Electronic Colloquium on Computational Complexity, Report No. 141 (2024)

Public Coin Interactive Proofs for Label-Invariant Distribution Properties

Tal Herman*
Weizmann Institute of Science

September 12, 2024

Abstract

Assume we are given sample access to an unknown distribution D over a large domain $[N]$. An emerging line of work has demonstrated that many basic quantities relating to the distribution, such as its distance from uniform and its Shannon entropy, despite being hard to approximate through the samples only, can be *efficiently and verifiably* approximated through interaction with an untrusted powerful prover, that *knows* the entire distribution [Herman and Rothblum, STOC 2022, FOCS 2023]. Concretely, these works provide an efficient proof system for approximation of any label-invariant distribution quantity (i.e. any function over the distribution that's invariant to a re-labeling of the domain $[N]$).

In our main result, we present the first efficient *public coin* AM protocol, for any label-invariant property. Our protocol achieves sample complexity and communication complexity of magnitude $\widetilde{O}(N^{2/3})$, while the proof can be generated in quasi-linear $\widetilde{O}(N)$ time.

On top of that, we also give a public-coin protocol for efficiently verifying the distance a between a samplable distribution D, and some explicitly given distribution Q.

*This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819702), from the Israel Science Foundation (grant number 5219/17), and from the Simons Foundation Collaboration on the Theory of Algorithmic Fairness.

ISSN 1433-8092

1 Introduction

Given sample access to a distribution, what can we learn about the distribution, and what is the complexity of learning? These questions are central to computer science and statistics and have guided a rich body of work with applications ranging many fields. An emerging line of work asks the following question:

What is the complexity of *verifying* claims about a samplable distribution?

That is, suppose there exists a powerful yet untrusted prover that claims to have drawn many samples from a distribution D, and concluded that it satisfies some condition, e.g. its support is of size at most K, its Shannon entropy is h, etc. Can a verifier interacting with the prover be convinced that the claim is (approximately) correct, while taking fewer samples and running in less time than required to compute these measures directly from samples?

This question was raised by Chiesa and Gur [CG18], and recently Herman and Rothblum [HR23] showed that a rich family of distribution properties, namely *label-invariant* distribution properties - those distribution measures that remain unchanged after permuting the domain (such as the distribution's support size and Shannon entropy) - have *(doubly)* efficient proof systems, that for natural problems, allow verification that is significantly faster than computation from samples only. These protocols are *private-coin* protocols, in which the verifier can draw samples from D, toss random coins, and choose whether to send them to the prover, or keep them hidden from it. Indeed, the protocols in [HR23] rely heavily on the fact that the verifier *hides* its random coin-tosses in order to perform the verification. In this work we explore *public-coin* protocols for verifying distribution properties, in which the verifier reveals to the prover every coin it tosses immediately upon drawing it. We construct efficient *public-coin* proof systems for label-invariant distribution properties, and more.

More concretely, we follow the definition of *public-coin* proof systems for distribution properties from Chiesa and Gur [CG18], in which the verifier can only send random coin tosses to the prover, and the samples they draw from D are independent from the transcript of the protocol, and are drawn only after the communication phase.

Our work studies the power of public-coin proof systems in the context of verifying properties of an unknown samplable distribution. We find this to be a foundational question: indeed, the power of public-coin proof systems has been a central question since they were first introduced [GMR85, BM88]. In the classical setting (verifying the membership of a fixed and known input in a language), Goldwasser and Sipser [GS86] showed how to convert general protocols into public-coin ones (albeit their transformation does not preserve the honest prover's running time [Vad00, AR21]). In our context, where the verifier only has sampling access to the unknown distribution, no such general transformation is known. Chiesa and Gur showed upper and lower bounds for public-coin interactive proofs for distribution properties. Beyond the foundational importance of public-coin protocols, they are also important for removing interaction using the Fiat-Shamir paradigm [FS86] and for transforming general protocol into zero-knowledge ones [GMW91, BGG$^+$88]

1.1 This Work: Public-coin Protocols for Label-Invariant Distribution Properties

Our main result is a new *public-coin* protocol for label-invariant distribution properties. We proceed to present this result, and put it into context with the private-coin setting of [HR23], and the other public-coin distribution verification protocols of [CG18].

A distribution property $\mathcal{P} = (\mathcal{P}_N)_{N\in\mathbb{N}}$ is an ensemble such that $\mathcal{P}_N$ is a set of distributions over domain $[N]$. We consider the distance of a distribution D over domain $[N]$ from the property by the *total variation* of D from the closest distribution to it in $\mathcal{P}_N$. A distribution property is said to be *label-invariant* if permuting the domain doesn't change $\mathcal{P}$. This family of distribution properties contains many natural properties, such as the property of being close to uniform over some subset of the domain, or having Shannon entropy roughly k.

Theorem 1.1 (Main result: public-coin IPs for label-invariant properties, informal)**.** *For every label-invariant distribution property $\mathcal{P}$ with a doubly-efficient approximate decision procedure,*[1] *there exists a 2-message* public-coin *interactive protocol as follows. The prover and the verifier both get as input an integer N and proximity parameters $\varepsilon_c, \varepsilon_f \in [0,1]$ where $\varepsilon_c < \varepsilon_f$, as well as sampling access to an unknown distribution D over support $[N]$, and the following properties hold:*

- *Completeness: if D is ε_c-close to the property (its total variation distance from the closest distribution in the property is at most ε_c), and the prover follows the protocol, then w.h.p. the verifier accepts.*
- *Soundness: if D is ε_f-far from the property (its total variation distance from every distribution in the property is at least ε_f), then w.h.p. no matter how the prover cheats, the verifier rejects.*
- *Doubly-efficient prover: Taking $\rho = \varepsilon_f - \varepsilon_c$, the honest prover's runtime and sample complexity are $\widetilde{O}(N) \cdot \mathsf{poly}(1/\rho)$.*
- *Efficient verification: the communication complexity and the verifier's sample complexity and runtime are all $\widetilde{O}\left(N^{2/3}\right) \cdot \mathsf{poly}(1/\rho)$.*

Public-coin verification vs. testing of label-invariant distribution properties. Observe that the protocol above allows us to efficiently approximate the distance of D from $\mathcal{P}$, by running a binary search with different values for $\varepsilon_c, \varepsilon_f$. Raskhodnikova et al. [RRSS09], and Valiant and Valiant [VV11] showed that approximating the distance between D and natural label-invariant distribution properties, given only black-box sample access to the distribution, requires $\Theta(N/\log N)$ samples. This includes approximating the distance from being uniform over the entire domain, from having entropy k, and more. Thus, our result demonstrates that public-coin verification can be more efficient than stand-alone computation with no access to a prover for these natural distribution problems.

Comparison to the secret-coin setting of [HR23]. Herman and Rothblum provided a *secret-coin* interactive proof for verifying membership in any label-invariant distribution property (that admits an efficient approximate decision procedure) with verifier sample complexity, runtime, communication complexity of magnitude $\widetilde{O}\left(\sqrt{N}\right)$, and only two messages. The first message in their protocol contains a tuple of elements in $[N]$, where each element was sampled with probability $\frac{1}{2}$ from the distribution D, and with probability $\frac{1}{2}$ was drawn uniformly from $[N]$. Crucially for

[1] See Definition 4.16. In a nutshell, these are label-invariant properties that can be efficiently decided from the *τ-approximate bucket-histogram of the distribution*, i.e. by only knowing how many elements have probability roughly $\frac{(1+\tau)^j}{N}$ for all j, see Definition 2.2. [HR22] showed that this assumption is quite mild, and many natural distribution properties admit such a procedure, the reader is referred to [HR22] for a deeper exploration of this notion.

their argument, the verifier doesn't share with the prover which samples were drawn according to which distribution, and later capitalizes on that fact to reject dishonest prover behavior.

In our public-coin protocol not only is the verifier required to share the random coin tosses, it also cannot send samples from D as part of the communication. Thus, theorem 2.1 achieves a similar result qualitatively to theirs, but using only public coins, at the cost of more samples and communication.

Comparison with Chiesa and Gur [CG18]. Chiesa and Gur provided *public-coin* protocols for any property with communication $c = \widetilde{O}(N)$, and verifier sample complexity $s = O\left(\sqrt{N}\right)$, by having the prover send an explicit description of the distribution, and the verifier use an *identity tester* from the distribution testing literature to check that the description matches the samplable distribution. Then, the verifier accepts if D is both close to the explicit distribution provided, and if this description is of a distribution inside the property. Moreover they also proved that for a distribution property that requires $\Omega(t)$ samples to test, any *public-coin* proof system for this property must satisfy $s \cdot c = \Omega(t)$. As mentioned above, verifying the distance from uniformity or approximating the entropy of a distribution requires $\widetilde{\Omega}(N)$ samples, and so, every AM protocol that verifies this property must also satisfy $s \cdot c = \widetilde{\Omega}(N)$. Our protocol for this problem achieves $c \cdot s = \widetilde{O}(N^{4/3})$, and the question of whether there exists a more efficient *public-coin* proof system for this problem remains open.

Obtaining *approximate tags* of elements in $[N]$. The method through which our protocol allows the verifier to verify *any* label invariant distribution property is by having the verifier uniformly draw elements from $[N]$, and verifiably obtain an approximation of the probability of each element according to D, that is correct on average (we call this a *uniformly drawn approximate tagged sample*). Formally, for some accuracy parameter $\sigma \in (0, 1)$, and a tuple $(z_i) \in [N]^s$, we define:

Definition 1.2 (σ-approximate tags for (z_i) with respect to D)**.** *σ-approximate tags for (z_i) with respect to D is a tuple $(\pi_i)_{i\in[s]} \in [0, 1]^s$ that satisfies the following inequality:*

$$\frac{1}{s}\sum_{i\in[s]}\left(1 - \min\left\{\frac{D(z_i)}{\pi_i}, \frac{\pi_i}{D(z_i)}\right\}\right) \leq \sigma \tag{1}$$

In other words, on *average*, $\pi_i \in [1 \pm \sigma]\, D(z_i)$. A uniformly drawn approximate tagged sample allows to approximate the probability histogram of a distribution, as explained in the following sections. Note that in [HR22] and [HR23] the authors obtain an approximate tagged sample drawn according to D, rather than from a uniformly drawn sample, and use it to approximate the probability histogram of D. Thus, upon obtaining the probability histogram, our approaches converge, and we follow these works to bridge the gap between obtaining a probability histogram of a distribution and the estimation of distance from a label-invariant property. Note that the main difficulty is obtaining the tagged sample, a task that without communication would've required $\widetilde{\Omega}(N)$ samples, and so, this paper will focus on this point.

Moreover, [HR22, HR23] not only contain secret coins, but also rely on the fact that the verifier can send samples from D to the prover. In this work, we allow the verifier to only send random coins, not even samples from D. This choice is justified in Chiesa and Gur [CG18], and allows

our protocol to utilize properties of public-coin protocols over other objects with different access models.

