Electronic Colloquium on Computational Complexity, Report No. 200 (2025)

On doubly-efficient interactive proofs for distributions

Oded Goldreich Guy N. Rothblum

December 3, 2025

Abstract

Interactive proofs of proximity for distributions, introduced by Chiesa and Gur (*ITCS18*) and extensively studied recently by Herman and Rothblum (*STOC22*, *FOCS23*, *FOCS24*), offer a way of verifying properties of distributions using less samples than required to test these properties.

We say that such an interactive proof system is doubly-efficient if the verifier's sample complexity is lower than the sample complexity of testing the property, and the honest-prover's sample complexity is lower than the sample complexity of learning the property. We prove a feasibility result for this notion. Specifically, we present properties of distributions for which the prover's sample complexity is close to the complexity of testing, whereas the sample complexity of the verifier is much lower.

Contents

1 Introduction

The model of *interactive proofs of proximity* (IPP) *for properties of distribution*, introduced by Chiesa and Gur [4], is a hybrid of the model of interactive proof systems and distribution testing. In an IPP (for distributions), the verifier is given a small sample from an unknown distribution, but can interact with an untrusted prover who knows the distribution. The goal of this interaction is to be convinced that the unknown distribution has some predetermined property. Hence, this proof system, comes with corresponding notions of completeness and soundness:

***Completeness*:** If the distribution has the predetermined property, then the verifier accepts with high probability (provided that the prover follows a presecribed property).

***Soundness*:** If the distribution is far from having the predetermined property, then the verifier rejects with high probability regardless of the strategy followed by the prover.

ISSN 1433-8092

The point of such an IPP is that it may allow the verifier to use a significantly smaller sample than the one required for testing (when not assisted by a prover). This remarkable potential has been materialized in a sequence of works by Herman and Rothblum [7, 8, 9].

In this paper we initiate a study of *doubly-efficient IPPs* (ds-IPP) *for properties of distributions.* In these IPPs the verifier is required to be more efficient than the corresponding tester while the "honest prover" (i.e., a prover satisfying the completeness condition) is required to be more efficient than the corresponding learning algorithm. Specifically:

- the verifier's sample complexity is lower than the sample complexity of testing the property;
- the honest-prover's sample complexity is lower than the sample complexity of learning the property.

We show that such ds-IPP exist, alas for unnatural properties of distributions. This stands in contrast to the fact, proved by Herman and Rothblum [10], that many natural properties of distributions that do have an IPP (in which the verification is more efficient than testing) do not have ds-IPPs.

We comment that IPPs for distributions are akin to IPPs for functions, where the primary complexity measure is the number of queries (rather than the number of sample-points as discussed above). Similarly, ds-IPPs for distributions are akin to IPPs for functions, whose study was initiated by Amir, Goldreich and Rothblum [1].

Some laconic credits. Distribution testing emerged explicitly in the works of Batu *et. al.* [3, 2]. Interactive proof of proximity (for functions) emerged explicitly in [11], whereas doubly-efficient interactive proofs (for functions) emerged explicitly in [6].

The rest of this paper. We present a feasibility result regarding the notion of ds-IPPs for distributions. Actually, we present two related properties of distributions for which the prover's sample complexity is close to the complexity of testing, whereas the sample complexity of the verifier is much lower. The basic idea is presented in Section 3, and its ramifications are presented in Section 4. (The main result is stated at the end of Section 4.) But, first, we formally present the relevant definitions.

2 IPPs and ds-IPPs for distributions

The notion of an interactive proof of proximity for properties of distributions is a hybrid of the notions of distribution testing and interactive proof systems. The basic mind-frame – of approximate decision and sub-linear complexity – is inherited from the former, whereas the actual setting – of verification via an interaction between a powerful but untrusted prover and a probabilistic verifier – is taken from the latter. In other words, a verifier is a tester that is assisted by interaction with an untrusted prover. In the context of properties of distributions, which is the focus of this work, we denote by $(\widetilde{P}, V(z))$ the output of the verifier V, on input z (i.e., a sample from a tested distribution), 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 1 (a verifier of proximity for the property $\mathcal{D}$): *Let $\mathcal{D} = \bigcup_{n\in\mathbb{N}} \mathcal{D}_n$ such that $\mathcal{D}_n$ is a set of distributions over $[n]$. A* verifier*, denoted V,* of sample complexity $s : \mathbb{N} \times [0,1] \to \mathbb{N}$ for the property $\mathcal{D}$ *is a probabilistic interactive machine that, on input parameters n and ϵ, and a sequence of $s(n,\epsilon)$ sample-points drawn from 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*

