instances consists of pairs of distributions (rather than of a single input distribution, which is tested against a fixed distribution (as discussed above)). Hence, we consider the promise problem that corresponds to EQUALITY, a very popular distribution testing problem.

Definition 1.2 (EQUALITY): *For a fixed n and $\epsilon > 0$, the instances are pairs of distributions over $[n]$, the* YES*-instances are (X, Y) such that $X \equiv Y$, and the* NO*-instances are (X, Y) such that X is ϵ-far from Y.*

The complement promise problem, called ϵ-FARNESS, consists of switching the YES and NO-instances. That is, the YES-instances of ϵ-FARNESS are pairs of distributions (X, Y) such that X is ϵ-far from Y, whereas the NO-instances are (X, Y) such that $X \equiv Y$.

1.2 Our results

We consider interactive proof systems for properties of distributions in which both the verifier and the prover can obtain samples from the unknown distributions (see definitions in Section 2). For promise problems that refer to a single distribution, we prove the following result.

Theorem 1.3 (interactive proof systems for ϵ-FARfromUNI): *For every $t = \omega(\epsilon^{-2} \cdot \log(1/\epsilon))$ and $m = \mathrm{poly}(1/\epsilon) \cdot n/t$ such that $t \leq m$, there exists an interactive proof system for ϵ-FARfromUNI in which the verifier has sample complexity t, the honest prover has sample complexity $\widetilde{O}(m/\epsilon^2)$, and the communication complexity is $O(t \cdot \log n)$. Furthermore, the prover is laconic* (i.e., it sends a constant number of bits).

We highlight the fact that the verifier in this proof system may use $o(\sqrt{n})$ samples. This stands in contrast to the fact (proved in [5]) that the verifier must use $\Omega(\sqrt{n})$ samples in any proof system for UNIFORMITY (i.e., the complement of ϵ-FARfromUNI). Furthermore, the honest prover in the proof system for FARfromUNI may use $o(n)$ samples. (Note that a total sample complexity of $\Omega(\sqrt{n})$ is essential (for a proof systems for FARfromUNI) because a tester may emulate both parties.)

The foregoing result extends to being far from any distribution D_n; that is, the promise problem in which the YES-instances are distributions that are ϵ-far from D_n, and the NO-instances are distributions that equal D_n (i.e., there is a single NO-instance – D_n itself). Hence, we obtain doubly-sublinear proof system for all these promise problems; that is, proof systems in which the verifier is more efficient than a decision procedure while the prover is more efficient than a corresponding learning algorithm.

Our main result is a doubly-sublinear proof system for ϵ-FARNESS. Here both the verifier and the prover can obtain samples from each of the two the unknown distributions (and the prover's goal is to convince the verifier that the two distributions are ϵ-far apart from one another, rather than identical).

Theorem 1.4 (interactive proof systems for ϵ-FARNESS): *For every $t = \Omega(1/\epsilon^2)$ and $m = \Omega(\epsilon^{-2} \cdot n/\sqrt{t})$ such that $t \leq m$, there exists an interactive proof system for ϵ-FARNESS in which the verifier has sample complexity t, the honest prover has sample complexity $\widetilde{O}(m/\epsilon^2)$, and the communication complexity is $O(t \cdot \log n)$. Furthermore, the prover is laconic* (i.e., it sends $O(1)$ bits).

In particular, for any $\beta \in [0, 1/3)$, we can use $t = O(n^{2\beta}/\epsilon^2)$ and $m = O(n^{1-\beta}/\epsilon^2)$. We highlight the fact that the verifier in this proof system may use $o(\sqrt{n})$ samples, and the honest prover in this system may use $o(n)$ samples. (In contrast, in any proof system for EQUALITY, the verifier must use $\Omega(\sqrt{n})$ samples, since UNIFORMITY is a special case of EQUALITY.) Note that a total sample complexity of $\Omega(n^{2/3})$ is essential (for a proof systems for FARNESS) because a tester may emulate both parties (whereas, by [2], the sample complexity of testing EQUALITY is $\Omega(n^{2/3})$).

As evident from the foregoing description, our main focus is on *doubly-sublinear interactive proofs*. Recall that these are proof systems in which the verifier is more efficient than a decision procedure, while the prover is more efficient than a corresponding learning algorithm. Our starting point is the fact, proved by Herman and Rothblum [17], that some natural (label-invariant) properties of distributions do not have doubly-sublinear interactive proofs of proximity, while they do have interactive proofs of proximity [15, 16]. Recently, Goldreich and Rothblum demonstrated the existence of doubly-sublinear interactive proofs of proximity for distributions [12], but their demonstration utilizes properties that are not natural. Our main contribution is presenting doubly-sublinear interactive proofs for natural (promise problems regarding) distributions.

Some laconic credits. Distribution testing emerged explicitly in the works of Batu *et. al.* [4, 3]. Interactive proof of proximity (for functions) emerged explicitly in [18], and interactive proof of proximity for distribitions where first considered in [5]. Doubly-efficient interactive proofs (for functions) emerged explicitly in [14], whereas doubly-efficient interactive proofs of proximity for distributions have been recently considered in [16]. The study of doubly-sublinear interactive proofs of proximity (for functions) was recently initiated in [1]. The notion of laconic provers emerged explicitly in [11].

Organization. We present two different proof systems for FARNESS. The first proof system (presented in Section 3) is obtained by a reduction to a tester of EQUALITY that uses a different number of samples from the two distributions. Such a tester, called asymmetric, was presented by Diakonikolas and Kane [6]. The second proof system (presented in Section 4) improves over the performance of the first one, but its analysis is more complex. (Specifically, the first proof system establishes Theorem 1.4 only for $t = \Omega(\sqrt{n}/\epsilon^2)$.) As for Theorem 1.3, it is proved in Section 5. But before all of these, in Section 2, by explicitly presenting the notion of an interactive proof for promise problems regarding distributions.

2 Interactive proofs for promise problems regarding distributions

The notion of an interactive proof for promise problems regarding distributions is a generalization of the notion of an interactive proof of proximity for distributions. We first present the generalized notion for a single distribution (as in Definition 1.1), and then extend it to handle a pair of distributions (as in Definition 1.2).

A key issue is that when we talk about "machines that are given an unknown distribution" (as their input), we mean that such a machine is given *oracle access* to a sampling devise for this distribution.[4] Hence, for a distribution X (over $[n]$) and a machine M, we let $M^X(n)$ denote the output of M when given a sampling device to X. Hence, we denote by $(\widetilde{P}, V^X(n))$ the output of a verifier V, when having access to a sampling device for X, after interacting with an arbitrary prover $\widetilde{P}$. Recall that a distribution X is said to be ϵ-far from a set of distribution $\mathcal{D}$ if, for every distribution Y in $\mathcal{D}$, the total variation distance between X and Y is greater than ϵ (i.e., $\sum_z |\Pr[X=z] - \Pr[Y=z]| > 2\epsilon$).