We also show that a uniformly drawn approximate tagged-sample can also be used to verify distribution properties that are *not* label-invariant. Specifically, we also show that for the well-studied problem of approximating the distance of D from an explicit distribution Q, an approximate tagged uniform sample is sufficient:

Theorem 1.3 (Tolerant Verification of Identity)**.** *Given an explicit description of distribution Q over $[N]$, parameters $0 < \varepsilon_c < \varepsilon_f <$, and sample access to distribution D over domain $[N]$, there exists a 2-message* public-coin *protocol, with verifier sample complexity and communication complexity $\widetilde{O}(N^{2/3}) \cdot \mathsf{poly}(\frac{1}{\varepsilon_f - \varepsilon_c})$ such that:*

- *If $\Delta_{SD}(D, Q) \leq \varepsilon_c$, the verifier accepts with high probability.*
- *If $\Delta_{SD}(D, Q) \geq \varepsilon_f$, the verifier rejects with high probability.*

1.2 Further Related Works

Interactive proof systems were introduced in the seminal work of Goldwasser, Micali and Rackoff [GMR85] in the context of proving computational statements about an input that is fully known to the prover and the verifier. In our work, the distribution can be thought of as the input, but it is not fully known to the verifier, and is accessed implicitly through samples. We aim for verification without examining the distribution in its entirety, using minimal resources (samples, communication, runtime, etc.).

Our work builds on a line of work that studied the power of sublinear time verifiers, who cannot read the entire input [EKR04, RVW13, GR18], on verifying properties of distributions using a small number of samples [CG18, HR22, HR23], and the rich literature of distribution testing, of which most notably, we extensively use the ideas of Batu and Canonne in [BC17], as explained in the technical overview. We also note that Herman and Rothblum [HR24] recently showed that a very rich family of distribution properties, those that can be decided by a *small* circuit from an explicit description of the distribution, can be doubly-efficiently verified with a*secret-coin* protocol.

2 Technical Overview

As discussed in the introduction above, the protocol behind Theorem 1.1 is based on obtaining verified $\Theta(\rho)$-*approximate tags with respect to* D for a sample uniformly drawn from $[N]$. In this section, we describe the public-coin protocol for obtaining this object. We then detail how this tagged sample can be leveraged to verify membership in label-invariant distribution properties.

Theorem 2.1. *[Informal] There exists a 2-message* public-coin *interactive protocol between a verifier and a (potentially malicious) prover, where the verifier receives as input parameters $\sigma \in (0, 0.1)$ and $N \in \mathbb{N}$, as well as sample access to a distribution D over domain $[N]$. The communication complexity, verifier sample complexity, and verifier runtime are all $s = \widetilde{O}\left(N^{2/3}\right) \mathsf{poly}(\sigma^{-1})$, the honest prover with the same input as the verifier has sample complexity and runtime $\widetilde{O}(N)\mathsf{poly}(\sigma^{-1})$. At the end of the interaction, the verifier rejects or outputs $(S_i) \in [N]^s$ that is drawn uniformly from $[N]$, and $(\pi_i) \in [0, 1]^s$ such that:*

- *If the prover is honest, for all $i \in [s]$, $\pi_i = D(S_i)$, and with probability at least 0.75, the verifier doesn't reject.*
- *Whatever strategy a dishonest prover follows, with probability at most 0.25 over the verifier's coin tosses and samples, the verifier accepts and outputs (π_i) such that doesn't satisfy Inequality (1).*

We outline the protocol behind Theorem 2.1. We highlight that some details are swept under the rug for sake of simplicity. In particular, we assume that $D(x) \leq \frac{1}{s}$ for all $x \in [N]$. After we present the protocol under this assumption, we discuss how to remove this assumption.

The communication phase. The verifier draws an i.i.d. sample $S = (S_i)$ of size $s = \widetilde{O}\left(N^{2/3}\right) \cdot \mathsf{poly}\left(\sigma^{-1}\right)$ uniformly from $[N]$, and sends the sample the prover. For each sample S_i received, the prover replies with π_i such that $\pi_i = D(S_i)$. Note that with high probability, due to the choice of s, there doesn't exist an element in $x \in [N]$ that was sampled more than 3 times,[2] and in general, the fraction of elements that were sampled twice or three times is very small with respect to s. Therefore, for sake of simplicity, assume that S contains only unique elements.

Moreover, since we assumed $D(x) \leq \frac{1}{s}$ for all $x \in [N]$, by choice of s, the sample S contains with overwhelming probability *many* samples uniformly distributed inside $\mathrm{Supp}(D)$.

Verifing the prover's message. The verifier divides the samples in S into *buckets* according to their alleged probability, where inside each bucket all the samples are claimed to have roughly the same mass. Concretely, for $\tau = O(\sigma^3)$, and for every j, denote by $B_j^S \subseteq [s]$ the collection of indices in S that the prover claimed have probability in the range $\left[\frac{(1+\tau)^j}{N}, \frac{(1+\tau)^{j+1}}{N}\right]$. The verifier then tests for every such j that the *average probability* of the elements in B_j^S is indeed roughly $\frac{(1+\tau)^j}{N}$, and that $D\big|_{B_j^S}$ is *close to uniform*:

- **Checking that the average mass is correct.** The verifier draws a fresh sample T, and checks that the empirical mass of B_j^S in T is roughly $s \cdot \left|B_j^S\right| \cdot \frac{(1+\tau)^j}{N}$, and rejects otherwise. Observe that for any distribution D, the true mass of B_j^S is $\sum_{k \in B_j^S} D(S_k)$. And so, by choice of s, since the empirical mass of B_j^S in T is strongly concentrated around its mean, if the test passes, then with high probability:

$$s \cdot \sum_{k \in B_j^S} D(x) \overset{\tau}{\approx} s \cdot \left|B_j^S\right| \cdot \frac{(1+\tau)^j}{N}$$

  Where for $\alpha \in (0,1)$ we use the notation $a \overset{\alpha}{\approx} b$ to indicate that $a \in (1 \pm \alpha)b$. We conclude that with high probability:

$$\mathbb{E}_{k \overset{uni}{\sim} B_j^S}\left[D(S_k)\right] \overset{O(\tau)}{\approx} \frac{(1+\tau)^j}{N} \tag{2}$$

[2] The probability that 4 samples collide is $\sum_{x \in [N]} D(x)^4 = \frac{1}{N^3}$ while there are only $\binom{s}{4} = O(N^{8/3})$ possible 4-tuples in the sample S.

- **Verifying that $D\big|_{B_j^S}$ is close to uniform.** The verifier draws another fresh D-sample T' of size s, and counts how many *3-way collisions* occur between elements in B_j^S and the two samples T, T', i.e. the number of 3-tuples $(k, r, r') \in [s]^3$ satisfy $k \in B_j^S$, $S_k = T_r = T'_{r'}$. If this quantity is far from $s^2 \cdot \left|B_j^S\right| \cdot \left(\frac{(1+\tau)^j}{N}\right)^2$, the verifier rejects. Similar to before, for any fixed pair of entries in T, T', $(r, r') \in [s]^2$, the true expected number of $k \in B_j^S$ for which $S_k = T_r = T_{r'}$ is $\sum_{k \in B_j^S} (D(S_k))^2$. The total expected number of such 3-tuples is $s^2 \cdot \sum_{k \in B_j^S} (D(S_k))^2$. This quantity is also strongly concentrated around its mean by choice of $s = \Theta(N^{2/3})\mathsf{poly}(\sigma^{-1})$. We conclude that if this test passed, then with high probability:

$$s^2 \cdot \sum_{k \in B_j^S} (D(S_k))^2 \overset{O(\tau)}{\approx} s^2 \left|B_j^S\right| \cdot \left(\frac{(1+\tau)^j}{N}\right)^2$$

And equivalently:

$$\mathbb{E}_{k \overset{uni}{\sim} B_j^S}\left[(D(S_k))^2\right] \overset{O(\tau)}{\approx} \left(\frac{(1+\tau)^j}{N}\right)^2 \tag{3}$$

We are thus left to argue that Equations (2) and (3) imply that $D\big|_{B_j^S}$ is close to uniform. Following Batu and Canonne [BC17], observe that:

$$\mathrm{Var}_{k \overset{uni}{\sim} B_j^S}\left[D(S_k)\right] = \mathbb{E}_{k \overset{uni}{\sim} B_j^S}\left[(D(S_k))^2\right] - \left(\mathbb{E}_{k \overset{uni}{\sim} B_j^S}\left[D(S_k)\right]\right)^2$$

And so, assuming Equations (2) and (3) hold, we get that $\mathrm{Var}_{k \overset{uni}{\sim} B_j^S}\left[D(S_k)\right] = O(\tau)\left(\mathbb{E}\left[D(x)\right]\right)^2$. Using Chebychev's Inequality:

$$\Pr_{k \overset{uni}{\sim} B_j^S}\left(\left|D(S_k) - \mathbb{E}_{k \overset{uni}{\sim} B_j^S}\left[D(S_k)\right]\right| \geq O\left(\sqrt{\frac{\tau}{\sigma}}\right) \cdot \mathbb{E}_{k \overset{uni}{\sim} B_j^S}\left[D(S_k)\right]\right) \leq O(\sigma) \tag{4}$$

From which we conclude all but σ-fraction of entries $i \in B_j^S$ satisfy:

$$\pi_i \overset{O(\tau)}{\approx} \frac{(1+\tau)^j}{N} \overset{O(\tau)}{\approx} \mathbb{E}_{k \overset{uni}{\sim} B_j^S}\left[D(S_k)\right] \overset{O(\sqrt{\tau/\sigma})}{\approx} D(S_i)$$

Where the first inequality stems from the definition of B_j^s, the second from Equation (2), and the last from Inequality (4). Plugging in $\tau = O(\sigma^3)$, we get: $\pi_i \overset{O(\sigma)}{\approx} D(S_i)$.

We thus showed that if both verifier tests pass, then with high probability over the randomness of the verifier, it holds that for every j, the tags over B_j^S are σ-approximately correct, from which Inequality (1) is inferred.

Assuming D contains no *heavy* elements. Observe that the probability of all elements with probability larger than $1/s$ can be well-approximated through their empirical mass in a sample of size $\widetilde{\Theta}(s)$ from D. Therefore, we can think of a verifier that estimates without need of a prover the mass of all such elements. This process is described in detail in [HR22], and we describe it shortly here. The reader is referred to their work for further detail. After receiving the prover's tags, the verifier performs the following step: the verifier draws a fresh D-sample, denoted $\mathcal{H}$, of size $\widetilde{O}(s)\mathsf{poly}(\sigma^{-1})$ from D. With high probability, by a *coupon-collector* argument, this set contains all elements with probability at least $\frac{1}{s}$ (if any exist).

The verifier tests the mass of $\mathcal{H}$ be drawing a fresh sample and examining the empirical mass of $\mathcal{H}$ in that new sample. If it is significant, i.e. $\Omega(\sigma)$, the verifier "learns" $D\big|_{\mathcal{H}}$ up to σ distance by subsampling from this distribution and running a *folklore distribution learner* (see Theorem 2.1). This requires $\widetilde{O}(s)\mathsf{poly}(\sigma^{-1})$ samples from D, and thus doesn't incur significant overhead to the sample complexity of the protocol. Thus, the verifier obtains an explicit description of the distribution $P_{\mathcal{H}}$, which is $O(\sigma)$-close to $D\big|_{\mathcal{H}}$. Since $\mathcal{H}$ is a set of size at most s, and the sample S was drawn drawn i.i.d. from $[N]$, with overwhelming probability it holds that $|S \cap \mathcal{H}| = O(N^{1/3}) = o(s)$, and in order to verify the prover's answer's in the protocol described above, the verifier can just "erase" every element in S that appeared in $\mathcal{H}$, and run the protocol presented above over just elements guaranteed with high probability to be of probability at most $1/s$, without affecting the correctness of the protocol. Thus, the verifier obtains full tags for $\mathcal{H}$, and tags for $S \setminus \mathcal{H}$. Later, the verifier can "fill-in" the missing parts in S to obtain a full tagged sample. If D is entirely supported over heavy elements, then the protocol can be avoided all together by also checking the mass of $\mathcal{H}$ is larger than $1 - O(\sigma)$, and ignoring the prover's message.