$$\Pr_{i_1,...,i_s\sim X}[(P, V(n,\epsilon; i_1,...,i_s)) = 1] \geq 2/3,$$

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

2. The verifier rejects distributions that are far from $\mathcal{D}$: *If X is ϵ-far from any distribution in $\mathcal{D}_n$* (i.e., X is ϵ-far from $\mathcal{D}$)*, then for every prover strategy $\widetilde{P}$ it holds that*

$$\Pr_{i_1,...,i_s\sim X}[(\widetilde{P}, V(n,\epsilon; i_1,...,i_s)) = 0] \geq 2/3,$$

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

(Hence, the sample complexity of (the verifier) V is a function (of n and ϵ) that specifies the number of sample-points that V gets, on input parameters n and ϵ.)

An alternative formulation specifies the (honest) prover strategy that is used in the completeness condition. This is called for when wishing to discuss the complexity of such honest prover strategies. Indeed, verifiers of proximity that admit an efficient proving strategy (in case X is in $\mathcal{D}$) are of natural interest, and are our actual focus. In these cases, we provide these prover strategies with samples of X and consider their sample complexity, which will be denoted $s' : \mathbb{N} \times [0,1] \to \mathbb{N}$. Hence, we use the following definition.

Definition 2 (an IPP for property $\mathcal{D}$): *For $\mathcal{D}$ and V as in Definition 1, we say that (P,V) is an* interactive proof of proximity (IPP) for $\mathcal{D}$ *if P is an interactive machine and Condition 1* (completeness) *holds with P replaced by $P(n,\epsilon; i'_1,...,i'_{s'(n)})$, for some $s' : \mathbb{N} \times [0,1] \to \mathbb{N}$; that is, for every $n \in \mathbb{N}$ and $\epsilon > 0$ and for every X in $\mathcal{D}_n$, it holds that*

$$\Pr_{i'_1,...,i'_{s'},i_1,...,i_s\sim X}[(P(n,\epsilon; i'_1,...,i'_{s'}), V(n,\epsilon; i_1,...,i_s)) = 1] \geq 2/3,$$

where $s' = s'(n,\epsilon)$ and $(s = s(n,\epsilon))$ *and $i'_1,...,i'_{s'}, i_1,...,i_s$ are drawn independently from the distribution X. In this case, we say that s' is the* sample complexity of (the honest prover) P.

We stress that V also satisfies Condition 2 (soundness), and its sample complexity is s.

We say that (P,V) is a doubly-sublinear IPP (ds-IPP) for $\mathcal{D}$ if (1) the sample complexity of V is sublinear in the sample complexity of *testing* $\mathcal{D}$, and (2) the sample complexity of P is sublinear in the sample complexity of *learning* $\mathcal{D}$. We stress that this definition is minimal: It only refers to the sample complexity of the two parties, and it only requires some advantage over the straightforward case.[1]

[1] Evidently, verification is reducible to testing, whereas proving is essentially reducible to learning the distribution (when considering the sample complexity only).

3 The basic idea

For $d \ll n$, we consider the set of degree d polynomials over a field of size n, denoted $\mathcal{F}$. For each such polynomial p, we let X_p be a distribution that is uniform over $S_p \stackrel{\rm def}{=} \{(e, p(e)) : e \in \mathcal{F}\}$, and let $\mathcal{D}$ be the set of all such distributions. We also consider $\mathcal{D}'$ that consist of all distributions with a support that is a subset of S_p for some degree d polynomial p. Note that both $\mathcal{D}$ and $\mathcal{D}'$ are properties of distributions that ranges over $\mathcal{F} \times \mathcal{F} \equiv [n]^2$.

We first note that testing $\mathcal{D}$ (resp., $\mathcal{D}'$) requires more than d sample-points. This is the case because when using at most d sample-points one cannot distinguish the following two random d-long sequences.