Definition 2.1 (a verifier for the promise problem $(\mathcal{D}, \mathcal{D}')$): *Let $\mathcal{D} = \bigcup_{n\in\mathbb{N}} \mathcal{D}_n$ and $\mathcal{D}' = \bigcup_{n\in\mathbb{N}} \mathcal{D}'_n$ such that $\mathcal{D}_n$ and $\mathcal{D}'_n$ are sets of distributions over $[n]$. A* verifier*, denoted V,* for the promise problem *$(\mathcal{D}, \mathcal{D}')$ is a probabilistic interactive machine that, on input $n \in \mathbb{N}$ and when given access to a sampling device for an unknown distribution X over $[n]$, satisfies the following two conditions after interacting with a potential prover.*

1. The verifier accepts distributions that belong to $\mathcal{D}$: *If X is in $\mathcal{D}_n$, then there exists a prover strategy P such that* $\Pr[(P, V^X(n))=1] \geq 2/3$.
2. The verifier rejects distributions that are in $\mathcal{D}'$: *If X is in $\mathcal{D}'_n$, then for every prover strategy $\widetilde{P}$ it holds that* $\Pr[(\widetilde{P}, V^X(n))=0] \geq 2/3$.

We say that V has sample complexity *$s : \mathbb{N} \to \mathbb{N}$ if on input n it activates the sampling device $s(n)$ times* (i.e., it obtains $s(n)$ samples from X).

Note that an interactive proof of proximity for $\mathcal{D}$ is viewed as a special case in which $\mathcal{D}'$ consists of the set of distributions that are ϵ-far from $\mathcal{D}$.

An alternative formulation of Definition 2.1 specifies the (honest) prover strategy that is used in the completeness condition (i.e., Condition 1). This is called for when wishing to discuss the complexity of such honest prover strategies. Indeed, verifiers that admit an efficient proving strategy (in case X is in $\mathcal{D}$) are of natural interest, and are our main focus. In these cases, we provide these prover strategies with samples of X and consider their sample complexity, which will be denoted $s' : \mathbb{N} \to \mathbb{N}$. Hence, we use the following definition.

Definition 2.2 (an interactive proof for the promise problem $(\mathcal{D}, \mathcal{D}')$): *For $(\mathcal{D}, \mathcal{D}')$ and V as in Definition 2.1, we say that (P, V) is an* interactive proof for *$(\mathcal{D}, \mathcal{D}')$ if P is an interactive machine and Condition 1* (completeness) *holds with P replaced by P^X; that is, for every $n \in \mathbb{N}$ and for every X in $\mathcal{D}_n$, it holds that* $\Pr[(P^X(n), V^X(n))=1] \geq 2/3$. *We say that P has* sample complexity *$s' : \mathbb{N} \to \mathbb{N}$ if on input n it activates the sampling device $s'(n)$ times* (i.e., it obtains $s'(n)$ samples from X).

We stress that V also satisfies Condition 2 (soundness).

We say that (P, V) is a doubly-sublinear interactive proof for $(\mathcal{D}, \mathcal{D}')$ if (1) the sample complexity of V is sublinear in the sample complexity of *deciding* $(\mathcal{D}, \mathcal{D}')$, and (2) the sample complexity of P is sublinear in the sample complexity of *learning* $\mathcal{D}$.

Lastly, we note that the definitions extend to pairs (or tuples) of input distributions, by providing the relevant machines with sampling devices to both (or all) input distributions. For sake of clarity, the resulting definition is spelled out next.

[4] We note that an alternative model in which one is given the sampling device itself emerged in the context of statistical zero-knowledge; see survey [13].

Definition 2.3 (Definition 2.2 adapted to pairs of distributions): *Let $\mathcal{D} = \bigcup_{n\in\mathbb{N}} \mathcal{D}_n$ and $\mathcal{D}' = \bigcup_{n\in\mathbb{N}} \mathcal{D}'_n$ such that $\mathcal{D}_n$ and $\mathcal{D}'_n$ are sets of pairs of distributions over $[n]$. We say that (P,V) is an* interactive proof for *$(\mathcal{D},\mathcal{D}')$ if V and P are interactive machines that, on input $n \in \mathbb{N}$ and when given access to sampling device for a pair of unknown distribution (X,Y) over $[n]$, satisfies the following two conditions.*

1. The verifier accepts distributions that belong to $\mathcal{D}$: *If (X,Y) is in $\mathcal{D}_n$, then it holds that* $\Pr[(P^{X,Y}(n), V^{X,Y}(n))=1] \geq 2/3$.
2. The verifier rejects distributions that are in $\mathcal{D}'$: *If (X,Y) is in $\mathcal{D}'_n$, then for every prover strategy $\widetilde{P}$ it holds that* $\Pr[(\widetilde{P}, V^{X,Y}(n))=0] \geq 2/3$.

We say that V (resp., *P*) *has* sample complexity $s : \mathbb{N} \to \mathbb{N}$ *if on input n it activates each sampling device $s(n)$ times* (i.e., it obtains $s(n)$ samples both from X and from Y).

We stress that $M^{X,Y}$ means that the machine M can activate the sampling devices of both X and Y (i.e., obtains samples from both X and Y). In the actual description of our protocols, we shall use less formal phrases; for example, we will say that the verifier "takes t samples from a distribution" (rather than saying that "it activates the sampling device t times").

3 Proving FARNESS via reduction to an asymmetric tester for EQUALITY

In this section, we use the asymmetric *tester* for EQUALITY provided in [6, Prop. 2.11]. On input parameters n and ϵ, this tester uses m_1 samples from one distribution and m_2 (additional) samples from each of the two distributions, where

$$m_2 = O(\epsilon^{-2} \cdot \max(n/\sqrt{m_1}, \sqrt{n})).$$

Focusing on $m_1 \leq n$, we get $m_2 = O(\epsilon^{-2} n/\sqrt{m_1})$. Hence, this tester, hereafter called the DK-tester, takes $m = m_1 + m_2$ samples from one distribution, and $t = m_2 < m$ samples from the other distribution. Assuming that $m_2 \ll m_1$ (so that $t \ll m$), this means that $t = O(\epsilon^{-2} n/\sqrt{m})$ (equiv., $m = \Omega(\epsilon^{-2} n/t)^2$), where $m < 2m_1 \leq 2n$. In the following proof system, the sample complexity of the verifier is t, whereas the sample complexity of the honest prover is m.

Protocol 3.1 (proof system for FARNESS, using an asymmetric tester for EQUALITY): *On input parameters n and ϵ, the parties proceed as follows.*

1. *The verifier selects at random one of the two distributions, takes a sample of size t from it, and sends the sample to the prover. Denote this sample by S_V.*

   (Recall that $t = \Omega(\sqrt{n}/\epsilon^2)$.)

