Electronic Colloquium on Computational Complexity, Report No. 321 (2026)

The Communication Complexity of Verifying Distribution Properties

Tal Herman*
UC Berkeley and RESI

Abstract

Suppose an untrusted powerful data analyst collected many samples from an unknown distribution, and claims that the distribution satisfies some property. For example, that it is close to uniform, or has large support. Can they provide a convincing proof to this effect? If so, what is the complexity of certifying this proof? Is it possible to be convinced of its correctness with fewer resources than required for the original computation?

A recent line of work shows that indeed many fundamental claims about distributions admit *proof systems* with an efficient verifier that draws only a few samples from the distribution. While the question of the optimal sample complexity required for such proofs has received much attention, their communication complexity remained essentially unaddressed, besides in the limited public-coin regime. In this work we give the first communication complexity lower bounds for general proofs for distributions, including secret-coin protocols and computationally sound systems with efficient honest provers, that apply to natural distribution properties.

We show a lower bound on the number of bits required for verifying fundamental quantities of a distribution, including its Shannon entropy, its distance from uniform, and its support size under a minimum-probability promise. For nearly optimal verifier sample complexity, we complement these lower bounds with matching communication upper bounds on the hard families, up to polylogarithmic factors and polynomial dependence on accuracy, using two-message proof systems with nearly optimal verifier and honest-prover sample complexity.

Our lower bound result reveals a connection between interactive proofs for distributions and fundamental problems in communication complexity.

1 Introduction

Suppose we are given i.i.d. samples from an unknown distribution, and are interested to learn some property of the distribution, for example, whether the distribution is close to uniform or whether it has large support. An untrusted data analyst claims to have drawn many samples from the distribution, run some computation over them, and wants to convince us that the distribution indeed admits these features. Can the analyst provide a proof of this claim? Can we verify the proof with fewer resources than required for replicating the analysis ourselves? Recent works study these questions and show that the answer is yes: many basic claims about distributions can be verified with much fewer resources than required for replication.

*This research is supported by The Research Pod on Resilience in Brain, Natural, and Algorithmic Systems at the Simons Institute.

ISSN 1433-8092

Concretely, verification is formalized through the framework of interactive proofs for distribution properties [GMR85, CG18]. In this setting, a probabilistic verifier can draw samples from an unknown distribution and interact with an untrusted prover, whose goal is to convince the verifier that the distribution satisfies some property. If the prover is honest and follows the protocol, the verifier should accept with high probability; but if the claim is false and the distribution is far (in total variation distance) from satisfying the property, then no matter what strategy the cheating prover employs, the verifier should reject with high probability (see Definition 3.4).

Indeed, rich families of distribution properties admit efficient proofs, as shown by recent works [HR22, HR23, Her24, HR24, HR25a, GHR25]. While the question of optimal sample complexity (of the verifier or honest prover) was addressed in several of these works [CG18, HR23, HR25a, HR25b], the question of the optimal communication complexity of this type of proof has so far been largely unexplored.

The only result showing lower bounds for the communication complexity for proofs for distributions is found in the paper that introduced the model by Chiesa and Gur [CG18]. In this work, the authors show a tradeoff between the verifier's sample complexity and the communication complexity of proofs for distributions, granted that the protocol is *public-coin*, that is, that the verifier is limited to only sending random bits. Not only that, they show that their lower bound does not apply to the secret-coin setting, and that the latter allows for more efficient proof systems than possible with public-coin. Indeed, save for a single result, [Her24], all the known upper bound results in this model are *secret-coin*. Thus, the question of the optimal *communication complexity* of general proofs for distributions, be they secret coin or computationally sound, remained open.

In this work, we address this question and show the first lower bounds on the communication complexity of general proofs for distributions, that also apply to secret-coin proofs and computationally sound argument systems with efficient honest provers. We do so by revealing a new connection between interactive proofs for distributions and well established problems in communication complexity. We also show that these lower bounds are tight by providing matching upper bound results.

1.1 Our Results

Our main result establishes new, nearly tight lower bounds on the communication complexity of verifying fundamental properties of distributions, including Shannon entropy, support size[1], and distance from uniformity. We describe the model in more detail and proceed to state the results.