1. A sequence of the form $((i_1, f(i_1)), ..., (i_d, f(i_d)))$, where f is a random function from $\mathcal{F}$ to $\mathcal{F}$ and $i_1, ..., i_d$ are uniformly and independently selected in $\mathcal{F}$.

2. A sequence of the form $((i_1, p(i_1)), ..., (i_d, p(i_d)))$, where p is a random degree d polynomial over $\mathcal{F}$ and $i_1, ..., i_d$ are uniformly and independently selected in $\mathcal{F}$.

The reason is that any d values of a random degree d polynomial are uniformly and independent distributed in $\mathcal{F}$, just as is the case with a random function.

On the other hand, assuming that $d = \omega(\sqrt{n})$, a ds-IPP for $\mathcal{D}$ proceeds as follows. When given a sampling access to X_p, the honest prover uses $O(d/\epsilon^2)$ sample-points of X_p in order to reconstruct the polynomial p, which it sends to the verifier. Upon receiving p, the verifier tests whether the input distribution X is uniform on S_p; that is, the verifier takes a sample of size $O(\sqrt{n}/\epsilon^2)$ from X, and invokes the uniformity tester on X, while using a 1-1 correspondence between $[n]$ and S_p. Needless to say, the verifier rejects if it sees any sample-point that is not in S_p. (Otherwise, it rules according to the verdict of the uniformity tester.) Note that the hypothesis that $d = \omega(\sqrt{n})$ implies that the sample complexity of the verifier is lower than the sample complexity of a tester.

A ds-IPP for $\mathcal{D}'$ can also be obtained. In this case the sample (of size $O(d/\epsilon^2)$) may not determine a unique degree d polynomial, but the honest prover may just select an arbitrary degree d polynomial p that is consistent with this sample, and sends p to the verifier. Upon receiving p, the verifier takes a sample of size $O(1/\epsilon^2)$ from the input distribution X, and accepts if and only if all sample-points are in S_p. Note that in this case, regardless of the relation between d and n, the sample complexity of the verifier is lower than the sample complexity of a tester.

Summary. Both $\mathcal{D}$ and $\mathcal{D}'$ require more than d sample-points for testing, but they both have IPPs in which the honest prover uses $O(d/\epsilon^2)$ sample-points. The verifier's sample complexity is $O(\sqrt{n}/\epsilon^2)$ in case of $\mathcal{D}$, and $O(1/\epsilon^2)$ in case of $\mathcal{D}'$. Recall that each distribution in $\mathcal{D}$ is uniform over a set of size n, whereas each distribution in $\mathcal{D}'$ has support size at most n.

4 Ramifications

Beyond being artificial, the foregoing examples have two deficiencies, which we fix next.

1. The complexity of testing $\mathcal{D}$ (resp., $\mathcal{D}'$) equals the one of learning distributions in $\mathcal{D}$ (resp., $\mathcal{D}'$), whereas the focus of property testing is on cases in which testing is significantly more efficient than learning (cf., e.g., [5]).

2. For constant $\epsilon > 0$, the communication complexity of our IPP is slightly larger than the sample complexity of testing; hence, in this case, the running-time of the verifier is not lower than the running-time of the tester.

In fixing these deficiencies we use properties that are even more artificial that the foregoing ones. Still this does not harm their demonstrative feature.

4.1 Making testing significantly easier than learning

For the class $\mathcal{D}$, this can be done by augmenting the class of distributions so that the complexity of learning the resulting class becomes higher than the complexity of testing, which is maintained. For example, we may replace S_p by $S_{p,g} = \{(e, p(e), g(e)) : e \in \mathcal{F}\}$ such that $g : \mathcal{F} \to \{0,1\}$ is an arbitrary Boolean function (and let $\mathcal{D}$ consist of the set of all distributions that are uniform over some $S_{p,g}$). Recall that the complexity of learning an arbitrary Boolean function over $\mathcal{F} \equiv [n]$ is $\Omega(n)$, but testing the class of all Boolean functions is trivial.