2. *The honest prover takes a sample of size m from the first distribution. Denoting this sample by S_P, the honest prover invokes the DK-tester with the samples S_V and S_P, and return 1 if the tester accepts and 2 otherwise.*[5]

   (Recall that $m = \max(t, \Omega(\epsilon^{-2} n/t)^2)$.)

[5] Alternatively, this step can be replaced by a more symmetric version. Specifically, in this alternative, the honest prover takes a sample of size m from each of the two distributions. Denoting these two samples by S_1 and S_2, for each $i \in \{1,2\}$, the honest prover invokes the DK-tester with the samples S_V and S_i, and responds with i if the i^{th} invocation accepts. (If both invocations decide identically, the honest prover responds with a random bit.)

3. *The verifier accepts if and only if the prover's answer fits the choice made by the verifier in Step 1; that is, the verifier accepts if it selected the first* (resp., second) *distribution and the prover answered 1* (resp., answered 2).

Soundness (with error probability 1/2) follows from the fact that, when the distributions are identical, the sample S_V reveals no information about the verifier's choice. On the other hand, completeness follows from the completeness and soundness of the DK-tester; that is, completeness (resp., soundness) guarantees that if the verifier chose the first (resp., second) distribution, then (whp) the DK-tester accepts (resp., rejects).

Summary: For every $t = \Omega(\epsilon^{-2}/\sqrt{n})$, the problem ϵ-FARNESS has a proof system in which the verifier's sample complexity is t, the honest prover's sample complexity is $m = O((\epsilon^{-2}n/t)^2)$, and the communication complexity is $O(t \log n)$. Specifically, for every constant $\alpha \in [0, 1/6]$ and $\epsilon > 0$, we can use $t = O(n^{0.5+\alpha}/\epsilon^2)$ and $m = O(n^{1-2\alpha}/\epsilon^2)$. Hence, this yields a doubly-sublinear interactive proof for FARNESS.

The main deficiency of the foregoing result is that it holds only for $t = \Omega(\sqrt{n}/\epsilon^2)$. As stated in the introduction, this deficiency is removed in the following section.

4 Proving FARNESS via a direct construction

In this section, we present a direct construction of an interactive proof system for FARNESS. This proof system establishes Theorem 1.4.

The basic intuition is that ϵ-FARNESS can be decided with a small but noticeable advantage over a coin toss based on two sample-points that are either drawn at random from one of the distributions or are each drawn from one of the two distributions. That is, in the first case we select at random one of the two distributions and take two sample-points from it, whereas in the second case we take a single sample-point from each of the two distributions. As shown next, the probability that these two sample-points are identical is different in the two cases.

Specifically, let us denote the two input distributions by $P = (p_i)_{i\in[n]}$ and $Q = (q_i)_{i\in[n]}$. Then, the probability of a collision in the first case is $\frac{1}{2} \cdot \sum_i p_i^2 + \frac{1}{2} \cdot \sum_i q_i^2$, whereas the probability of collision in the second case is $\sum_i p_i q_i$. The key observation is that

$$\frac{1}{2} \cdot \sum_{i\in[n]} p_i^2 + \frac{1}{2} \cdot \sum_{i\in[n]} q_i^2 \;-\; \sum_{i\in[n]} p_i q_i \;=\; \frac{1}{2} \cdot \sum_{i\in[n]} (p_i - q_i)^2 \;\geq\; 0, \tag{1}$$

which is greater than $2\epsilon^2/n$ in case $\sum_{i\in[n]} |p_i - q_i| > 2\epsilon$ (i.e., if P is ϵ-far from Q).

It is tempting to say that $O(n/\epsilon^2)$ pairs of sample-points suffice in order to distinguish between the case that P is ϵ-far from Q and the case that $P = Q$. This is indeed true if $p_i + q_i = O(1/n)$ for every $i \in [n]$, but this condition does not hold in general. Furthermore, recall that the sample complexity of deciding FARNESS is actually $\Omega(n^{2/3})$, which is indeed due to the violation of the foregoing condition.

For $t, m \in \mathbb{N}$, we aim at a proof system in which the verifier has sample complexity t and the honest prover has sample complexity $m \gg t$, we first observe that $m = \Omega(n^{2/3})$ must hold, since a tester can emulate the interaction between the verifier and the honest prover. Next, we observe that if we use a prover of sample complexity $O(m/\epsilon^2)$, then we may assume that $p_i + q_i = o(1/m)$ for

every i, because i's that violate this condition can be either ignored (if $|p_i - q_i| = o(\epsilon^2/m)$) or used directly (if $|p_i - q_i| = \Omega(\epsilon^2/m)$).[6] We shall show that, under this assumption (i.e., $p_i + q_i = o(1/m)$ for every i), FARNESS has a proof system in which the verifier uses t samples and the honest prover uses $m \gg t$ samples, provided that $\sqrt{t} \cdot m = \Omega(n/\epsilon^2)$. (In contrast, satisfying $t \cdot m = \Omega(n/\epsilon^2)$ does not suffice.)[7]

4.1 The basic protocol

For parameters m and t, the protocol is as follows, assuming that each element in each of the distributions appears in it with probability at most $0.01/m$ (or so).

Protocol 4.1 (interactive proof system for FARNESS, with parameters t and m): *The parties proceed as follows.*

1. *The verifier sends t samples, denoted $s_1, ..., s_t$, to the prover, where each sample is selected at random from one of the two distributions. That is, for each $j \in [t]$, the verifier selects $b_j \in \{1, 2\}$ uniformly at random, and sets s_j to be a sample from the b_j^{th} distribution.*

2. *The honest prover takes a sample of size m from each of the two distributions. For each $(j, k) \in [t] \times [m]$, let $\chi_{j,k}(b) = 1$ if s_j appears as the k^{th} element in its b^{th} sample* (and $\chi_{j,k}(b) = 0$ otherwise). *For each $j \in [t]$, the prover sends $v_j \stackrel{\text{def}}{=} \sum_{k \in [m]} \chi_{j,k}(1) - \sum_{k \in [m]} \chi_{j,k}(2)$ to the verifier.*[8]

   Intuitively, v_j represents a vote regarding b_j; a positive (resp., negative) *vote represents a surplus of collisions with the first* (resp., second) *sample. If the two distributions are far apart, then we expect that v_j be positively correlated with $(-1)^{b_j - 1}$.*

3. *The verifier accepts if $v \stackrel{\text{def}}{=} \sum_{j \in [t]} (-1)^{b_j - 1} \cdot v_j > 0$, rejects if $v < 0$, and accepts with probability $1/2$ if $v = 0$.*

   That is, the verifier adds v_j to the sum if $b_j = 1$, and adds $-v_j$ to the sum otherwise (i.e., if $b_j = 2$).