Verifying label-invariant distribution properties. In order to verify *label-invariant* distribution properties, it suffices to know the *probability histogram* of the distribution, i.e., how many elements have probability p for every $p \in [0,1]$. Herman and Rothblum [HR22] observed that for many natural properties an approximation of this histogram is sufficient, and define the *τ-bucket histogram* as follows:

Definition 2.2 (τ-bucket histogram of D)**.** *For any $j \in \{\ldots,-1,0,1,\ldots,\frac{\log N}{\tau}\}$, the j'th bucket of D over domain $[N]$ is:*

$$B_j^D = \left\{ x :\in Supp(D) : D(x) \in \left[\frac{(1+\tau)^j}{N}, \frac{(1+\tau)^{j+1}}{N} \right) \right\}$$

The τ-bucket histogram of D is the tuple $\Big((j, D(B_j^D))\Big)_{j:B_j^D \neq \phi}$.

In [HR22] the authors focus their attention on those label-invariant distribution properties for which the information $\Big((j, D(B_j^D))\Big)_{j:B_j^D \neq \phi}$ is sufficient in order to efficiently approximate the *distance* (in total-variation) of D from the property. They say that such properties admit an *efficient approximate decision procedure*, and show that many natural label-invariant problems are of this type, including the property of having Shannon entropy roughly k, or being close to uniform over some set of size $M \leq N$.

In our protocol the verifier obtains a uniformly drawn tagged sample[3]. We argue that this tagged sample allows the verifier to compute an approximation of the bucket histogram of D: if our

[3]Here we differ from [HR22] that obtain a D-sampled tagged sample, i.e. (z_i) in their case was drawn from D.

protocol didn't end in rejection, then with high probability, the tags are roughly correct. In other words, for every j, $\frac{|B_j^S|}{s}$ is the empirical mass of B_j^D in the uniform sample S. Since we expect there to be about $\frac{|B_j^D|}{N}$-fraction of samples in S that landed in B_j^D, we conclude that:

$$\frac{\left|B_j^S\right|}{s} \approx \frac{\left|B_j^D\right|}{N}$$

And since $D(B_j^D) \approx \left|B_j^D\right| \cdot \frac{(1+\tau)^j}{N}$, if we set $p_j = \left(\frac{|B_j^S|}{s} \cdot N\right) \cdot \frac{(1+\tau)^j}{N}$, then $D(B_j^D) \approx p_j$, and we get with high probability, a τ-histogram which is $O(\sigma)$ close to the true histogram of D in the following sense: there exists a distribution D' with histogram exactly $((j, p_j))$ that is $O(\sigma)$-close to D in total variation distance. Thus, using the decision procedure, the verifier decides whether $((j, p_j))$ is consistent with some distribution close to $\mathcal{P}$, and thus, conclude whether D is far from the property, or close to it.

3 Preliminaries

For an integer $n \in \mathbb{N}$, we use $[n]$ to denote the set $\{1, \ldots, n\}$.

Definition 3.1. *The total variation distance (alt. statistical distance) between distributions P and Q over a finite domain X is defined as:*

$$\Delta_{SD}(P, Q) = \frac{1}{2} \sum_{x \in X} |P(x) - Q(x)|$$

Theorem 3.2 (Folklore distribution learner [Gol17])**.** *There exists an algorithm that given sample access to a distribution P over the domain $[N]$, and an accuracy parameter $\alpha \in (0, 1)$, it runs in time $\widetilde{O}(N/\alpha^2)$, takes $O(N/\alpha^2)$ samples, and with probability at least 0.99 outputs a full description of a distribution P_{approx} such that $\Delta_{SD}(P, P_{approx}) \leq \alpha$.*

Definition 3.3 (Distribution property)**.** *We say the $\mathcal{P} = (\mathcal{P}_N)_{N \in \mathbb{N}}$ is a distribution property if $\mathcal{P}_N \subseteq \Delta_N$, where Δ_N is the set of all distributions over domain $[N]$.*

Definition 3.4 (Distribution tester for property $\mathcal{P}$)**.** *Let $\mathcal{P}$ be a distribution property. A tester T of property $\mathcal{P}$ is a probabilistic oracle machine, that on input parameters N and ε, and oracle access to a sampling device for a distribution D over a domain of size $[N]$, outputs a binary verdict that satisfies the following two conditions:*

1. *If $D \in \mathcal{P}_N$, then $\Pr(T^D(N, \varepsilon) = 1) \geq 2/3$.*
2. *If $\Delta_{SD}(D, \mathcal{P}_N) > \varepsilon$, then $\Pr(T^D(N, \varepsilon) = 0) \geq 2/3$.*

In the context of this work, the relevant distance measure is *statistical distance* as defined above. An extension of this definition, introduced by Parnas, Ron, and Rubinfeld [PRR06] is the following:

Definition 3.5 ($(\varepsilon_c, \varepsilon_f)$-tolerant distribution property tester)**.** *For parameters $\varepsilon_c, \varepsilon_f \in [0, 1]$ such that $\varepsilon_c < \varepsilon_f$, a $(\varepsilon_c, \varepsilon_f)$-tolerant tester T of property Π is a probabilistic oracle machine, that on inputs $N, \varepsilon_c, \varepsilon_f$ and given oracle access to a sampling device for distribution D over a domain of size N, outputs a binary verdict that satisfies the following two conditions:*

1. *If* $\delta(D, \Pi_N) \leq \varepsilon_c$, *then* $\Pr(T^D(N, \varepsilon_c, \varepsilon_f) = 1) \geq 2/3$.
2. *If* $\delta(D, \Pi_N) \geq \varepsilon_f$, *then* $\Pr(T^D(N, \varepsilon_c, \varepsilon_f) = 0) \geq 2/3$.

Note that a tolerant distribution test is for some property Π is at least as hard as a standard non-tolerant tester for the same property.

Definition 3.6 (Proof system for tolerant distribution testing problems)**.** *A proof system for a tolerant distribution testing problem $\mathcal{P}$ with parameters ε_c and ε_f is a two-party game, between a verifier executing a probabilistic polynomial time strategy V, and a prover that executes a strategy P. Given that both V and P have black-box sample access to distribution D over the domain $[N]$, and are given N, the interaction should satisfy the following conditions:*

- ***Completeness:*** *For every D over domain of size at most N, such that $\Delta_{SD}(D, \mathcal{P}_N) \leq \varepsilon_c$, the verifier V, after interacting with the prover P, accepts with probability at least $2/3$.*
- ***Soundness:*** *For every D over domain of size at most N such that $\Delta_{SD}(D, \mathcal{P}_N) \geq \varepsilon_f$, and every cheating strategy P^*, the verifier V, after interacting with the prover P^*, rejects with probability at least $2/3$.*

The complexity measures associated with the protocol are: the sample complexity of the verifier as as the honest prover (strategy P), the communication complexity, the runtime of both agents, and the round complexity (how many messages were exchanged).

Definition 3.7 (Label invariant distribution property)**.** *A distribution property $\mathcal{P}$ is called label invariant if for all $N \in \mathbb{N}$, it holds that any permutation σ over N elements satisfies that $D \in \mathcal{P}_N$ if and only if $\sigma(D) \in \mathcal{P}_N$.*

4 Public Coin Protocol for Verified Tagged Sample

Using the same approach as Herman and Rothblum [HR22], we provide an algorithm to obtain a *tagged sample* assuming that the samplable distribution D satisfies that for every $x \in [N]$, $D(x) \leq \frac{1}{s}$, where $s = O\left(\frac{\log N}{\varepsilon^5} \cdot N^{2/3}\right)$. In Section 2 we discuss why we can assume this without loss of generality.

Theorem 4.1. *There exists 2-message* `AM` *interactive protocol between an honest verifier and a (potentially malicious) prover, where the verifier receives as input parameters $\sigma \in (0, 0.1)$ and $100 < N \in \mathbb{N}$, as well as sample access to a distribution D over domain $[N]$. Set $\tau = \frac{\sigma^3}{8000}$. Assume $D(x) \leq \frac{1}{s}$ for $s = O\left(\frac{\log N}{\varepsilon^5} \cdot N^{2/3}\right)$. The communication complexity, verifier sample complexity, and verifier runtime are all s. Given sample access to the distribution D, the honest prover requires with high probability $\widetilde{O}(N)\,\mathsf{poly}(\sigma^{-1})$ samples and runtime.*

At the end of the interaction, the verifier rejects or outputs $((z_i, \pi_i))_{i \in [s]}$ where $(z_i)_{i \in [s]}$ is a sample of size s drawn uniformly i.i.d. from $[N]$ and:

- **Completeness.** *If the prover is honest, then with probability at least 0.75, the verifier doesn't reject, and $((z_i, \pi_i))_{i \in [s']}$ satsifies $\frac{1}{s}\sum_{i \in [s]: \pi_i \geq \frac{\sigma}{1000N}} \left(1 - \min\left\{\frac{\pi_i}{D(z_i)}, \frac{D(z_i)}{\pi_i}\right\}\right) = O(\tau)$, while $\frac{1}{s}\sum_{i \in [s]: \pi_i \leq \frac{\sigma}{1000N}} D(z_i) \leq \frac{\sigma}{50N}$.*

- **Soundness.** *Whatever strategy a dishonest prover follows, with probability at most* 0.25 *over the verifier's coin tosses and samples, they accept and* $((z_i, \pi_i))_{i\in[s']}$ *satisfies:*

$$\frac{1}{s}\sum_{i\in[s]:\pi_i\geq\frac{\sigma}{1000N}}\left(1-\min\left\{\frac{\pi_i}{D(z_i)},\frac{D(z_i)}{\pi_i}\right\}\right)\geq\sigma \tag{5}$$

or

$$\frac{1}{s}\sum_{i\in[s]:\pi_i\leq\frac{\sigma}{1000N}}D(z_i)\geq\frac{\sigma}{10N} \tag{6}$$

Note that we use the convention that $\min\left\{\frac{\pi_i}{D(z_i)},\frac{D(z_i)}{\pi_i}\right\}=1$ *if* $\pi_i=0$ *and* $D(z_i)\neq 0$*, or* $\pi_i\neq 0$ *and* $D(z_i)=0$.

We show that Protocol 4.1.1 satisfies the conditions of Theorem 4.1.

Protocol 4.1.1: Public-Sample Tagged Sample Retrieval Protocol

Input: parameters $N\in\mathbb{N}$, $\sigma\in(0,1)$, as well as sample access to distribution D over domain $[N]$ such that for all $x\in[N]$, $D(x)\leq\frac{1}{s}$ for $s=O\left(\frac{\log N}{\varepsilon^5}N^{2/3}\right)$.