The model. A distribution property $\mathcal{P} = (\mathcal{P}_N)_N$ is a sequence of collections of distributions, where $\mathcal{P}_N$ contains distributions over domain $[N]$, analogous to a language for strings. The measure of distance we consider in this work is *total variation distance* (see Definition 3.1). A *tolerant proof system* for $\mathcal{P}$ (following [GMR85, CG18]) is a protocol between a verifier V and a prover, both with sampling access to an unknown distribution D over $[N]$. Given proximity parameters $\delta_c < \delta_f \in [0, 1)$, the protocol must satisfy:

- **Completeness:** There exists a prover strategy P such that if D is δ_c-close to $\mathcal{P}_N$, at the end of the interaction V accepts with high probability; and

[1]For support-size estimation, we assume every nonzero probability is at least $1/(2N)$ see [HR22, Claims 2.5 and 6.10].

- **Soundness:** If D is δ_f-far from $\mathcal{P}_N$, then no matter what strategy the prover employs (even of unbounded complexity) V rejects with high probability.

The computational resources of interest for such proofs are: the verifier's sample complexity, the communication complexity (total bits exchanged), and the number of rounds, as well as the honest prover's sample complexity and runtime. We now state our results. We say that a proof system verifies some distribution quantity (like the Shannon entropy of the distribution), if given a claim about the value of said quantity, the verifier accepts distributions consistent with that claim, and rejects distributions that are far from it in total variation distance (i.e. far from any such distribution).

Theorem 1.1 (Entropy verification communication lower bound, informal)**.** *Any proof system for verifying (up to constant distance) the Shannon entropy of a distribution, its distance from uniform, or its support size, with optimal verifier sample complexity of $\widetilde{O}(\sqrt{N})$, must incur* communication complexity *of magnitude $\Omega\left(N \cdot \|D\|_2^2\right)$, where D is the input distribution and $\|D\|_2^2 = \sum_{x \in Supp(D)} D^2(x)$.*

This lower bound holds also for argument systems, in which the soundness condition is against computationally bounded adversaries, and under cryptographic assumptions. Even though this is stated for protocols where the verifier's sample complexity is optimal (Chiesa and Gur [CG18] show $\Omega\left(\sqrt{N}\right)$ verifier samples are necessary for verifying these quantities), this same lower bound extends to proof systems where the verifier sample complexity is even larger, up to $o\left(\min\left\{\frac{1}{\|D\|_2^2}, N \cdot \|D\|_2\right\}\right)$, which can be as large as $o(N^{2/3})$ (more on this in Section 2.1 and in the discussion below about the range of quantities related to $\|D\|_2$).

To show that this lower bound is tight, we also present a complementary upper bound that matches the communication complexity lower bound (up to polynomial dependence on the distance parameter and $\log N$) for general label-invariant properties (and thus in particular for the estimation of Shannon entropy, of the distance from uniform, and of the support size of the distribution), while maintaining verifier sample complexity. This protocol is based on the general protocol for label-invariant properties by Herman and Rothblum [HR23].

Theorem 1.2 (Communication-efficient IPs for label-invariant properties - informal)**.** *Every label-invariant distribution property $\mathcal{P}$ admitting the doubly-efficient approximate decision procedure of [HR23, Definition 6.5] admits a tolerant interactive proof system with the following histogram-verification resources:*[2]

- *Efficient verification: the verifier's sample complexity and runtime are $\widetilde{O}\left(\sqrt{N} \cdot (\delta_f - \delta_c)^{-4}\right)$.*
- ***Low communication:** The protocol consists of 2 messages, and the communication complexity is $\widetilde{O}\left(N \cdot \|D\|_2^2 \cdot (\delta_f - \delta_c)^{-6}\right)$.*
- *Doubly-efficient prover: The honest prover's sample complexity is $\widetilde{O}\left(N(\delta_f - \delta_c)^{-4}\right)$.*

Optimal protocols for label-invariant properties. This protocol is optimal, up to dependence on the distance parameter and polylog$N$ factors in all quantities of interest, and for all the properties mentioned above. Concretely, previous works established that the verifier's sample complexity and

[2] See Section 5 for accuracy requirements and property-dependent costs.

runtime, as well as the honest prover's sample complexity are optimal [CG18, HR23], and this work shows that the communication complexity can also be made optimal. Moreover, this protocol is considerably more efficient than the standalone tester for many label-invariant properties, that required sample complexity of $\widetilde{\Omega}(N)$. Restricting ourselves to non-interactive proof systems, i.e. MA proofs, by Chiesa and Gur's tradeoff [CG18, Theorem 2] we know that any proof system with $\widetilde{O}(\sqrt{N})$ verifier samples requires $\widetilde{\Omega}(\sqrt{N})$ communication. Thus, our protocol can improve on this communication bound with only 2 messages, meaning that this proof system is also optimal with respect to the round complexity (if we fix optimal verifier sample complexity and optimal communication complexity).