We first observe that if the two distributions are identical, then, no matter how the prover behaves, the verifier accepts with probability $1/2$. This is the case because the s_j's reveal no information about the b_j's. Hence, for each choice of $(v_1, ..., v_t)$, for some $p \in [0, 0.5]$, the value of $\sum_{j \in [t]} (-1)^{b_j - 1} \cdot v_j$ is positive with probability p, negative with probability p, and 0 with probability $1 - 2p$.

Next, we shall show that if the distributions are ϵ-far apart, then the honest prover convinces the verifier with high constant probability. Intuitively, as shown next, if the two distributions are far part, then collisions with the same distribution are more likely than collisions with the other distribution. Hence, each s_j is more likely to collide with the honest prover's sample of the b_j^{th}

[6] Details are postponed to the end of Section 4.1.

[7] The failure of the latter intuitive requirement (i.e., $t \cdot m = \Omega(n/\epsilon^2)$) is not surprising given that a tester can emulate the interaction between the verifier and the honest prover, whereas (by [2]) a tester must have sample complexity $\Omega((n/\epsilon^2)^{2/3})$. Indeed, note that $t = m = o((n/\epsilon^2)^{2/3})$ violates $\sqrt{t} \cdot m = \Omega(n/\epsilon^2)$ but not $t \cdot m = O(n/\epsilon^2)$).

[8] It may be more natural to have the honest prover sends the index of the sample with which s_j collides, but this presupposes a single collision. The foregoing step avoids this assumption, and provides the full tally of collisions for each s_j.

distribution than with the sample of the $(3-b_j)^{\text{th}}$ distribution. The quantitative version of this assertion depends not only on the distance between the two distributions but also on an upper bound on the max-norm of the two distributions. The latter bound is captured by the second condition below, where $\tau > 0$ is a free parameter to be set to a small constant.

Lemma 4.2 (completeness of Protocol 4.1): *Denoting the first distribution by $P = (p_i)_{i\in[n]}$ and the second distribution by $Q = (q_i)_{i\in[n]}$, suppose that the following three conditions hold:*

- *The total variation distance betwen P and Q is greater than ϵ; that is, $\sum_{i\in[n]} |p_i - q_i| > 2\epsilon$.*
- *For each $i \in [n]$, it holds that $p_i, q_i < \tau/m$.*
- $\sqrt{t} \cdot m \geq 2n/\epsilon^2$.

Then, the verifier accepts with probability $1 - O(\tau)$.

Looking at the third condition, note that in the uninteresting (for us) case of $t = m$ it implies $m \geq (2n/\epsilon^2)^{2/3}$, which matches the sample complexity bound in [8]. More interestingly, for any $\beta \in (0, 1/3)$, we can use $t = n^{2\beta}$ and $m = O(n^{1-\beta}/\epsilon^2)$.

Proof: Let s'_k (resp., s''_k) denote the k^{th} element in the sample of the first (resp., second) distribution taken by the prover. Note that $\Pr[s'_k = s_j | b_j = 1] = \sum_i p_i^2$ and $\Pr[s'_k = s_j | b_j = 2] = \sum_i p_i q_i$; likewise, $\Pr[s''_k = s_j | b_j = 2] = \sum_i q_i^2$ and $\Pr[s''_k = s_j | b_j = 1] = \sum_i q_i p_i$. Also note that each of the three sums is upper-bounded by τ/m, since $p_i, q_i < \tau/m$ for each i.

For each $j \in [t]$ and $k \in [m]$, recall that $\chi_{j,k}(1) = 1$ if $s_j = s'_k$ (and $\chi_{j,k}(1) = 0$ otherwise), and $\chi_{j,k}(2) = 1$ if $s_j = s''_k$, where s_j is the j^{th} element in the sample taken by the verifier and that this element is taken from the b_j^{th} distribution. Hence, $\chi_{j,k}(b_j)$ is the indicator of the event in which either $s_j = s'_k$ and $b_j = 1$ or $s_j = s''_k$ and $b_j = 2$. (Likewise, $\chi_{j,k}(3 - b_j)$ is the indicator of the event in which either $s_j = s''_k$ and $b_j = 1$ or $s_j = s'_k$ and $b_j = 2$.) Recalling that b_j is uniformly distributed in $\{1, 2\}$, it follows that

$$\begin{aligned}
\Pr[\chi_{j,k}(b_j) = 1] &= \frac{1}{2} \cdot \sum_{i\in[n]} p_i^2 + \frac{1}{2} \cdot \sum_{i\in[n]} q_i^2 \\
\Pr[\chi_{j,k}(3 - b_j) = 1] &= \frac{1}{2} \cdot \sum_{i\in[n]} p_i q_i + \frac{1}{2} \cdot \sum_{i\in[n]} q_i p_i
\end{aligned}$$

Now, let

$$\zeta_{j,k} = (-1)^{b_j - 1} \cdot (\chi_{j,k}(1) - \chi_{j,k}(2)) = \chi_{j,k}(b_j) - \chi_{j,k}(3 - b_j). \tag{2}$$

In particular, $\zeta_{j,k} = 1$ if $\chi_{j,k}(b_j) = 1$ and $\chi_{j,k}(3 - b_j) = 0$, whereas $\zeta_{j,k} = -1$ if $\chi_{j,k}(3 - b_j) = 1$ and $\chi_{j,k}(b_j) = 0$. Actually, the case $\chi_{j,k}(b_j) = \chi_{j,k}(3 - b_j) = 0$ is far more likely, and in this case $\zeta_{j,k} = 0$. (Indeed, $\zeta_{j,k} = 0$ also holds in the more rare case in which $\chi_{j,k}(b_j) = \chi_{j,k}(3 - b_j) = 1$.) Combining the foregoing, we get

$$\begin{aligned}
\text{Exp}[\zeta_{j,k}] &= \text{Exp}[\chi_{j,k}(b_j)] - \text{Exp}[\chi_{j,k}(3 - b_j)] \\
&= \frac{1}{2} \cdot \left(\sum_{i\in[n]} p_i^2 + \sum_{i\in[n]} q_i^2 \right) - \sum_{i\in[n]} p_i q_i \\
&= \frac{1}{2} \cdot \sum_{i\in[n]} (p_i - q_i)^2,
\end{aligned}$$

which is greater than $2\epsilon^2/n$ since $\sum_i |p_i - q_i| > 2\epsilon$. Letting $\mu \stackrel{\text{def}}{=} \text{Exp}[\zeta_{j,k}]$, we have

$$\text{Exp}\left[\sum_{(j,k)\in[t]\times[m]} \zeta_{j,k}\right] = t \cdot m \cdot \mu. \tag{3}$$