1. V: draw s uniformly from $[N]$. Denote the sample $(S_i)_{i\in[s]}$. Reject if there exists $x\in[N]$ such that x appears more than $\log N$ times in S. Otherwise, send (S_i) to P.
2. P: set $\tau=\frac{\sigma^3}{80000}$. For every $i\in[s]$, if $D(S_i)\geq\frac{\sigma}{100N}$, send π_i such that $\pi_i=D(S_i)$, otherwise, send $\pi_i=0$.
3. V: for every j set $S^j=\left\{i\in[s]:\pi_i\in\left[\frac{e^{j\tau}}{N},\frac{e^{(j+1)\tau}}{N}\right)\right\}$. Draw two fresh samples of size s from D, $T=(T_i)_{i\in[s]}$ and $T'=(T'_i)_{i\in[s]}$. For every j such that $\left|S^j\right|\geq e^{-j\tau}\cdot s\cdot\frac{\varepsilon\cdot\tau}{100\log N}$ and $\frac{e^{j\tau}}{N}\geq\frac{\sigma}{1000N}$, set:

$$\widetilde{C}_j^{pair}=\left|\left\{(k,r)\in[s]^2:k\in S^j,\ S_k=T_r\right\}\right|$$

$$\widetilde{C}_j^{triple}=\left|\left\{(k,r,r')\in[s]^3:k\in S^j,\ S_k=T_r=T'_r\right\}\right|$$

   Reject unless for all such j:

$$\left|\widetilde{C}_j^{pair}-s\cdot\left|S^j\right|\cdot\frac{e^{j\tau}}{N}\right|\leq 4\tau\cdot s\cdot\left|S^j\right|\cdot\frac{e^{j\tau}}{N} \tag{7}$$

   And

$$\left|\widetilde{C}_j^{triple}-s^2\cdot\left|S^j\right|\cdot\left(\frac{e^{j\tau}}{N}\right)^2\right|\leq 4\tau\cdot s^2\cdot\left|S^j\right|\cdot\left(\frac{e^{j\tau}}{N}\right)^2 \tag{8}$$

4. V: denote $S^{-\infty}=\{i\in[s]:\pi_i=0\}$. Reject unless $\widetilde{C}_{-\infty}^{pair}\leq s\cdot|S^{-\infty}|\cdot\frac{\sigma}{50N}$.
5. V: Output $((S_i,\pi_i))_{i\in[s]}$

4.1 Protocol 4.1.1 is Complete

We first show that Step 1 of Protocol 4.1.1 does not result in rejection.

Claim 4.2. *With probability at least* 0.99 *over the choice of* S*, there doesn't exist an element* $x \in [N]$ *that was sampled more than* 3 *times in* S*, and the verifier doesn't reject after Step 1 of Protocol 4.1.1.*

Proof. Fix $x \in [N]$ and $i_1, i_2, i_3, i_4 \in [s]$ such that for all $k, k' \in [\log N]$, $i_k \neq i_{k'}$. Note that:

$$\Pr_S \left(S_{i_1} = S_{i_2} = S_{i_3} = S_{i_4}\right) = \left(\frac{1}{N}\right)^4$$

There are $\binom{s}{4}$ possible choices for $i_1, i_2, i_3, i_4 \in [s]$. Therefore, the probability that there exists some set of 4 indices whose respective samples equal x is at most:

$$\binom{s}{4} \cdot \frac{1}{N^4} \leq \left(\frac{s}{N}\right)^4 \leq \frac{1}{N^{4/3}}$$

Taking the union bound over all possible $x \in [N]$ yields the desired result. □

Next, we argue that if the prover is honest, with high probability, the verifier collision tests don't result in rejection.

Claim 4.3. *Assuming the verifier didn't reject after Step 1 and that the prover is honest, then with probability at least* 0.8 *over the choice of* T, T' *the verifier doesn't reject.*

Proof. For every j such that $\frac{e^{j\tau}}{N} \geq \frac{\sigma}{1000N}$ and $\left|S^j\right| \geq e^{-j\tau} \cdot s \cdot \frac{\varepsilon \cdot \tau}{100 \log N}$ by Propositions A.1 and A.2 and choice of s, it also holds that:

$$\mathbb{E}\left[\widetilde{C}_j^{pair}\right] = s\left(\sum_{i \in S^j} D(S_i)\right) \geq s \cdot \left|S^j\right| \cdot \frac{e^{j\tau}}{N} \geq \frac{300 \log^2 N}{\tau^3}$$

$$\mathbb{E}\left[\widetilde{C}_j^{triple}\right] \geq s^2 \sum_{i \in S^j} (D(S_i))^2 = s^2 \cdot \left|S^j\right| \cdot \left(\frac{e^{j\tau}}{N}\right)^3 \geq \frac{300 \log^2 N}{\tau^3}$$

And so, since there are at most $2 \log N/\tau$ buckets for which $\frac{e^{j\tau}}{N} \geq \frac{\sigma}{1000N}$, we conclude from Propositions A.1 and A.2 that with probability at least 0.8 over the choice of T, T' for all j as described in statement it holds that:

$$\left|\widetilde{C}_j^{triple} - \mathbb{E}\left[\widetilde{C}_j^{triple}\right]\right| \leq \mathbb{E}\left[\widetilde{C}_j^{triple}\right] \cdot \sqrt{\frac{300 \log^2 N}{\tau \cdot \mathbb{E}\left[\widetilde{C}_j^{triple}\right]}} \leq \left(e^{2\tau} - 1\right) s^2 \left|S^j\right| \left(\frac{e^{j\tau}}{N}\right)^2 \cdot \tau \leq 4\tau s^2 \left|S^j\right| \left(\frac{e^{j\tau}}{N}\right)^2$$

And similarly:

$$\left|\widetilde{C}_j^{pair} - \mathbb{E}\left[\widetilde{C}_j^{pair}\right]\right| \leq \mathbb{E}\left[\widetilde{C}_j^{pair}\right] \cdot \sqrt{\frac{300 \log^2 N}{\tau \cdot \mathbb{E}\left[\widetilde{C}_j^{pair]}\right]}} \leq \left(e^{\tau} - 1\right) s \left|S^j\right| \frac{e^{j\tau}}{N} \cdot \tau \leq 4\tau s \left|S^j\right| \frac{e^{j\tau}}{N}$$

□

Claim 4.4. *If the prover is honest, with high probability over T, the final verifier test passes with high probability, and:*

$$\frac{1}{s}\sum_{i\in[s]:\pi_i<\frac{\sigma}{1000N}} D(S_i) \leq \frac{\sigma}{10N} \tag{9}$$

Proof. Since the prover is honest, $\mathbb{E}\left[\widetilde{C}_{-\infty}\right] = s\cdot\sum_{i\in S^{-\infty}} D(S_i) \leq s\cdot|S^{-\infty}|\cdot\frac{\sigma}{1000N}$, and so, by Markov's Inequality, with probability at least 0.95, $\widetilde{C}_{-\infty}\leq s\cdot|S^{-\infty}|\cdot\frac{\sigma}{50N}$, and the final test passes. Moreover, Inequality (9) holds. □

Remark 4.5 (Honest prover complexity)**.** *For sake of simplicity we assume the honest prover in Protocol 4.1.1 knows $D(S_i)$ exactly. However, this is not necessary. A prover that approximates this quantity for every sample up to sufficient accuracy using only $\widetilde{O}(N)\mathsf{poly}(\tau^{-1})$ samples suffices. See Remark 4.14 in [HR23] for a detailed discussion.*

4.2 Protocol 4.1.1 is Sound

Note that by Claim 4.2, regardless of the prover's response, the verifier rejects after Step 1 with probability at most 0.01, and so, throughout this section, we assume that Step 1 passed, and S doesn't contain elements appearing more than 4 times, even when not stated explicitly.

First, we address the *last* verifier test:

Claim 4.6. *For every index j such that $\frac{e^{j\tau}}{N}\geq\frac{\sigma}{1000N}$ and $\left|S^j\right|\geq e^{-j\tau}\cdot s\cdot\frac{\varepsilon\cdot\tau}{100\log N}$, with probability at least 0.98 over the choice of T and T', either the verifier rejects, or it holds that:*

$$\mathbb{E}\left[\widetilde{C}_j^{pair}\right]\geq\frac{300\log^2 N}{\tau^3} \tag{10}$$

and

$$\mathbb{E}\left[\widetilde{C}_j^{triple}\right]\geq\frac{300\log^2 N}{\tau^3} \tag{11}$$

Proof. Fix some j_0 such that $\left|S^{j_0}\right|\geq s\cdot\frac{\varepsilon\tau}{100\log N}$, $\frac{e^{j_0\tau}}{N}\geq\frac{\varepsilon}{100N}$, and also $\mathbb{E}\left[\widetilde{C}_{j_0}^{triple}\right]<\frac{300\log^2 N}{\tau^3}$. By Markov's Inequality, with probability at least 0.99:

$$\widetilde{C}_{j_0}^{triple}\leq 100\mathbb{E}\left[\widetilde{C}_{j_0}^{triple}\right]\leq\frac{30000\log^2 N}{\tau^3}$$

However, the verifier rejects unless:

$$\widetilde{C}_{j_0}^{triple}\geq(1-4\tau)\,s^2\left|S^{j_0}\right|\left(\frac{e^{j_0\tau}}{N}\right)^2\geq\frac{1}{2}s^3\cdot\frac{\tau\varepsilon}{100\log N}\cdot\left(\frac{\sigma}{1000N}\right)^2\geq s^3\cdot\frac{\tau\varepsilon^3}{2\cdot100^3N^3\log N}>\frac{30000\log^2 N}{\tau^3}$$

Where the last inequality is justified since $s\geq\frac{300\log N}{\tau^{4/3}\varepsilon}N^{2/3}$. We thus conclude that for every j such that $v_{j_0}\geq\frac{\varepsilon\tau}{100\log N}$, $\frac{e^{j_0\tau}}{N}\geq\frac{\varepsilon}{100N}$, either $\mathbb{E}\left[\widetilde{C}_j^{triple}\right]\geq\frac{300\log^2 N}{\tau^3}$ or the verifier reject with probability at least 0.99. An analogous argument can be made w.r.t. to $\widetilde{C}_j^{pair}$. Taking the union bound over both these events yields the required result. □

Claim 4.7. *With probability at least* 0.8 *over the choice of* T *and* T'*, for every* j *such that* $\left|S^j\right| \geq e^{-j\tau} \cdot s \cdot \frac{\varepsilon \cdot \tau}{100 \log N}$ *and* $\frac{e^{j\tau}}{N} \geq \frac{\sigma}{1000N}$*, and for which Inequalities* (10) *and* (11) *hold, it further holds that:*

$$\left| \widetilde{C}_j^{pair} - s \sum_{i \in S^j} D(S_i) \right| \leq 4\tau \cdot s \sum_{i \in S^j} D(S_i) \tag{12}$$

As well as:

$$\left| \widetilde{C}_j^{triple} - s^2 \sum_{i \in S^j} (D(S_i))^2 \right| \leq 4\tau \cdot s^2 \sum_{i \in S^j} (D(S_i))^2 \tag{13}$$

Proof. By Propositions A.1 and A.2 it holds that with probability 0.8 over the choice of T and T' for every j such that $\left|S^j\right| \geq s \cdot \frac{\varepsilon \cdot \tau}{100 \log N}$ and $\frac{e^{j\tau}}{N} \geq \frac{\varepsilon}{100N}$, the following holds:

$$\left| \widetilde{C}_j^{pair} - \mathbb{E}\left[\widetilde{C}_j^{pair}\right] \right| \leq \mathbb{E}\left[\widetilde{C}_j^{pair}\right] \cdot \sqrt{\frac{300 \log^2 N}{\tau \cdot \mathbb{E}\left[\widetilde{C}_j^{pair}\right]}}$$

$$\left| \widetilde{C}_j^{triple} - \mathbb{E}\left[\widetilde{C}_j^{triple}\right] \right| \leq \mathbb{E}\left[\widetilde{C}_j^{triple}\right] \cdot \sqrt{\frac{300 \log^2 N}{\tau \cdot \mathbb{E}\left[\widetilde{C}_j^{triple}\right]}}$$