On the quantity $\left(N \cdot \|D\|_2^2\right)$. $\|D\|_2^2 = \sum_{x \in \mathrm{Supp}(D)} D^2(x)$, also called the *collision probability* of the distribution, captures (among other things) the probability that two samples drawn according to D collide and land on the same domain element. Thus, if we draw a sample of size $\Theta\left(\sqrt{N}\right)$, since every two samples collide with probability $\|D\|_2^2$, and there are $\Theta\left(N\right)$ possible pairs in the sample, the quantity $\Theta\left(N \cdot \|D\|_2^2\right)$ is the magnitude of the expected number of *colliding pairs* in the sample. For $\sigma = \delta_f - \delta_c$, first consider the bounded-probability case $D(x) \leq \frac{\sigma^2}{\sqrt{N}}$ for every $x \in [N]$.[3] Then $\left(N \cdot \|D\|_2^2 \cdot \sigma^{-6}\right) = O\left(\sqrt{N} \cdot \sigma^{-4}\right)$, but it can also be as low as $\mathrm{poly}\left(\log N, \sigma^{-1}\right)$, if, for example, D is guaranteed to satisfy $D(x) \leq \frac{\mathrm{polylog}(N)}{N}$ for every $x \in [N]$. This communication lower bound thus suggests that the verifier needs to communicate information proportional to the number of colliding pairs they exhibit in the sample, and no less. Indeed, in the following sections we explore the connection between the set of colliding pairs in the verifier's sample and the communication complexity of the protocols.

Comparison to [HR23]. To illustrate the scale of improvement in communication complexity compared to the previous construction of Herman and Rothblum [HR23], take for example the construction behind the lower bound of Valiant and Valiant [VV10]. In this work, Valiant and Valiant construct two families of distributions over domain $[N]$, both are closed under permutation (i.e. label-invariant families), but one contains distributions that are close to uniform over the entire domain, while the other contains distributions close to being uniform over some half the domain. Still, they show that given samples from a distribution promised to have come from one of these families, telling to which one it belongs requires $\Omega(N/\log N)$ samples. This result implies that estimating the Shannon entropy of a distribution, its distance from uniform or support size just from samples, must require $\Omega(N/\log N)$ samples. Both Theorem 1.2 and the result in [HR23] allow a verifier with only $\widetilde{O}\left(\sqrt{N}\right)$ samples to distinguish between the two families, however, in [HR23] this will require $\widetilde{\Theta}\left(\sqrt{N}\right)$ bits of communication; while the communication complexity for verifying this same task via Theorem 1.2 will amount to only $\mathrm{polylog}(N)$ bits! This is because any distribution from either family in the Valiant and Valiant construction satisfies that for every $x \in [N]$, $D(x) \leq \frac{\mathrm{polylog}(N)}{N}$, which implies that $\|D\|_2^2 \leq \frac{\mathrm{polylog}(N)}{N}$.

[3] Remark A.15 accounts for learning the heavier elements and verifying the remaining part, preserving both resource bounds and the stated accuracy exponents.

Public-coin vs. secret-coin protocol for distributions. Chiesa and Gur [CG18] showed that any public-coin proof system (i.e. a proof system where the verifier is only allowed to send random bits to the prover) with verifier sample complexity s and communication complexity c, for a distribution property with optimal *tester* sample complexity *opt* must satisfy $s \cdot c = \Omega(opt)$. They also showed that there are properties for which secret coin proof systems achieve parameters that break this bound, and are strictly stronger. The protocol in [HR23] provided a secret-coin protocol for general distribution properties, yet its parameters do not exceed the lower bound for public coin proofs. In this work, we show that private coin proof systems are strictly stronger than public coin proof systems for the rich family of label-invariant properties, including, as stated above, even the family of hard distributions introduced by Valiant and Valiant [VV10].

From IPs to communication complexity. In interactive proofs for distribution properties, a verifier draws samples from an unknown distribution and interacts with an untrusted prover. In standard two-party communication complexity, two honest parties each hold an explicit input string, with query access to the string, and exchange messages to compute a function of their inputs. Earlier work connects distribution testing to communication complexity to prove *sample* lower bounds [BCG17]. We show how a proof system yields an honest-party communication protocol with comparable communication complexity. This transfers established communication lower bounds on strings to the communication required for verifying distribution properties, resolving the previously open question for general proof systems for these properties, including protocols in which the verifier uses private randomness.