The verifier in the ds-IPP proceeds as the basic one, except that it also checks that sample-points that are equal on the first element (i.e., e) are also equal on the third element (i.e., $g(e)$). Given that the verifier takes $O(\sqrt{n}/\epsilon^2)$ sample-points anyhow, this allows checking that the distribution of the third element fits *some* function $g : [n] \to \{0,1\}$. (In the analysis, one may assume that the tested distribution over triples induces a uniform distribution on the first element.)[2]

For the class $\mathcal{D}'$, we can augment the distribution on pairs by an arbitrary distribution on an additional bit (equiv., the foregoing function g can be replaced by an arbitrary random process $G : [n] \to \{0,1\}$). In this case, we do not augment the basic verifier. (Indeed, the same idea can be applied to obtain an alternative definition of $\mathcal{D}$, but in this case the distributions in $\mathcal{D}$ are not necessarily uniform over some n-subset.)

4.2 Significantly reducing the communication complexity of our IPP

Here a more significant modification is required. Rather than using a single univariate polynomial of degree $d = \omega(\sqrt{n})$ over a field of size n, we shall use $t = n/|\mathcal{F}|$ univariate polynomials of degree $d = O(\log n)$ over $\mathcal{F}$ such that $|\mathcal{F}| \gg d$ (e.g., $|\mathcal{F}| \geq 10d$). Specifically, for $\overline{p} = (p_1, ..., p_t)$, where each p_i is a degree d polynomial over $\mathcal{F}$, we consider the set

$$S_{\overline{p}} \stackrel{\text{def}}{=} \{(i, e, p_i(e)) : i \in [t] \ \&\ e \in \mathcal{F}\}$$

and let $\mathcal{D}$ (resp., $\mathcal{D}'$) be the set of all distributions that are uniform over some $S_{\overline{p}}$ (resp., have a support that is a subset of some $S_{\overline{p}}$). Note that $|S_{\overline{p}}| = t \cdot |\mathcal{F}| = n$, and that we can pick $\mathcal{F}$ such that $|\mathcal{F}| = O(\log n)$.

We first observe that that $\mathcal{D}$ (resp., $\mathcal{D}'$) cannot be tested using $o(t \cdot d)$ sample-points. This holds because using $m = o(t \cdot d)$ sample-points does not allow to distinguish the following two distributions over m-long sequences.

1. A sequence of the form $((i_1, e_1, f(e_1, i_1)), ..., (i_m, e_m, f(i_m, e_m)))$, where f is a random function from $[t] \times \mathcal{F}$ to $\mathcal{F}$ and $(i_1, e_1), ..., (i_m, e_m)$ are uniformly and independently selected in $[t] \times \mathcal{F}$.

[2] In this case, if the third element is ϵ-far from being determined by the first element, then a pair of samples detects this non-determination with probability at lesat ϵ/n.

2. A sequence of the form $((i_1, e_1, p_{i_1}(e_1)), ..., (i_m, e_m, p_{i_m}(e_m)))$, where the p_i's are random degree d polynomial over $\mathcal{F}$ and $(i_1, e_1), ..., (i_m, e_m)$ are uniformly and independently selected in $[t] \times \mathcal{F}$.

The reason is that m sample-points of the form (i, e, v) are highly unlikely to include more than d triples with the same first element (i.e., same i), and otherwise the sample-points that correspond to each $i \in [t]$ are uniformly and independently distributed in $\{i\} \times \mathcal{F} \times \mathcal{F}$.

On the other hand, assuming $t = \omega(\sqrt{n})$, a ds-IPP for $\mathcal{D}$ proceeds as follows. The honest prover uses $O(t \cdot d/\epsilon^2)$ sample-points of $X_{\overline{p}}$ in order to reconstruct $\overline{p} = (p_1, ..., p_t)$, but it does not send $\overline{p}$ to the verifier (in this case). The verifier uses $O(\sqrt{n}/\epsilon^2)$ sample-points from the input distribution X in order to check whether the first two elements of X are uniformly distributed in $[t] \times \mathcal{F}$. In addition, for each sample-point (i, e, v), the verifier sends i to the prover, who replies with a degree d polynomial $\widetilde{p}_i$, and the verifier checks that $\widetilde{p}_i$ (is a degree d polynomial that) satisfies $\widetilde{p}_i(e) = v$. (Indeed, the honest prover sets $\widetilde{p}_i \leftarrow p_i$.)