Moreover, from the same propositions we know that:

$$\mathbb{E}\left[\widetilde{C}_j^{pair}\right] = s \sum_{i \in S^j} D(S_i)$$

$$\mathbb{E}\left[\widetilde{C}_j^{triple}\right] = s^2 \sum_{i \in S^j} (D(S_i))^2$$

We thus conclude that for all the j as specified above:

$$\left| \widetilde{C}_j^{pair} - s \sum_{i \in S^j} D(S_i) \right| \leq \mathbb{E}\left[\widetilde{C}_j^{pair}\right] \cdot \sqrt{\frac{300 \log^2 N}{\tau \cdot \mathbb{E}\left[\widetilde{C}_j^{pair}\right]}} \leq \tau s \sum_{i \in S^j} D(S_i)$$

Where the last inequality above stems from the assumption that Inequality (11) holds. Similarly:

$$\left| \widetilde{C}_j^{triple} - s^2 \sum_{i \in S^j} (D(S_i))^2 \right| \leq \tau s^2 \sum_{i \in S^j} (D(S_i))^2$$

□

Claim 4.8. *Assuming the verifier didn't reject, with probability at least* 0.8 *over the choice of* T *and* T'*, for every* j *such that* $\left|S^j\right| \geq e^{-j\tau} \cdot s \cdot \frac{\varepsilon \cdot \tau}{100 \log N}$ *and* $\frac{e^{j\tau}}{N} \geq \frac{\sigma}{1000N}$*, and for which Inequalities* (10) *and* (11) *hold. It further holds that:*

$$\frac{1}{|S^j|} \sum_{i \in S^j} D(S_i) \in \frac{e^{j\tau}}{N} \left[1 - 10\tau, 1 + 10\tau\right] \tag{14}$$

$$\frac{1}{|S^j|}\sum_{i\in S^j}(D(S_i))^2 \in \left(\frac{e^{j\tau}}{N}\right)^2 [1-10\tau, 1+10\tau] \tag{15}$$

Proof. By Claim 4.7, with probability at least 0.8 over the choice of T and T', for every j such that and $\left|S^j\right| \geq e^{-j\tau}\cdot s\cdot\frac{\varepsilon\cdot\tau}{100\log N}$ and $\frac{e^{j\tau}}{N}\geq\frac{\sigma}{1000N}$, and for which Inequalities (10) and (11) hold, Inequalities (12) and (13) hold.

Furthermore, if the verifier didn't reject, for all such j, Inequalities (7) and (8) holds as well for all such j. Putting it all together, we get that:

$$\left|s\cdot\left|S^j\right|\cdot\frac{e^{j\tau}}{N} - s\sum_{i\in S^j}D(S_i)\right| \leq \left|s\cdot\left|S^j\right|\cdot\frac{e^{j\tau}}{N} - \widetilde{C}_j^{pair}\right| + \left|\widetilde{C}_j^{pair} - s\sum_{i\in S^j}D(S_i)\right| \tag{16}$$

$$\leq 4\tau s\cdot\left|S^j\right|\cdot\frac{e^{j\tau}}{N} + 4\tau s\sum_{i\in S^j}D(S_i) \tag{17}$$

Rearranging Inequality (16):

$$s\sum_{i\in S^j}D(S_i) \in s\cdot\left|S^j\right|\cdot\frac{e^{j\tau}}{N}\left[\frac{1-4\tau}{1+4\tau},\frac{1+4\tau}{1-4\tau}\right]$$

Likewise:

$$\left|s^2\cdot\left|S^j\right|\left(\frac{e^{j\tau}}{N}\right)^2 - s^2\sum_{i\in S^j}(D(S_i))^2\right| \leq +4\tau s^2\cdot\left|S^j\right|\left(\frac{e^{j\tau}}{N}\right)^2 + 4\tau s^2\sum_{i\in S^j}(D(S_i))^2 \tag{18}$$

Similarly, for Inequality (18):

$$s^2\sum_{i\in S^j}(D(S_i))^2 \in s^2\cdot\left|S^j\right|\left(\frac{e^{j\tau}}{N}\right)^2\left[\frac{1-4\tau}{1+4\tau},\frac{1+4\tau}{1-4\tau}\right]$$

And through the relation $\frac{1-4\tau}{1+4\tau}\geq 1-10\tau$ and $\frac{1+4\tau}{1-4\tau}\leq 1+10\tau$ that holds for all $\tau>0$, we get the desired result.

□

Definition 4.9. *Define the distribution U_{S^j} to be the uniform distribution over S^j.*

Claim 4.10. *Assuming the verifier didn't reject, with probability at least 0.8 over the choice of T and T', for every j such that $\left|S^j\right| \geq e^{-j\tau}\cdot s\cdot\frac{\varepsilon\cdot\tau}{100\log N}$ and $\frac{e^{j\tau}}{N}\geq\frac{\sigma}{1000N}$, and for which Inequalities (10) and (11) hold. It further holds that:*

$$\mathbb{E}_{i\sim U_{S^j}}[D(S_i)] \in \frac{e^{j\tau}}{N}[1-10\tau, 1+10\tau]$$

$$Var_{i\sim U_{S^j}}[D(S_i)] \leq 60\tau\left(\mathbb{E}_{i\sim U_{S^j}}[D(S_i)]\right)^2$$

Proof. With high probability, for all j as specified in the claim statement, by Claim 4.8:

$$\mathbb{E}_{i\sim U_{S^j}}\left[D(S_i)\right] = \sum_{i\in S_i}\frac{1}{|S^j|}D(S_i) \in \frac{e^{j\tau}}{N}\cdot[1-10\tau, 1+10\tau]$$

Furthermore:

$$\begin{aligned}
\mathbb{E}_{i\sim U_{S^j}}\left[\left(D(S_i)\right)^2\right] &= \frac{1}{|S^j|}\sum_{i\in S^j}\left(D(S_i)\right)^2 && (19)\\
&\leq (1+10\tau)\left(\frac{e^{j\tau}}{N}\right)^2 && (20)\\
&\leq (1+10\tau)\left(\mathbb{E}_{i\sim U_{S^j}}\left[D(S_i)\right]\right)^2\frac{1}{(1-10\tau)^2} && (21)\\
&\leq (1+40\tau)\left(\mathbb{E}_{i\sim U_{S^j}}\left[D(S_i)\right]\right)^2 && (22)
\end{aligned}$$

And so, we conclude that:

$$\begin{aligned}
\mathrm{Var}_{i\sim U_{S^j}}\left[D(S_i)\right] &= \mathbb{E}_{i\sim U_{S^j}}\left[\left(D(S_i)\right)^2\right] - \left(\mathbb{E}_{i\sim U_{S^j}}\left[D(S_i)\right]\right)^2\\
&\leq (1+40\tau)\left(\mathbb{E}_{i\sim U_{S^j}}\left[D(S_i)\right]\right)^2 - (1-20\tau)\left(\mathbb{E}_{i\sim U_{S^j}}\left[D(S_i)\right]\right)^2\\
&\leq 60\tau\left(\mathbb{E}_{i\sim U_{S^j}}\left[D(S_i)\right]\right)^2
\end{aligned}$$

□

Claim 4.11. *Assuming the verifier didn't reject, with probability at least* 0.8 *over the choice of* T *and* T'*, for every* j *such that* $\left|S^j\right| \geq e^{-j\tau}\cdot s\cdot\frac{\varepsilon\cdot\tau}{100\log N}$ *and* $\frac{e^{j\tau}}{N}\geq\frac{\sigma}{1000N}$*, and for which Inequalities* (10) *and* (11) *hold, it further holds that:*

$$\mathbb{E}_{i\sim U_{S^j}}\left[\min\left\{\frac{D(S_i)}{\pi_i}, \frac{\pi_i}{D(S_i)}\right\}\right] \geq 1-\frac{\sigma}{50} \tag{23}$$

Proof. By Claim 4.10, for every j as specificied in the claim statement, it holds that:

$$\mathbb{E}_{i\sim U_{S^j}}\left[D(S_i)\right] \in \frac{e^{j\tau}}{N}\left[1-10\tau, 1+10\tau\right]$$

$$\mathrm{Var}_{i\sim U_{S^j}}\left[D(S_i)\right] \leq 60\tau\left(\mathbb{E}_{i\sim U_{S^j}}\left[D(S_i)\right]\right)^2$$

Therefore, through Chebychev's Inequality:

$$\begin{aligned}
\Pr_{i\sim U_{S^j}}\left(\left|D(S_i)-\mathbb{E}\left[D(S_i)\right]\right| \geq \sqrt{\frac{6000\tau}{\sigma}}\cdot\mathbb{E}\left[D(S_i)\right]\right) &\leq \frac{60\tau\left(\mathbb{E}_{i\sim U_{S^j}}\left[D(S_i)\right]\right)^2}{\left(\mathbb{E}_{i\sim U_{S^j}}\left[D(S_i)\right]\right)^2\cdot 6000\tau/\sigma}\\
&\leq \frac{\sigma}{100}
\end{aligned}$$

Observe that with probability at least $1 - \frac{\sigma}{100}$ over the choice of $i \sim U_{S^j}$ it holds that:

$$\begin{aligned}
|D(S_i) - \pi_i| &\leq |D(S_i) - \mathbb{E}[D(S_i)]| + \left|\mathbb{E}[D(S_i)] - \frac{e^{j\tau}}{N}\right| + \left|\frac{e^{j\tau}}{N} - \pi_i\right| \\
&\leq \sqrt{\frac{6000\tau}{\sigma}} \cdot \mathbb{E}[D(S_i)] + \left|\mathbb{E}[D(S_i)] - \frac{e^{j\tau}}{N}\right| + (e^{\tau} - 1) \cdot \frac{e^{j\tau}}{N} \\
&\leq \sqrt{\frac{6000\tau}{\sigma}} \cdot \frac{e^{j\tau}}{N}(1 + 10\tau) + 12\tau \cdot \frac{e^{j\tau}}{N} \\
&\leq \left(2\sqrt{\frac{6000\tau}{\sigma}} + 12\tau\right)\frac{e^{j\tau}}{N} \\
&\leq e^{\tau}\left(2\sqrt{\frac{6000\tau}{\sigma}} + 12\tau\right)\pi_i \\
&\leq \left(3\sqrt{\frac{6000\tau}{\sigma}} + 12\tau\right)\pi_i
\end{aligned}$$

Where the second to last inequality stems from the fact that by definition for all $i \in S^j$, $\pi_i \in \left[\frac{e^{j\tau}}{N}, \frac{e^{(j+1)\tau}}{N}\right]$. We conclude that for all such i it holds that:

$$\frac{D(S_i)}{\pi_i} \in \left[1 - 3\sqrt{\frac{6000\tau}{\sigma}} - 12\tau, 1 + 3\sqrt{\frac{6000\tau}{\sigma}} + 12\tau\right]$$

By choice of τ, this implies that with probability at least $1 - \frac{1}{100\sigma}$ over the choice of $i \sim U_{S^j}$, it holds that:

$$\frac{D(S_i)}{\pi_i} \in \left[1 - \frac{\sigma}{100}, 1 + \frac{\sigma}{100}\right]$$