Beyond label-invariant properties. On top of the implication to label-invariant properties, our lower bound construction also implies communication lower bounds for some *label-sensitive* tasks. Consider the task of *tolerant identity verification*: given samples from distribution D over domain $[N]$, distance parameters $\delta_c < \delta_f \in (0,1)$, as well as an explicit description of a distribution Q over domain $[N]$ - i.e. a string of size $(N \cdot \log N)$ specifying $Q(x)$ for every $x \in [N]$, the verifier needs to accept w.h.p. if $\delta_{\mathtt{TV}}(D,Q) \leq \delta_c$; and reject w.h.p. if $\delta_{\mathtt{TV}}(D,Q) \geq \delta_f$, no matter what cheating strategy the prover might employ. We show:

Theorem 1.3 (Communication lower bound for tolerant identity verification)**.** *Consider tolerant identity verification with completeness threshold $\delta_c = 1/16$ and soundness threshold $\delta_f = 5/64$: the verifier is given an arbitrary explicit reference distribution Q over $[N]$ and sample access to an unknown distribution D over $[N]$. Any proof system for this problem using*

$$s(N) = o\left(\min\left\{\frac{1}{\|D\|_2^2},\, N\|D\|_2\right\}\right)$$

verifier samples must communicate $\Omega\big(N\|D\|_2^2\big)$ bits. In particular, this communication complexity lower bound holds for every protocol using $O(\sqrt{N})$ verifier samples (i.e. protocols with optimal verifier sample complexity).

Therefore, as our general upper bound results also allow tolerant identity verification with the same parameters (see Proposition 5.1) as in Theorem 1.2, this work also implies an optimal interactive proof for this task as well.

Another label-sensitive property for which we get a communication lower bound is *monotonicity verification*. Let MON_n be the distributions D on $\{0,1\}^n$ satisfying $D(x) \geq D(y)$ whenever $x \leq y$

coordinatewise. Standalone monotonicity testing requires $2^{\Omega(n)}$ samples even for a sufficiently small constant proximity parameter [AGP+19, Corollary 13]. For $N = 2^n$, we obtain a worst-case communication lower bound of $\Omega(N\|D\|_2^2)$ for tolerant monotonicity verification with $\widetilde{O}(\sqrt{N})$ verifier samples. See Section 4.4 for more detail.

More related works. Our work builds on a the study of the power of sublinear time verifiers, who cannot read the entire input [EKR04, RVW13, GR18], and on results in property testing [GGR98], and in particular, distribution testing - a line of reasearch that investigates algorithms that upon granted samples from an unknown distribution determine (without help of a prover) whether the distribution admits a property or not [GR00, BFR+00]. In particular, we consider in our work *tolerant* verification that is based on tolerant property testing as introduced by [PRR06]. The reader is referred to the distribution testing survey by Canonne [Can22]. We also note that the setting of verifying some properties of distributions (in particular, closeness verification and entropy verification) was also researched in earlier works, in the white-box setting, where the distribution is given as a explicit sampling circuit, rather than a black-box sampling device [Vad99, GV99] - a related model, that offers considerably stronger tools for verification.

On top of the works mentioned above, we highlight a few recent works that deal with adjacent questions. Most closely to this work, Herman and Rothblum [HR25b] showed a lower bound on the *sample complexity* of the *honest prover* in proofs for label invariant properties, demonstrating that even for *easy* label invariant properties (i.e. those where the optimal tester requires $o(N)$ samples), the honest prover strategy still requires many samples, thus providing a deeper perspective of the parameter regime for proofs for label-invariant properties. Biswas et al. [BBCS26] also consider proofs for label invariant distribution properties and construct proof systems that achieve better parameters than presented here, breaking the lower bound, by providing the verifier with added power: for any two elements in the domain of their choosing, the verifier can also know which is more probable, and by how much. Lastly, Gopalan et al. [GMR+25] investigated the communication complexity required for Alice and Bob, holding distributions X and Y respectively, to compute $\mathbb{E}_{x\sim X, y\sim Y}[f(x,y)]$. Despite some similarities, the connection between their models and ours is unclear.

AI methodology. This paper was fully written before any AI was used and the ideas and proofs were originally all conceived by the author. The author used AI tools to make their proofs and explanations more concise, while correcting mistakes identified by AI. The statements of claims and lemmas, as well as the strategy of the proofs, are mostly written by the author, while the proofs themselves heavily employed AI tools. AI also helped significantly in writing the preliminaries section, while Sections 1 and 2 were written by the author entirely (save for very minor adjustments). The model used for writing was ChatGPT 6.1 Sol. Claude Opus 5.5 was used to review the final paper for mistakes and clarity. The author take full responsibility for any mistakes that remain in the paper, and for any ideas that were not properly credited to their original sources.