We stress that the prover does not send the entire $\overline{p}$ to the verifier, but rather sends only the parts of $\overline{p}$ that were specified by the verifier. Hence, the communication complexity of this ds-IPP is almost-linear in the sample complexity of the verifier (i.e., it is $O(\epsilon^{-2}\sqrt{n} \cdot d \cdot \log n)$ rather than $O(\epsilon^{-2}t \cdot d \cdot \log n)$).

A ds-IPP for $\mathcal{D}'$ can also be obtained. In this case a sample (of size $O(t \cdot d/\epsilon^2)$) may not determine a unique sequence of t polynomials (of degree d), but the honest prover may just select an arbitrary sequence $\overline{p} = (p_1, ..., p_t)$ that is consistent with this sample. The verifier takes $O(1/\epsilon^2)$ sample-points from the input distribution X, sends the corresponding first elements to the prover, who (again) replies with the corresponding degree d polynomials, and the verifier accepts if and only if the polynomials sent by the prover fit the corresponding other elements. That is, for each sample-point (i, e, v), the verifier sends i to the prover, who replies with p_i (and the verifier checks that $p_i(e) = v$ (and that p_i is a polynomial of degree d)).

Summary. Both $\mathcal{D}$ and $\mathcal{D}'$ require $\Omega(t \cdot d)$ sample-points for testing, but they both have IPPs in which the honest prover uses $O(t \cdot d/\epsilon^2)$ samples. The verifier's sample complexity is $O(\sqrt{n}/\epsilon^2)$ in case of $\mathcal{D}$, and $O(1/\epsilon^2)$ in case of $\mathcal{D}'$. The communication complexity is $O(d \cdot \log n)$ times larger than the sample complexity of the verifier. Recall that each distribution in $\mathcal{D}$ is uniform over a set of size $t \cdot |\mathcal{F}| = n$, and that we may use $|\mathcal{F}| \gg O(d) = O(\log n)$. In particular, for $|F| = \sqrt{n}$, it holds that $t \cdot d = \widetilde{O}(\sqrt{n})$.

Combining both augmentations. Using the suggestion made in the last paragraph of Section 4.1, we can augment the forgoing $\mathcal{D}$ (resp., $\mathcal{D}'$) by replacing $S_{\overline{p}}$ with $S_{\overline{p},G}$ such that

$$S_{\overline{p},G} \stackrel{\rm def}{=} \{(i, e, p_i(e), G(i, e)) : i \in [t] \;\&\; e \in \mathcal{F}\}$$

where $G : [n] \times \mathcal{F} \to \{0, 1\}$ is an arbitrary random process. In this case, the complexity of testing remains $O(t \cdot d/\epsilon^2)$, but the complexity of learning is $\Omega(n)$. Note that $[t] \times \mathcal{F}^2 \times \{0, 1\}$ has size $2n^2/t$. Hence, using $d = O(\log n)$ and a 1-1 correspondence between $[t] \times \mathcal{F}^2 \times \{0, 1\}$ and $[2n^2/t]$, we obtained (via $\mathcal{D}'$) –

Theorem 3 (main result): *For every $t : \mathbb{N} \to \mathbb{N}$ such that $t(n) \leq n/O(\log n)$, there exists a property of distributions over $[2n^2/t(n)]$ that satisfy the following*

1. *The sample complexity of learning this property is* $\Omega(n)$.

2. *The sample complexity of testing this property is* $\Theta(T(n)/\epsilon^2)$, *where* $T(n) = t(n) \cdot \log_2 n$.

3. *The property has a ds-IPP in which the verifier takes* $O(1/\epsilon^2)$ *samples, the prover takes* $O(T(n)/\epsilon^2)$ *samples, and the communication complexity is* $O(\epsilon^{-2} \cdot \log^2 n)$.

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, 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.

[3] 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.

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

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

[6] 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.

[7] Tal Herman and Guy N. Rothblum. Verifying the unseen: interactive proofs for label-invariant distribution properties. In *54th STOC*, pages 1208–1219, 2022.

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

[9] Tal Herman and Guy N. Rothblum. Interactive Proofs for General Distribution Properties. In *65th FOCS*, pages 528–529, 2024.

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

[11] 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.

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il