We shall show that, with high probability, $\sum_{(j,k)\in[t]\times[m]} \zeta_{j,k} > 0$. For starters, recall that $\text{Var}[\zeta_{j,k}] < \text{Exp}[\zeta_{j,k}^2]$ and observe that $\text{Exp}[\zeta_{j,k}^2] = \Pr[|\zeta_{j,k}|=1]$, whereas

$$\begin{aligned}
\Pr[|\zeta_{j,k}|=1] &\leq \Pr[\chi_{j,k}(b_j)=1] + \Pr[\chi_{j,k}(3-b_j)=1] \\
&= \sum_{i\in[n]} \left(\frac{1}{2}\cdot p_i^2 + \frac{1}{2}\cdot q_i^2\right) + \sum_{i\in[n]} p_i q_i \\
&= \frac{1}{2}\cdot \sum_{i\in[n]} (p_i + q_i)^2
\end{aligned}$$

which is at most $2\tau/m$ since $|p_i + q_i| \leq 2\tau/m$. Hence, $\text{Var}[\zeta_{j,k}] < 2\tau/m$. Applying Chebyshev's inequality, we have

$$\Pr\left[\sum_{(j,k)\in[t]\times[m]} \zeta_{j,k} < tm\cdot\mu/2\right] \leq \frac{\text{Var}[\sum_{(j,k)\in[t]\times[m]} \zeta_{j,k}]}{(tm\cdot\mu/2)^2} \tag{4}$$

Letting $\overline{\zeta}_{j,k} \stackrel{\text{def}}{=} \zeta_{j,k} - \mu$, we get

$$\begin{aligned}
\text{Var}\left[\sum_{(j,k)\in[t]\times[m]} \zeta_{j,k}\right] &= \text{Exp}\left[\left(\sum_{(j,k)\in[t]\times[m]} \overline{\zeta}_{j,k}\right)^2\right] &(5)\\
&= \sum_{(j_1,k_1),(j_2,k_2)\in[t]\times[m]} \text{Exp}[\overline{\zeta}_{j_1,k_1}\cdot\overline{\zeta}_{j_2,k_2}] \\
&= \sum_{(j,k)\in[t]\times[m]} \text{Var}[\zeta_{j,k}] + \sum_{(j_1,k_1,j_2,k_2)\in T'\cup T''} \text{Exp}[\overline{\zeta}_{j_1,k_1}\cdot\overline{\zeta}_{j_2,k_2}] &(6)
\end{aligned}$$

where $T' \stackrel{\text{def}}{=} \{(j,k_1,j,k_2) \in ([t]\times[m])^2 : k_1 \neq k_2\}$ and $T'' \stackrel{\text{def}}{=} \{(j_1,k,j_2,k) \in ([t]\times[m])^2 : j_1 \neq j_2\}$. Recalling that $\text{Var}[\zeta_{j,k}] < 2\tau/m$, we have

$$\sum_{(j,k)\in[t]\times[m]} \text{Var}[\zeta_{j,k}] < 2t\cdot\tau. \tag{7}$$

For each $(j,k_1,j,k_2) \in T'$, we have

$$\begin{aligned}
\text{Exp}[\overline{\zeta}_{j,k_1}\cdot\overline{\zeta}_{j,k_2}] &< \text{Exp}[\zeta_{j,k_1}\cdot\zeta_{j,k_2}] \\
&\leq \Pr[\zeta_{j,k_1} = \zeta_{j,k_2} \neq 0],
\end{aligned}$$

where the last inequality relies on $\zeta_{j,k} \in \{-1,0,1\}$. The key observation is that $\Pr[\zeta_{j,k_1} = \zeta_{j,k_2} \neq 0]$ is upper-bounded by $\sum_i \frac{p_i+q_i}{2} \cdot (p_i+q_i)^2$, since this event requires s_j to collide with both the k_1^{th} and k_2^{th} sample taken by the prover (from either of the two distributions). Using $\sum_i \frac{p_i+q_i}{2} \cdot (p_i+q_i)^2 \leq \max_i\{p_i+q_i\}^2 \leq (2\tau/m)^2$, we get

$$\sum_{(j,k_1,j,k_2)\in T'} \text{Exp}[\overline{\zeta}_{j,k_1} \cdot \overline{\zeta}_{j,k_2}] \leq t \cdot m^2 \cdot (2\tau/m)^2. \tag{8}$$

Likewise, for each $(j_1,k,j_2,k) \in T''$ we have

$$\begin{aligned} \text{Exp}[\overline{\zeta}_{j_1,k} \cdot \overline{\zeta}_{j_2,k}] &< \text{Exp}[\zeta_{j_1,k} \cdot \zeta_{j_2,k}] \\ &\leq \Pr[\zeta_{j_1,k} = \zeta_{j_2,k} \neq 0], \end{aligned}$$

which is at most $(2\tau/m)^2$. Hence,

$$\sum_{(j_1,k,j_2,k)\in T'} \text{Exp}[\overline{\zeta}_{j_1,k} \cdot \overline{\zeta}_{j_2,k}] \leq t^2 \cdot m \cdot (2\tau/m)^2. \tag{9}$$

Combining Eq. (5) & (6) with Eq. (7)–(9), we get

$$\text{Var}\left[\sum_{(j,k)\in[t]\times[m]} \zeta_{j,k}\right] \leq 2t \cdot \tau + 4t \cdot \tau^2 + 4t^2 \cdot \tau^2/m. \tag{10}$$

Combining Eq. (4) with Eq. (10), we get

$$\begin{aligned} \Pr\left[\sum_{(j,k)\in[t]\times[m]} \zeta_{j,k} < tm \cdot \mu/2\right] &\leq \frac{(2t + 4t\tau + (4t^2\tau/m)) \cdot \tau}{(tm\mu/2)^2} \\ &= \frac{O(t) \cdot \tau}{(tm\mu)^2} + \frac{O(t^2/m) \cdot \tau^2}{(tm\mu)^2} \end{aligned}$$

which is $O(t\tau/(tm\mu)^2)$ provided $t = O(m/\tau)$. Assuming $t \leq (tm\mu)^2$ (equiv., $\sqrt{t} \cdot m \geq 1/\mu$), we get an error probability of $O(\tau)$. Recalling that $\mu \geq \epsilon^2/2n$, we need $\sqrt{t} \cdot m \geq 2n/\epsilon^2$, which is gauarntred by the third condition in the lemma. ■

Getting rid of the assumption $\max_i\{\max(p_i,q_i)\} \leq \tau/m$**.** Essentially, the honest prover can identify i such that $p_i > \tau/m$ (or $q_i > \tau/m$) and act accordingly. Specifically, such i's can be ignored by the honest prover (if their contribution to the distance between P and Q is small) or used by it (otherwise), depending on the case. Details follow.