Next, since for all i by definition $\min\left\{\frac{D(S_i)}{\pi_i}, \frac{\pi_i}{D(S_i)}\right\} \leq 1$, we get that for all j as specified in the claim statment, with probability at least 0.8 over the choice of T and T' if the verifier didn't reject, it holds that:

$$\mathbb{E}_{i \sim U_{S^j}}\left[\min\left\{\frac{D(S_i)}{\pi_i}, \frac{\pi_i}{D(S_i)}\right\}\right] \geq \frac{\sigma}{100} + \left(1 - \frac{\sigma}{100}\right)\left(1 - \frac{\sigma}{100}\right) \geq 1 - \frac{\sigma}{50}$$

□

Claim 4.12. *Assume the prover's tags satisfy the following inequality:*

$$\frac{1}{s}\sum_{i \in [s]: \pi_i \geq \frac{\sigma}{1000N}}\left(1 - \min\left\{\frac{\pi_i}{D(z_i)}, \frac{D(z_i)}{\pi_i}\right\}\right) \geq \sigma \tag{24}$$

Then, there exists some j_0 such that $\left|S^{j_0}\right| \geq s \cdot e^{-j_0\tau} \cdot \frac{\varepsilon \cdot \tau}{100 \log N}$ and $\frac{e^{j_0\tau}}{N} \geq \frac{\varepsilon}{100N}$, and:

$$\mathbb{E}_{i \sim U_{S^{j_0}}}\left[\min\left\{\frac{D(S_i)}{\pi_i}, \frac{\pi_i}{D(S_i)}\right\}\right] \leq 1 - 0.7\sigma \tag{25}$$

Proof. We decompose the sum in Inequality (5) according to alleged buckets as follows:

$$\begin{aligned}\sigma &\leq \frac{1}{s}\sum_{i\in[s]:\pi_i\geq\frac{\sigma}{1000N}}\left(1-\min\left\{\frac{\pi_i}{D(z_i)},\frac{D(z_i)}{\pi_i}\right\}\right)\\ &=\frac{1}{s}\sum_{j:|S^j|\neq\phi}\sum_{i\in S^j}\left(1-\min\left\{\frac{\pi_i}{D(z_i)},\frac{D(z_i)}{\pi_i}\right\}\right)\\ &=\sum_{j:|S^j|\neq\phi}\frac{|S^j|}{s}\cdot\frac{1}{|S^j|}\sum_{i\in S^j}\left(1-\min\left\{\frac{\pi_i}{D(z_i)},\frac{D(z_i)}{\pi_i}\right\}\right)\\ &=\sum_{j:|S^j|\neq\phi}\frac{|S^j|}{s}\cdot\mathbb{E}_{i\sim U_{S^j}}\left[1-\min\left\{\frac{D(S_i)}{\pi_i},\frac{\pi_i}{D(S_i)}\right\}\right]\end{aligned}$$

Define $J=\left\{j:|S^j|\geq e^{-j\tau}\cdot s\cdot\frac{\varepsilon\cdot\tau}{100\log N},\frac{e^{j\tau}}{N}\geq\frac{\sigma}{1000N}\right\}$, and denote $\sum_{j\notin J}\frac{|S^j|}{s}=\alpha$. Define next $J^c=\left\{j:0<|S^j|<e^{-j\tau}\cdot s\cdot\frac{\varepsilon\cdot\tau}{100\log N},\frac{e^{j\tau}}{N}\geq\frac{\sigma}{1000N}\right\}$. Observe that:

$$\sum_{j\in J^c}\frac{|S^j|}{s}\leq\frac{1}{s}\sum_{j\in J_1^c}e^{-j\tau}\cdot s\cdot\frac{\varepsilon\cdot\tau}{100\log N}\leq\sum_{j\in J_1^c}\frac{100}{\sigma}\cdot\frac{\varepsilon\cdot\tau}{100\log N}\leq\frac{\sigma}{20}$$

Then:

$$\sum_{j\in J}\frac{|S^j|}{s}\cdot\mathbb{E}_{i\sim U_{S^j}}\left[1-\min\left\{\frac{D(S_i)}{\pi_i},\frac{\pi_i}{D(S_i)}\right\}\right]\geq 0.7\sigma$$

Consider thus the distribution $\mathcal{B}$ that assigns to every $j\in J$ the probability $\left|\frac{S^j}{s\cdot(1-\alpha)}\right|$, and 0 otherwise. Then:

$$\mathbb{E}_{j\sim\mathcal{B}}\left[\mathbb{E}_{i\sim U_{S^j}}\left[1-\min\left\{\frac{D(S_i)}{\pi_i},\frac{\pi_i}{D(S_i)}\right\}\right]\right]\geq\sigma-\frac{\sigma}{20}\geq 0.9\sigma$$

And so it must hold that there exists some $j_0\in J$ such that:

$$\mathbb{E}_{i\sim U_{S^{j_0}}}\left[1-\min\left\{\frac{D(S_i)}{\pi_i},\frac{\pi_i}{D(S_i)}\right\}\right]\geq 0.9\sigma$$

Finally, this implies that for j_0:

$$\mathbb{E}_{i\sim U_{S^{j_0}}}\left[\min\left\{\frac{D(S_i)}{\pi_i},\frac{\pi_i}{D(S_i)}\right\}\right]\leq 1-0.9\sigma$$

□

Claim 4.13. *With high probability over the choice of S,T,T', if Inequality (5) holds, then, with high probability, the verifier rejects.*

Proof. Assume the prover's response $(\pi_i)_{i\in[s]}$ satisfies Inequality (24). Then, by Claim 4.12, it holds that there exists some j_0 such that $\left|S^{j_0}\right| \geq s \cdot \frac{\varepsilon\cdot\tau}{100\log N}$ and $\frac{e^{j_0\tau}}{N} \geq \frac{\varepsilon}{100N}$, and for which:

$$\mathbb{E}_{i\sim U_{S^{j_0}}}\left[\min\left\{\frac{D(S_i)}{\pi_i}, \frac{\pi_i}{D(S_i)}\right\}\right] \leq 1 - 0.9\sigma \tag{26}$$

Next, by Claim 4.6, with probability at least 0.98 over the choice of T, T', Inequalities (10) and (13) hold for j_0. Then, assuming the verifier didn't reject, by Claim 4.11 it holds that with probability at least 0.8 over the choice of T, T' that:

$$\mathbb{E}_{i\sim U_{S^{j_0}}}\left[\min\left\{\frac{D(S_i)}{\pi_i}, \frac{\pi_i}{D(S_i)}\right\}\right] \geq 1 - \frac{\sigma}{50} \tag{27}$$

Note that Inequality (26) and Inequality (27) contradict one another, from which we conclude that if the prover's response satisfies Inequality (24), then with probability at least 0.75 over the choice of S, T, T', the verifier should reject.

□

Finally, concerning the final verifier test:

Claim 4.14. *If the prover's answer didn't result with the verifier rejecting the test in Step 4 of Protocol 4.1.1, then with probability at most* 0.01*, Inequality* (6) *holds.*

Proof. By Proposition A.1 $\mathbb{E}\left[\widetilde{C}_{-\infty}\right] = s\sum_{i:\pi_i<\frac{\sigma}{1000N}} D(x)$. Thus, assuming that Inequality (6) holds, every entry in T has probability at least $\sum_{i\in[s]:\pi_i<\frac{\sigma}{1000N}} D(x) \geq s\cdot\frac{\sigma}{10N}$ of landing on $S^{-\infty}$, and by Hoeffdings Inequality, this will yield:

$$\widetilde{C}_{-\infty} \in (1+\frac{1}{\sqrt{s}})s^2\cdot\frac{\sigma}{10N} > s\cdot\left|S^{-}\right|\cdot\frac{\sigma}{50N}$$

And the verifier rejects with high probability. □

4.3 From verified uniform tagged sample to property verification

Lemma 4.15. *Fix two distributions D, Q over domain $[N]$, and parameter $\sigma \in (0,1)$. Let $(z_i)_{i\in[s]}$ be a sample of size $s = \widetilde{O}(N^{2/3})\mathsf{poly}(\sigma^{-1})$ drawn uniformly from $[N]$. There exists an algorithm that runs in time $O(s)$ and outputs $\delta \in [0,1]$, such that $|\delta - \Delta_{SD}(Q,D)| = O\left(\sigma + \frac{1}{\sqrt{s}}\right)$, given the following input:*

- *The sample $(z_i)_{i\in[s]}$.*
- *$(\pi_i)_i \in [0,1]^s$, that satisfy the following two inequalities:*

$$\frac{1}{s}\sum_{i\in[s]:\pi_i\geq\frac{\sigma}{1000N}}\left(1-\min\left\{\frac{\pi_i}{D(z_i)}, \frac{D(z_i)}{\pi_i}\right\}\right) \leq \sigma \tag{28}$$

$$\frac{1}{s}\sum_{i\in[s]:\pi_i\leq\frac{\sigma}{1000N}} D(z_i) \leq \frac{\sigma}{10N} \tag{29}$$

- $Q(z_i)$, *for all* $i \in [s]$.

Proof. Consider the following algorithm: for every $i \in [S]$, set $\theta'_{z_i} = \frac{|\pi_i - Q(z_i)|}{2}$, and output $\delta = \frac{1}{s}\sum_{i\in[s]} \theta'_{z_i}$. We show that this algorithm satisfies the conditions of the lemma.

For every $x \in [N]$ define $\theta_x = \frac{|D(x)-Q(x)|}{2}$. Observe that by definition, $\Delta_{\mathrm{SD}}(D,Q) = \mathbb{E}_{x\sim U_{[N]}}[\theta_x]$. Since the sample (z_i) was drawn i.i.d., the collection (θ_x) is independent. By Hoeffding's Inequality:

$$\Pr_S\left(\left|\frac{1}{s}\sum_{i\in[s]}\theta_{z_i} - \Delta_{\mathrm{SD}}(D,Q)\right| > \frac{2}{\sqrt{s}}\right) \leq 2e^{-8} < 0.01$$

And so, with probability at least 0.99 over the choice of (z_i):

$$\left|\frac{1}{s}\sum_{i\in[s]}\theta_{z_i} - \Delta_{\mathrm{SD}}(D,Q)\right| \leq \frac{2}{\sqrt{s}} \tag{30}$$

By assumption over (π_i) and the Triangle Inequality:

$$\left|\frac{1}{s}\sum_{i\in[s]}\theta_{z_i} - \frac{1}{s}\sum_{i\in[s]}\theta'_{z_i}\right| \leq \frac{1}{s}\left|\sum_{i\in[s]}\left(\frac{|D(z_i)-Q(z_i)|}{2} - \frac{|\pi_i - Q(z_i)|}{2}\right)\right| \tag{31}$$

$$\leq \frac{1}{2s}\sum_{i\in[s]}\left|\left(|D(z_i)-Q(z_i)| - |\pi_i - Q(z_i)|\right)\right| \tag{32}$$

$$\leq \frac{1}{2s}\sum_{i\in[s]}\left|\left(D(z_i)-Q(z_i)\right) - \left(\pi_i - Q(z_i)\right)\right| \tag{33}$$

$$= \frac{1}{2s}\sum_{i\in[s]}\left|D(z_i) - \pi_i\right| \tag{34}$$

For every i such that $D(z_i) \neq 0$, it holds that save for at most σ-fraction of $i \in [s]$, $\pi_i \in (1 \pm O(\sigma))\, D(z_i)$, and for every i such that $D(z_i) \neq 0$, it must hold that $\frac{1}{s}\sum_{i\in[s]:D(z_i)=0}\pi_i \leq \sigma$. And so:

$$\frac{1}{2s}\sum_{i\in[s]}|D(z_i)-\pi_i| \leq \frac{1}{2s}\sum_{i\in[s]:D(z_i)\neq 0} D(z_i)\left|1 - \frac{\pi_i}{D(z_i)}\right| + \frac{1}{2s}\sum_{i\in[s]:D(z_i)=0}\pi_i \tag{35}$$

$$\leq \frac{1}{2}\left(\frac{1}{s}\sum_{i\in[s]:D(z_i)\overset{O(\sigma)}{\approx}\pi_i} D(z_i)\left|1 - \frac{\pi_i}{D(z_i)}\right| + \frac{1}{s}\sum_{i\in[s]:D(z_i)\overset{O(\sigma)}{\not\approx}\pi_i} D(z_i)\left|1 - \frac{\pi_i}{D(z_i)}\right| + \sigma\right) \tag{36}$$

$$\leq \frac{1}{2}\left(O(\sigma) + O(\sigma) + \sigma\right) \tag{37}$$

$$= O(\sigma) \tag{38}$$

We thus conclude that with high probability over (z_i), the algorithm yields δ such that: $|\delta - \Delta_{\mathrm{SD}}(D,Q)| = O(\sigma + \frac{1}{\sqrt{s}})$ □

An immediate corollary of this lemma is Theorem 1.3. We note here that this method can also be leveraged to achieve an efficient protocol for identity testing from an approximate tagged sample *drawn according to* D, and so, can be also implemented on the output of [HR23] without incurring further overhead.

We now address the question of verification of label-invariant distribution problems. First, we recall the following definition:

Definition 4.16 (Efficient approximate decision procedure, [HR22])**.** *A distribution property $\mathcal{P}$ has a* μ-efficient approximate decision procedure *if there exists a polynomial-time procedure A as follows. A gets as input the domain size N, a distance parameter $\sigma \in (0,1)$, and a histogram $(m_j)_j$ satisfying $\sum_j \left| m_j - Q(B_j^Q) \right| \leq \mu$. For every integer N, every distribution D over $[N]$ and every $\sigma > 0$:*

- *If Q is in $\mathcal{P}$, then A accepts the $(m_j)_j$.*
- *A rejects every $(m_j)_j$ histogram that is consistent with a distribution that is not σ-close to $\mathcal{P}$.*

Corollary 4.17. *Let $\mathcal{P}$ be a label-invariant distribution property, $0 \leq \varepsilon_c < \varepsilon_f \leq 1$ distance parameters, and assume $\mathcal{P}$ admits an efficient τ-approximate decision procedure, where $\tau = O\left(\varepsilon_f - \varepsilon_c\right)^3$. Given sample access to distribution D over domain $[N]$, there exists a 2-message* public-coin *protocol with verifier sample complexity and communication complexity $\widetilde{O}(N^{2/3}) \cdot \mathsf{poly}(\tau^{-1})$, such that:*

- ***Completeness.*** *If $\Delta_{SD}(D, \mathcal{P}) \leq \varepsilon_c$, the verifier accepts with high probability.*
- ***Soundness.*** *If $\Delta_{SD}(D, \mathcal{P}) \geq \varepsilon_f$, the verifier rejects with high probability.*

We outline how to obtain a protocol for every label-invariant distribution property admitting an efficient decision procedure from a uniform verified tagged sample. Generally, we follow [HR22]. The reader is referred to their work for further detail on efficient decision procedures, as well as examples for such procedures for natural label-invariant properties, such as those relating to Shannon entropy, support size, and distance from uniformity. We note that the main obstacle in the protocol behind the above corollary, addressed by this paper in a novel way, is obtaining a good approximation of the probability according to D of randomly chosen elements in the domain. Recall that without communication, this task requires $\widetilde{O}(N)$ samples and runtime from the verifier.

We provide an outline the protocol behind Corollary 4.17. The verifier and the prover run Protocol 4.1.1 over distribution D with distance parameter $\sigma = \frac{\varepsilon_f - \varepsilon_c}{3}$, and with the following addition: the prover also sends, alongside $(\pi_i)_{i \in [s]}$, the tags $(q_i)_{i \in [s]}$, such that for all $i \in [s]$, $q_i = Q(i)$, for some distribution $Q \in \mathcal{P}$. The verifier performs the following checks:

- The verifier runs the tests outlined in Protocol 4.1.1 with respect to (π_i), and rejects w.h.p. if prover tags satisfy Inequalities (5) or (6).
- The verifier uses (q_i) to compute the bucket histogram of distribution Q. This is done by noting that the size of every bucket of significant mass j of Q can be approximated to high accuracy from a uniform tagged sample (q_i). Then, the mass of each bucket be approximated as well by multiplying the size by $\frac{e^{j\tau}}{N}$. Note that this process yields a probability histogram for Q that is accurate with high probability up to τ multiplicative factor. Then, the verifier runs the τ-approximate decision procedure with distance parameter σ, to check that indeed $Q \in \mathcal{P}_N$, and reject if it's far. Note that if indeed $Q \in \mathcal{P}_N$, and since the histogram is τ accurate, the verifier accepts with high probability, and rejects if Q is not inside the property.

- If non of the above tests failed, the verifier estimates the distance between Q and D using (π_i) and (q_i) as outlined in Lemma 4.15, and rejects unless this estimate is smaller than $\varepsilon_c + O(\tau)$.

If the all tests passed, then with high probability it holds that Q is τ-close to $\mathcal{P}$, and that $\Delta_{\mathrm{SD}}(Q, D) \leq \varepsilon_c + O(\tau)$, and the conditions of Corollary 4.17 hold. If D is ε_f far from the property, and the histogram of Q is consistent with a histogram that passes the efficient decision procedure, then by assumption, it holds that Q is $\varepsilon_f - \tau$ far from D, and so the distance test will fail. We omit further detail, as the process of verifying membership in distribution property from approximate histogram is outlined in [HR22].

References

[AR21] Gal Arnon and Guy N. Rothblum. On prover-efficient public-coin emulation of interactive proofs. In Stefano Tessaro, editor, *2nd Conference on Information-Theoretic Cryptography, ITC 2021, July 23-26, 2021, Virtual Conference*, volume 199 of *LIPIcs*, pages 3:1–3:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

[BC17] Tugkan Batu and Clément L. Canonne. Generalized uniformity testing. In Chris Umans, editor, *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pages 880–889. IEEE Computer Society, 2017.

[BGG+88] Michael Ben-Or, Oded Goldreich, Shafi Goldwasser, Johan Håstad, Joe Kilian, Silvio Micali, and Phillip Rogaway. Everything provable is provable in zero-knowledge. In Shafi Goldwasser, editor, *Advances in Cryptology - CRYPTO '88, 8th Annual International Cryptology Conference, Santa Barbara, California, USA, August 21-25, 1988, Proceedings*, volume 403 of *Lecture Notes in Computer Science*, pages 37–56. Springer, 1988.

[BM88] László Babai and Shlomo Moran. Arthur-merlin games: A randomized proof system, and a hierarchy of complexity classes. *J. Comput. Syst. Sci.*, 36(2):254–276, 1988.

[CG18] Alessandro Chiesa and Tom Gur. Proofs of proximity for distribution testing. In Anna R. Karlin, editor, *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, volume 94 of *LIPIcs*, pages 53:1–53:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018.

[EKR04] Funda Ergün, Ravi Kumar, and Ronitt Rubinfeld. Fast approximate probabilistically checkable proofs. *Inf. Comput.*, 189(2):135–159, 2004.

[FS86] Amos Fiat and Adi Shamir. How to prove yourself: Practical solutions to identification and signature problems. In Andrew M. Odlyzko, editor, *Advances in Cryptology - CRYPTO '86, Santa Barbara, California, USA, 1986, Proceedings*, volume 263 of *Lecture Notes in Computer Science*, pages 186–194. Springer, 1986.

[GMR85] Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The knowledge complexity of interactive proof-systems (extended abstract). In Robert Sedgewick, editor, *Proceedings of the 17th Annual ACM Symposium on Theory of Computing, May 6-8, 1985, Providence, Rhode Island, USA*, pages 291–304. ACM, 1985.

[GMW91] Oded Goldreich, Silvio Micali, and Avi Wigderson. Proofs that yield nothing but their validity for all languages in NP have zero-knowledge proof systems. *J. ACM*, 38(3):691–729, 1991.

[Gol17] Oded Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017.

[GR18] Tom Gur and Ron D. Rothblum. Non-interactive proofs of proximity. *Comput. Complex.*, 27(1):99–207, 2018.

[GS86] Shafi Goldwasser and Michael Sipser. Private coins versus public coins in interactive proof systems. In Juris Hartmanis, editor, *Proceedings of the 18th Annual ACM Symposium on Theory of Computing, May 28-30, 1986, Berkeley, California, USA*, pages 59–68. ACM, 1986.

[HR22] Tal Herman and Guy N. Rothblum. Verifying the unseen: interactive proofs for label-invariant distribution properties. In Stefano Leonardi and Anupam Gupta, editors, *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*, pages 1208–1219. ACM, 2022.

[HR23] Tal Herman and Guy N. Rothblum. Doubley-efficient interactive proofs for distribution properties. In *64th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2023, Santa Cruz, CA, USA, November 6-9, 2023*, pages 743–751. IEEE, 2023.

[HR24] Tal Herman and Guy N. Rothblum. Interactive proofs for general distribution properties. *Electron. Colloquium Comput. Complex.*, pages TR24–094, 2024.

[PRR06] Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *J. Comput. Syst. Sci.*, 72(6):1012–1042, 2006.

[RRSS09] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam D. Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM J. Comput.*, 39(3):813–842, 2009.

[RVW13] Guy N. Rothblum, Salil P. Vadhan, and Avi Wigderson. Interactive proofs of proximity: delegating computation in sublinear time. In Dan Boneh, Tim Roughgarden, and Joan Feigenbaum, editors, *Symposium on Theory of Computing Conference, STOC'13, Palo Alto, CA, USA, June 1-4, 2013*, pages 793–802. ACM, 2013.

[Vad00] Salil P. Vadhan. On transformation of interactive proofs that preserve the prover's complexity. In F. Frances Yao and Eugene M. Luks, editors, *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing, May 21-23, 2000, Portland, OR, USA*, pages 200–207. ACM, 2000.

[VV11] Gregory Valiant and Paul Valiant. The power of linear estimators. In Rafail Ostrovsky, editor, *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*, pages 403–412. IEEE Computer Society, 2011.