Let $H = \{i \in [n] : \max(p_i,q_i) > \tau/m\}$ and $D = \{i \in H : |p_i - q_i| > \tau\epsilon/4m\}$. If $\sum_{i\in D} |p_i - q_i| > \epsilon/4$, then the prover notifies the verifier that this is the case. In that case, rather than executing Protocol 4.1, the verifier selects at random one of the two distributions, sends the prover a single sample from this distribution, and asks the prover to tell the origin of this sample point. The prover guesses that the sample point came from P (resp., Q) if hits the set $S = \{i \in D : p_i > q_i\}$ (resp.,

$D \setminus S$), and answers at random otherwise (i.e., if the sample point is not in D). Observe that the probability that the honest prover is correct equals

$$\begin{aligned}
&\frac{1}{2}\cdot\left(\sum_{i\in S} p_i + \sum_{i\in[n]\setminus D} p_i\cdot 0.5\right) + \frac{1}{2}\cdot\left(\sum_{i\in D\setminus S} q_i + \sum_{i\in[n]\setminus D} q_i\cdot 0.5\right)\\
&= \frac{1}{4}\cdot\sum_{i\in[n]\setminus D}(p_i+q_i) + \frac{1}{4}\cdot\sum_{i\in D}(p_i+q_i) + \frac{1}{4}\cdot\sum_{i\in S}(p_i-q_i) + \frac{1}{4}\cdot\sum_{i\in D\setminus S}(q_i-p_i)\\
&= \frac{1}{4}\cdot\sum_{i\in[n]}(p_i+q_i) + \frac{1}{4}\cdot\sum_{i\in D}|p_i-q_i|\\
&> \frac{1}{2}+\frac{1}{4}\cdot\frac{\epsilon}{4}
\end{aligned}$$

Amplification by $O(1/\epsilon^2)$ repetitions is employed in order to boost the gap between the completeness and soundness bounds from $\Omega(\epsilon)$ to any desired constant level (e.g., 1/3).

Otherwise (i.e., $\sum_{i\in D}|p_i-q_i| \leq \epsilon/4$), when executing Protocol 4.1, the prover just ignores each $i \in H$ (where ignoring means reporting no collisions with such a sample), while relying on the fact that their total contribution to $\sum_i |p_i - q_i|$ is at most

$$\sum_{i\in D}|p_i-q_i| \;+\; \sum_{i\in H\setminus D}|p_i-q_i| \;<\; \frac{\epsilon}{4} \;+\; \frac{1}{\tau/m}\cdot\frac{\tau\epsilon}{4m}$$

which equals $\epsilon/2$.

We warn that the foregoing description is a bit hand-waved, because the prover does not know whether a given i is in H and ditto in D (and in S). Nevertheless, approximate decisions are good enough. Specifically, for each $i \in [n]$, it suffices to estimate p_i (resp., q_i) upto an additive deviation of (say) $0.01\tau\epsilon\cdot\max(p_i, 1/m)$ (resp., $0.01\tau\epsilon\cdot\max(q_i, 1/m)$). Such an estimate can be obtained if the honest prover uses a sample of size $\widetilde{O}(m/\epsilon^2)$ rather than of size m. Hence, we obtain

Theorem 4.3 (an interactive proof system for ϵ-FARNESS:) *For every $t = \Omega(1/\epsilon^2)$ and $m = \Omega(\epsilon^{-2}\cdot n/\sqrt{t})$ such that $t \leq m$, there exists an interactive proof system for ϵ-FARNESS in which the verifier has sample complexity t, the honest prover has sample complexity $\widetilde{O}(m/\epsilon^2)$, and the communication complexity is $O(t\cdot\log n)$.*

This falls short of establishing Theorem 1.4, because the prover is not laconic (i.e., the prover sends $O(t\log n)$ bits).[9] This gap is closed in the next subsection.

4.2 A reduction yielding a laconic prover

Suppose that we have an arbitrary interactive proof system for FARNESS in which the verifier uses t samples from each distribution, and the prover uses m samples from each distribution. For simplicity, assume that this proof system has error probability at most 1/3 in both completeness and soundness conditions. We construct an interactive proof system of similar sample complexities, but

[9] We can easily improve the length of the message sent by the prover (in Protocol 4.1) by observing that almost all v_j's are zero. Hence, rather than sending $(v_1, ..., v_t)$, the prover may send $(i, v_i)_{i:v_i\neq 0}$, which means that it need only sent $O(\epsilon^{-2}\log n)$ bits.

in which the prover sends a single bit. The system proceeds as follows, when the input distributions are $P = (p_i)_{i\in[n]}$ and $Q = (q_i)_{i\in[n]}$.

1. The (new) verifier selects one of the two distributions, uniformly at random, takes a sample of size t from it, and sends the sample to the prover. Denote this sample S_V.

   That is, the (new) verifier selects $b \in \{1,2\}$ uniformly at random, and lets S_V be a t-size sample of P if $b = 1$, and a t-size sample of Q otherwise (i.e., if $b = 2$).

2. The (new) honest prover emulates both parties in the original interactive proof system, while providing the original verifier with S_V as the first sample, instead of a t-sized sample of P. We stress that the second t-sized sample given to the original verifier is always drawn from Q, whereas the two m-sized samples given to the original honest prover are drawn from P and Q, respectively.

   Hence, when $b = 1$, the (new) honest prover emulates the execution of the original interactive proof system on the input (P, Q), and otherwise (i.e., when $b = 2$) it emulates an execution in which the original verifier gets two t-sized samples of Q.

   If the original verifier (emulated by the new honest prover) accepts, then the (new) honest prover sends 1 to the (new) verifier. Otherwise (i.e., the emulated original verifier rejects), it sends 2.

3. The (new) verifier accepts if and only if the prover guessed correctly the identity of the distribution it chose in Step 1 (i.e., if the (new) prover's message equals b).

Evidently, if $P = Q$, then the (new) verifier accepts with probability at most $1/2$, because in this case S_V carries no information about the (new) verifier's choice (i.e., b). On the other hand, we claim that if P is far from Q, then the probability that the verifier accepts when interacting with the honest prover is at least $2/3$. Specifically, letting η_c and η_s denote the completeness and soundness error of the original protocol, we claim that the completeness error of the new protocol is $\frac{1}{2}\cdot\eta_c + \frac{1}{2}\cdot\eta_s$. This is show by considering the initial choice of the new verifier.

- With probability $1/2$, the sample S_V is taken from P (equiv., $b = 1$). In this case, the new honest prover emulates the real execution of the original interactive proof system on samples drawn from P and Q. By the completeness condition of the original proof system, the original verifier accepts with probability at least $1 - \eta_c$ (since P is far from Q), and the new honest prover sends 1, which leads the new verifier to accept.

- Otherwise (i.e., with probability $1/2$), the sample S_V is taken from Q (equiv., $b = 2$). In this case, the new honest prover emulates an execution of the original proof system when the original verifier gets two t-sized samples of Q. By the soundness condition of the original proof system, the original verifier rejects with probability at least $1 - \eta_s$ (regardless of the behavior of the (original) prover), and the new honest prover sends 2, which leads the new verifier to accept.

Hence, the new verifier accepts with probability $\frac{1}{2}\cdot(1-\eta_c) + \frac{1}{2}\cdot(1-\eta_s)$. It follows that the new proof system has completeness error $(\eta_c + \eta_s)/2 \leq 1/3$, and soundness error $1/2$. Instantiating the resulting proof system with the proof system of Theorem 4.3 (and using error reduction), we establish Theorem 1.4.

5 A proof system for the complement of uniformity

In this section, we prove Theorem 1.3. Recall that we wish to accept distributions that are ϵ-far from U_n and reject U_n (i.e., the uniform distribution over $[n]$). Here we use the fact that a distribution of the former type has collision probability exceeding $(1+4\epsilon^2)/n$, whereas U_n has collision probability $1/n$.

Aiming at an interactive proof system in which the verifier has sample complexity t and the honest prover has sample complexity $m \gg t$, we first observe that $m = \Omega(n^{1/2})$, since a tester can emulate the interaction between the verifier and the honest prover. Analogously to Section 4, we also observe that we may assume that the input distribution $P = (p_i)_{i\in[n]}$ satisfies $p_i \leq 1/m$ for every i, because i's that violate this condition can be either ignored (if $|p_i - (1/n)| = o(\epsilon^2/m)$) or used directly (if $|p_i - (1/n)| = \Omega(\epsilon^2/m)$).

We shall first show that, under this assumption (i.e., $p_i \leq 1/m$ for every i), `FARfromUNI` has an interactive proof system in which the verifier uses t samples and the honest prover uses $m \gg t$ samples, provided that $t \cdot m = \Omega(n/\epsilon^4)$. (Later, as in the case of `FARNESS`, we shall get rid of the foregoing assumption, and also derive a laconic prover.)

Protocol 5.1 (interactive proof for `FARfromUNI`, with parameters t and m): *The parties proceed as follows.*

1. *The verifier sends t samples, denoted $s_1, ..., s_t$, to the prover, where each sample is selected at random either from P or from the uniform distribution. That is, for each $j \in [t]$, the verifier selects $b_j \in \{0,1\}$ uniformly at random, and sets s_j to be a sample of P if $b_j = 1$ and uses a uniformly distributed $s_j \in [n]$ otherwise.*

2. *The honest prover takes a sample of size m from P. For each $j \in [t]$, the honest prover sends the number of occurrences of s_j in this sample.[10] Let us denote this tally by v_j.*

3. *The verifier accepts if $\sum_{j\in[t]}(-1)^{b_j-1} \cdot v_j > 0$, and rejects otherwise.*

   That is, the j^{th} term in the sum is v_j if $b_j = 1$, and is $-v_j$ otherwise (i.e., if $b_j = 0$).

   Alternatively, for sake of easier analysis, we may have the verifier accept if and only if the following three conditions hold.

   (a) $\sum_{j\in[t]:b_j=1} v_j > t_1 m \cdot (1+3\epsilon^2)/n$, *where* $t_1 \stackrel{\text{def}}{=} |\{j \in [t] : b_j = 1\}|$.
   (b) $\sum_{j\in[t]:b_j=0} v_j < (t-t_1)m \cdot (1+\epsilon^2)/n$.
   (c) *For every $j \in [t]$, it holds that $v_j \in \{0,1,...,O(\log n)\}$.*

If P equals the uniform distribution, then the s_j's reveal no information about the b_j's. In this case, assuming $t = \omega(\epsilon^{-2} \cdot \log(1/\epsilon))$ (equiv., $\epsilon^2 \cdot t = \omega(\log t)$), for every choice of $v_1, ..., v_t \in \{0,1,...,O(\log n)\}$ and $\sigma \in \{0,1\}$, for a random choice of $b_1, ..., b_t \in \{0,1\}$, the sum $\sum_{j\in[t]:b_j=\sigma} v_j$ is concentrated around its expectation (which is $\sum_{j\in[t]} v_j/2$).[11] However, Conditions 3a and 3b are in conflict (since the former requires a noticeably larger sum than the latter). It follows that if P is uniform over $[n]$, then the verifier accepts with probability $o(1)$.

[10] Again, it may be more natural to have the honest prover indicate whether a collision occurred, but this presupposes a single collision.

[11] In particular, wvhp, $t_1 = (1 \pm o(1)) \cdot t/2$.

We shall show that if $P = (p_i)_{i\in[n]}$ is ϵ-far from the uniform distribution, then the honest prover convinces the verifier with probability $1-o(1)$. Essentially, it suffices to show that, for every distribution $Q = (q_i)_{i\in[n]}$ (where we actually care about P and about the uniform distribution), *with high probability, the number of collisions between $t' \approx t/2$ samples of P and m samples of Q approximates $t'm \cdot \sum_i p_i q_i$*, provided the latter is $\Omega(1/n)$. Applying the claim with $Q = P$, it follows that Condition 3a is satisfied with high probability, because $\sum_i p_i^2 > (1+4\epsilon^2)/n$. Likweise, applying the claim with $Q = U_n$, it follows that Condition 3b is satisfied with high probability, because $\sum_i p_i/n = 1/n$. (Furthermore, using $p_i \leq 1/m$ for every $i \in [n]$, Condition 3c is satisfied with high probability.) For sake of good order, we restate the claim before proving it.

Claim 5.2 (on approximating the number of collisions): *Let $P = (p_i)_{i\in[n]}$ and $Q = (q_i)_{i\in[n]}$ be two distributions such that $\sum_{i\in[n]|} p_i q_i = \Omega(1/n)$ and $p_i, q_i < 1/m$ for every $i \in [n]$. Then, for every $\eta > 0$, with probability $1 - O(n/(\eta^2 \cdot tm))$, the number of collisions between a t-sized sample of P and an m-sized sample of Q is $(1 \pm \eta) \cdot tm \cdot \sum_i p_i q_i$.*

Proof: For each $j \in [t]$ and $k \in [m]$, let $\zeta_{j,k} = 1$ if the j^{th} sample of the verifier (i.e., s_j) equals the k^{th} sample of the prover. We shall show that, with high probability, $\sum_{(j,k)\in[t]\times[m]} \zeta_{j,k} = (1 \pm \eta) \cdot \sum_i p_i q_i$. Specifically, letting $\mu \stackrel{\text{def}}{=} \text{Exp}[\zeta_{j,k}] = \sum_i p_i q_i = \Omega(1/n)$ and applying Chebyshev's inequality, we have

$$\Pr\left[\left|\sum_{(j,k)\in[t]\times[m]} \zeta_{j,k} \;-\; tm\cdot\mu\right| \;>\; \eta\cdot tm\cdot\mu\right] \leq \frac{\text{Var}[\sum_{(j,k)\in[t]\times[m]}\zeta_{j,k}]}{(\eta\cdot tm\cdot\mu)^2} \tag{11}$$