A Collision Tests Analysis

A.1 Twoway Collisions

Proposition A.1. *Assume that for every $x \in [N]$, $D(x) \leq \frac{1}{s}$. For every sample S such that for every $i \in [s]$, the element S_i appears at most $\log N$ times in S, with probability at least $1 - \frac{\tau}{100 \log N}$ over the choice of the sample T, it holds that:*

$$\mathbb{E}\left[\widetilde{C}_j^{pair}\right] = s \sum_{i \in S^j} D(S_i)$$

As well as:

$$\left|\widetilde{C}_j^{pair} - \mathbb{E}\left[\widetilde{C}_j^{pair}\right]\right| \leq \mathbb{E}\left[\widetilde{C}_j^{pair}\right] \cdot \sqrt{\frac{300 \log^2 N}{\tau \cdot \mathbb{E}\left[\widetilde{C}_j^{pair}\right]}}$$

Proof. For every $k, r \in [s]$ denote by $C_{k,r}$ the indicator of the event $\{S_k = T_r\}$. Observe that $\widetilde{C}_j^{pair} = \sum_{k \in S^j} \sum_{r \in [s]} C_{k,r}$, and that $\mathbb{E}_T[C_{k,r}] = D(S_k)$. By the linearity of expectation:

$$E\left[\widetilde{C}_j^{pair}\right] = \sum_{k \in S^j} \sum_{r \in [s]} \mathbb{E}[C_{k,r}] = \sum_{k \in S^j} \sum_{r,r' \in [s]} D(S_k) = s \sum_{k \in S^j} D(S_k) \tag{39}$$

In order to prove concentration we show that $\mathrm{Var}_T\left[\widetilde{C}_j^{pair}\right]$ is small. Note that:

$$\mathrm{Var}_T\left[\widetilde{C}_j^{pair}\right] = \sum_{(k,r) \in [s]^2} \sum_{(k',r') \in [s]^2} \mathrm{Cov}\left[C_{k,r}, C_{k',r'}\right]$$

In order to bound the variance consider the following case analysis for $\mathrm{Cov}\left[C_{k,r}, C_{k',r'}\right]$:

- **Type I.** Assume $S_k \neq S_{k'}$. Then, if $r \neq r'$, the random variables $C_{k,r}$ and $C_{k',r'}$ are independent, and $\mathrm{Cov}\left[C_{k,r}, C_{k',r'}\right] = 0$; otherwise, $r = r'$, in which case, it cannot be that $C_{k,r} = C_{k',r'} = 1$ simultaneously, and $\mathrm{Cov}\left[C_{k,r}, C_{k',r'}\right] < 0$.
- **Type II.** Assume $S_k = S_{k'}$. Then, if $r = r'$ we get that $\mathrm{Cov}\left[C_{k,r}, C_{k',r'}\right] = \Pr_T(S_r = S_k) = D(S_k) = E[C_{k,r}]$. Otherwise, if $r \neq r'$, $\mathrm{Cov}\left[C_{k,r}, C_{k',r'}\right] = \Pr_T(S_r = S_{r'} = S_k) = (D(S_k))^2 \leq \frac{1}{s} D(S_k)$

Where the last inequality stems from the assumption that $D(x) \leq \frac{1}{s}$ for all $x \in [N]$. We thus conclude that:

$$\mathrm{Var}\left[\widetilde{C}_j\right] \leq \sum_{k \in S^j} \sum_{k' \in S^j : S_k = S_{k'}} \sum_{r \in [s]} E[C_{k,r}] + \sum_{k \in S^j} \sum_{k' \in S^j : S_k = S_{k'}} \sum_{r \neq r' \in [s]} \frac{1}{s} D(S_k) \leq \log N \sum_{k \in S^j} \sum_{r \in [s]} E[C_{k,r}] = \log N \mathbb{E}\left[\widetilde{C}_j\right]$$

The desired result is thus achieved through Chebychevs' Inequality. □

A.2 Threeway Collisions

Proposition A.2. *Assume that for every $x \in [N]$, $D(x) \leq \frac{1}{s}$. For every sample $S = (S_i)_{i\in[s]}$ such that for every $i \in [s]$, the element S_i appears in at most $\log N$ locations in S, with probability at least $1 - \frac{\tau}{100 \log N}$ over the choice of the samples T, T', it holds that for any set of bucket indices J of size at most $2\frac{\log N}{\tau}$, for every $j \in J$:*

$$\mathbb{E}\left[\widetilde{C}_j^{triple}\right] = s^2 \sum_{i\in S^j} (D(x))^2$$

As well as:

$$\left|\widetilde{C}_j^{triple} - \mathbb{E}\left[\widetilde{C}_j^{triple}\right]\right| \leq \mathbb{E}\left[\widetilde{C}_j^{triple}\right] \cdot \sqrt{\frac{300 \log^2 N}{\tau \cdot \mathbb{E}\left[\widetilde{C}_j^{triple}\right]}}$$

Proof. For every $k, r, r' \in [s]$ denote by $C_{k,r,r'}$ the indicator of the event $\left\{S_k = T_r = T'_{r'}\right\}$. Observe that $\widetilde{C}_j^{triple} = \sum_{k\in S^j} \sum_{r,r'\in[s]} C_{k,r,r'}$, and that $\mathbb{E}_{T,T'}\left[C_{k,r,r'}\right] = (D(S_k))^2$. By the linearity of expectation:

$$E\left[\widetilde{C}_j^{triple}\right] = \sum_{k\in S^j} \sum_{r,r'\in[s]} E_{T,T'}\left[C_{k,r,r'}\right] = \sum_{k\in S^j} \sum_{r,r'\in[s]} (D(S_k))^2 = s^2 \sum_{k\in S^j} (D(S_k))^2 \tag{40}$$

Next, we show that for every $j \in J$ the random variable $\widetilde{C}_j^{triple}$ is well concentrated around its mean. In order to do so, we bound the variance of $\widetilde{C}_j^{triple}$. Note that:

$$\mathrm{Var}\left[\widetilde{C}_j^{triple}\right] = \sum_{\substack{(k_0,r_0,r'_0)\in[s]^3 \\ (k_1,r_1,r'_1)\in[s]^3}} \mathrm{Cov}\left[C_{k_0,r_0,r'_0}, C_{k_1,r_1,r'_1}\right]$$

And so, in order to bound the variance, consider the following case analysis for the pair $((k_0, r_0, r'_0), (k_1, r_1, r'_1))$:

- **Type I.** $S_{k_0} \neq S_{k_1}$, then: either $r_0 \neq r_1$ and $r'_0 \neq r'_1$ in which case $\mathrm{Cov}\left[C_{k_0,r_0,r'_0}, C_{k_1,r_1,r'_1}\right] = 0$ as the variables are independent; or $r_0 = r_1$ or $r'_0 = r'_1$, in which case since $S_{k_0} \neq S_{k'_1}$, it cannot be that $C_{k_0,r_0,r'_0} = 1$ and $C_{k_1,r_1,r'_1} = 1$ simultaneously, which means that $\mathrm{Cov}\left[C_{k_0,r_0,r'_0}, C_{k_1,r_1,r'_1}\right] < 0$.
- **Type II.** $S_{k_0} = S_{k_1}$ and $(r_0, r'_0) = (r_1, r'_1)$, then $\mathrm{Cov}\left[C_{k_0,r_0,r'_0}, C_{k_1,r_1,r'_1}\right] = Var\left[C_{k_0,r_0,r'_0}\right] \leq \mathbb{E}\left[C_{k_0,r_0,r'_0}\right]$.
- **Type III.**:
  - **Type IIIa.** $S_{k_0} = S_{k_1}$ and $r_0 = r_1 = r$, however $r'_0 \neq r'_1$, then:

$$\mathrm{Cov}\left[C_{k_0,r,r'_0}, C_{k_1,r,r'_1}\right] \leq \mathbb{E}\left[C_{k_0,r,r'_0} \cdot C_{k_1,r,r'_1}\right] = (D(S_{k_0}))^3$$

– **Type IIIb.** $S_{k_0} = S_{k_1}$ and $r'_0 = r'_1 = r'$, however $r_0 \neq r_1$, then:

$$\mathrm{Cov}\left[C_{k_0,r,r'_0}, C_{k_1,r,r'_1}\right] \leq \mathbb{E}\left[C_{k_0,r_0,r'} \cdot C_{k_1,r_1,r'}\right] = \left(D(S_{k_0})\right)^3$$

Since all pairs of indicators of Type I do not contribute to the variance, we are left to quantify how many pairs of indicators are there of Type II and Type III. Fix $k_0 \in [s]$, and denote $A_{k_0} = \{i \in [s] : S_i = S_{k_0}\}$.

- **Type II.** By assumption over S, $|A_{k_0}| \leq \log N$, and so, there are at most $\log N$ options for k_1. Then, there are s^2 ways to pick (r, r'). Therefore, k_0 participates in at most $s^2 \cdot \log N$ pairs of Type II.
- **Type III.** This type is divided into two symmetric sub-types. As above, for a fixed k_0, there are at most $\log N$ possible values for k_1. Then, there are s^3 ways to pick r, r'_0, r'_1. Therefore, k_0 participates in at most $2 \cdot s^3 \cdot \log N$ pairs of Type IIIa. Type IIIb is the symmetric where both triplets agree on r', but have two different values r_0 and r_1.

First, we calculate the contribution of all the *Type II* pairs to the variance:

$$\begin{aligned}
\sum_{(k_0,r,r')\in[s]^3} \sum_{k_1 \in A_{k_0}} \mathrm{Cov}\left[C_{k_0,r,r'}, C_{k_1,r,r'}\right] &\leq \sum_{(k_0,r,r')\in[s]^3} \sum_{k_1 \in A_{k_0}} \mathbb{E}\left[C_{k_0,r,r'}\right] && (41)\\
&\leq \log N \sum_{(k_0,r,r')\in[s]^3} \mathbb{E}\left[C_{k_0,r,r'}\right] && (42)\\
&= \log N \cdot \mathbb{E}\left[\sum_{(k_0,r,r')\in[s]^3} C_{k_0,r,r'}\right] && (43)\\
&= \log N \cdot \mathbb{E}\left[\widetilde{C}_j^{triple}\right] && (44)
\end{aligned}$$

As for the *Type IIIa* pairs:

$$\begin{aligned}
\sum_{(k_0,r,r'_0,r'_1)\in[s]^4} \sum_{k_1 \in A_{k_0}} \mathrm{Cov}\left[C_{k_0,r,r'_0}, C_{k_1,r,r'_1}\right] &\leq \sum_{(k_0,r,r'_0,r'_1)\in[s]^4} \sum_{k_1 \in A_{k_0}} \left(D(S_{k_0})\right)^3 && (45)\\
&\leq \log N \sum_{(k_0,r,r'_0,r'_1)\in[s]^4} \left(D(S_{k_0})\right)^3 && (46)\\
&\leq s \cdot \log N \sum_{(k_0,r,r'_0,)\in[s]^3} \left(D(S_{k_0})\right)^3 && (47)\\
&\leq s \cdot \log N \sum_{(k_0,r,r'_0,)\in[s]^3} \left(D(S_{k_0})\right)^2 \cdot \frac{1}{s} && (48)\\
&\leq \log N \sum_{(k_0,r,r'_0,)\in[s]^3} \mathbb{E}\left[C_{k_0,r,r'_0}\right] && (49)\\
&= \log N \cdot \mathbb{E}\left[\widetilde{C}_j^{triple}\right] && (50)
\end{aligned}$$

Similarly, all *Type IIIb* contribute at most $\log N \cdot \mathbb{E}\left[\widetilde{C}_j^{triple}\right]$ to the variance as well. We thus conclude that:

$$\mathrm{Var}\left[\widetilde{C}_j^{triple}\right] \leq 3 \log N \cdot \mathbb{E}\left[\widetilde{C}_j^{triple}\right]$$

Therefore, using Chebichev's Inequality:

$$\Pr_{T,T'}\left(\left|\widetilde{C}_j^{triple} - \mathbb{E}\left[\widetilde{C}_j^{triple}\right]\right| \geq \sqrt{\frac{300\log^2 N}{\tau}\cdot\mathbb{E}\left[\widetilde{C}_j^{triple}\right]}\right) \leq \frac{3\log N\cdot\mathbb{E}\left[\widetilde{C}_j^{triple}\right]}{\frac{300\log^2 N}{\tau}\cdot\mathbb{E}\left[\widetilde{C}_j^{triple}\right]} \tag{51}$$

$$\leq \frac{\tau}{100\log N} \tag{52}$$

Taking union bound over all $j \in J$ yields the desired result. □

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il