The analysis proceeds very much as in the proof of Lemma 4.2, where the exception is that in the current case we have better upper bounds on the variance of $\zeta_{j,k}$ as well as on the covariances of ζ_{j_1,k_1} and ζ_{j_2,k_2}. Specifically, $\text{Var}[\zeta_{j,k}] < \text{Exp}[\zeta_{j,k}^2] = \text{Exp}[\zeta_{j,k}] = \mu$, whereas $\text{Exp}[\zeta_{j_1,k_1}\zeta_{j_2,k_2}]$ equals $\Pr[\zeta_{j_1,k_1} = \zeta_{j_2,k_2} = 1]$, which in turn is upper-bounded by $\mu \cdot \max_i\{\max(p_i, q_i)\} \leq \mu/m$ (provided $(j_1, k_1) \neq (j_2, k_2)$). Hence, we get

$$\text{Var}\left[\sum_{(j,k)\in[t]\times[m]} \zeta_{j,k}\right] \;\leq\; mt\cdot\mu + (m^2t + t^2m)\cdot\mu/m \tag{12}$$

which equals $(1+1+(t/m))\cdot mt\cdot\mu \leq 3mt\cdot\mu$ (since $t \leq m$). Combining this with Eq. (11), we get a error probability bound of

$$\begin{aligned}\frac{\text{Var}[\sum_{(j,k)\in[t]\times[m]}\zeta_{j,k}]}{(\eta\cdot tm\cdot\mu)^2} &\leq \frac{3mt\cdot\mu}{(\eta\cdot tm\cdot\mu)^2}\\ &= \frac{3}{\eta^2\cdot tm\cdot\mu}\end{aligned}$$

and the claim follows since $\mu = \Omega(1/n)$. ■

Conclusion. Using Claim 5.2 (with $\eta = \epsilon^2$), we obtain the desired interactive proof system (i.e., the verifier uses $t = \Theta(\epsilon^{-2}\log n)$ samples whereas the honest prover uses $m = \Theta(\epsilon^{-4}n/t)$ samples),

under the assumption that $p_i \leq 1/m$ for every $i \in [n]$.[12], Next, we get rid of the latter assumption by applying the same strategy as in the end of Section 4.1, where here $\tau = 1$ (and $q_i = 1/n$ for every i). Hence, we obtain

Theorem 5.3 (an interactive proof for `FARfromUNI`:) *For every $t = \omega(\epsilon^{-2} \cdot \log(1/\epsilon))$ and $m = \mathrm{poly}(1/\epsilon) \cdot n/t$ such that $t \leq m$, there exists an interactive proof system for ϵ-`FARfromUNI` in which the verifier has sample complexity t, the honest prover has sample complexity $\widetilde{O}(m/\epsilon^2)$, and the communication complexity is $O(t \cdot \log n)$.*

Lastly, we can obtain a laconic prover by following the strategy of Section 4.2. Specifically, here the new verifier selects S_V to be either a t-sized sample of P or a t-sized sample of U_n; that is, if $b = 1$ then S_V is a sample of the input distribution P and otherwise it is a sample of the uniform distribution. As in Section 4.2, the new honest prover sends 1 if and only if the original interactive proof (emulated by it) accepts. This establishes Theorem 1.3.

Generalization to being far from any fixed distribution. Using the techniques in [9, Sec. 11.2.2], for any fixed distribution Q (over $[n]$), the foregoing interactive proofs can be generalized to interactive proofs for the problem of being far from Q. Specifically, this is done by using the reduction of testing equality to a fixed distribution (i.e., to Q) to testing uniformity, where these reductions operate by distance preserving "filters" (see [9, Sec. 11.2.2]).

References

[1] Noga Amir, Oded Goldreich and Guy N. Rothblum. Doubly Sub-linear Interactive Proofs of Proximity. In *16th ITCS*, LIPIcs, Volume 325, pages 6:1–6:25, 2025.

[2] Tugkan Batu and Clement L. Canonne. Generalized Uniformity Testing. In *58th FOCS*, pages 880–889, 2017.

[3] Tugkan Batu, Lance Fortnow, Eldar Fischer, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *42nd IEEE Symposium on Foundations of Computer Science*, pages 442–451, 2001.

[4] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. *Journal of the ACM*, Vol. 60 (1), pages 4:1–4:25, 2013. Preliminary version in *41st FOCS*, pages 259–269, 2000.

[5] Alessandro Chiesa and Tom Gur. Proofs of Proximity for Distribution Testing. In *9th ITCS*, LIPIcs, Volume 94, pages 53:1–53:14, 2018.

[6] Ilias Diakonikolas and Daniel Kane. A New Approach for Testing Properties of Discrete Distributions. In *ECCC*, TR16-074, 2016.

[7] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Collision-Based Testers are Optimal for Uniformity and Closeness. *ECCC*, TR16-178, 2016.

[12] It may be possible to improve the dependence on ϵ by generalizing the better treatment of [7] (see [10]), which refers to collisions within a single sample.

[8] Ilias Diakonikolas, Daniel Kane, and Alistair Stewart. Sharp Bounds for Generalized Uniformity Testing. In *ECCC*, TR17-132, 2017.

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

[10] Oded Goldreich. On the Optimal Analysis of the Collision Probability Tester (an Exposition). In *Computational Complexity and Property Testing*, LNCS 12050, pages 296–305, 2020.

[11] Oded Goldreich and Johan Hastad. On the Complexity of Interactive Proofs with Bounded Communication. *Information Processing Letters*, Vol. 67 (4), pages 205–214, 1998.

[12] Oded Goldreich and Guy N. Rothblum. On doubly-efficient interactive proofs for distributions. In *ECCC*, TR25-200, 2025.

[13] Oded Goldreich and Salil P. Vadhan. On the Complexity of Computational Problems Regarding Distributions. In *Studies in Complexity and Cryptography*, LNCS 6650, pages 390–405, Springer, 2011.

[14] Shafi Goldwasser, Yael Tauman Kalai, and Guy N. Rothblum. Delegating Computation: Interactive Proofs for Muggles. *Journal of the ACM*, Vol. 62 (4), pages 27:1–27:64, 2015. Preliminary version in *40th STOC*, 2008.

[15] Tal Herman and Guy N. Rothblum. Verifying the unseen: interactive proofs for label-invariant distribution properties.

[16] Tal Herman and Guy N. Rothblum. Doubley-Efficient Interactive Proofs for Distribution Properties. In *64th FOCS*, pages 743–751, 2023.

[17] Tal Herman and Guy N. Rothblum. Proving Natural Distribution Properties is Harder than Testing Them. In *ECCC*, TR25-152, 2025.

[18] Guy N. Rothblum, Salil Vadhan, and Avi Wigderson. Interactive Proofs of Proximity: Delegating Computation in Sublinear Time. In *45th ACM Symposium on the Theory of Computing*, pages 793–802, 2013.

[19] Roei Tell. On Being Far from Far and on Dual Problems in Property Testing. In *ECCC*, TR15-072, 2015. Extended abstract in *7th ITCS*, pages 103–110, 2016.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il