Acknowledgements. I would like to thank Satchit Sivakumar for his help improving this manuscript, Nathaniel Harms for explaining communication complexity ideas to me, and Guy Rothblum for general support and the improvement in the dependence on the distance parameter of the upper bound.

2 Technical Overview

This section is divided into two parts. Section 2.1 contains an overview of the main result: a communication complexity lower bound for proof systems for label invariant properties; and in Section 2.2 we review the protocol behind Theorem 1.2, that allows low-communication verification of general label-invariant properties.

2.1 Communication Complexity Lower Bounds for Label-Invariant Properties

Our lower bound for the properties mentioned in the previous section is obtained by showing that there exists a label-invariant property for which any proof system must require many bits of communication. Since this property can be reduced to estimation of Shannon entropy, distance from uniform or support size, a lower bound for it implies a lower bound for those three properties (and more).

The formal argument behind the lower bound is obtained via a new connection to a fundamental problem in communication complexity, namely the `Gap-Disjointness-Problem` (`GapDisj`) in which Alice and Bob each hold a set of elements, and wish to determine whether the intersection between their input sets is empty or contains at least g-many elements for some choice of gap parameter g (see Definition 3.13).

We first introduce the new promise problem, which we call the `Noisy Collisions Problem` (`NCP`), give intuition as to why it requires many bits of communication, and then review the lower bound proof which is a reduction from `GapDisj` to `NCP`.

2.1.1 The Noisy Collisions Problem

Consider the following two distributions $D_0(N)$ and $D_1(N)$ defined over domain $[N]$:

- $D_0(N)$ is a distribution with half its mass on elements with probability $\frac{1}{2N^{0.6}}$ (i.e. it has $N^{0.6}$ elements with said probability), and half its mass split uniformly on the rest of the domain, i.e. $N - N^{0.6}$ elements, each with probability $\frac{1}{2(N-N^{0.6})} \approx \frac{1}{2N}$.

- $D_1(N)$ is a distribution that also has half its mass on $N^{0.6}$ heavy elements with probability $\frac{1}{2N^{0.6}}$, but the other half is split uniformly over $N/3 - N^{0.6}$ light elements, each with probability $\frac{1}{2(N/3-N^{0.6})}$. The remaining $2N/3$ elements have probability zero.

For both distributions we refer to the elements with probability $\frac{1}{2N^{0.6}}$ as the *heavy* elements, and the elements with probability $\Theta\left(\frac{1}{N}\right)$, as the *light* elements. Given samples from a distribution D, that is guaranteed to be equal to one of these distributions - we want to determine whether $D = D_0$ or $D = D_1$. If we know the set of *light* elements for D_1 this is easy - draw $O(1)$ samples from the distribution, and if we don't see a single *light* sample that landed outside the support of D_1, we know with high confidence that $D = D_1$. This problem becomes interesting when all we know about D_0 and D_1 is the information provided above, i.e., only the label-invariant information about these distributions.

In this case, all a sample-based algorithm can do is count how many elements appear once, twice, three times, and so on in the sample, and decide using this information. In this setting, given $o\left(\sqrt{N}\right)$ samples, it's impossible to distinguish between these families. With $\Omega\left(\sqrt{N}\right)$ samples, *light*

elements begin to collide and appear twice, and the distribution D_0 will exhibit considerably fewer *light collisions* than D_1.

However, this is still not good enough, as the variance in the number of *heavy collisions* (the number of heavy elements that appear twice) will mask the expected difference in total collisions. If we could know which of the elements that appear twice is light and which is heavy, $O(\sqrt{N})$ samples would have indeed been sufficient; without this information, $\widetilde{O}(N^{0.6})$ samples suffice by first identifying the heavy elements and then testing the light part (Claim 4.3).

Now, fix some distributions $D_0(N)$ and $D_1(N)$ over domain $[N]$ as above, and consider the families of distributions $\mathcal{F}_0(N)$ and $\mathcal{F}_1(N)$ defined by taking the closure under permutations of $D_0(N)$ and $D_1(N)$ respectively (see Definition 3.5). We define:

$$\texttt{NCP}(N) = (\mathcal{F}_0(N), \mathcal{F}_1(N))$$

We think of $\mathcal{F}_0$ as the YES case, and $\mathcal{F}_1$ as the NO case. Observe that these families are: (i) label invariant; (ii) $\Omega(1)$-far in total variation distance; (iii) can be distinguished by a standalone algorithm using $\widetilde{O}(N^{0.6})$ samples; (iv) differ in many label-invariant features such as: support size, distance from $U_{[N]}$, entropy, and more. The latter implies that any protocol between a verifier and untrusted prover that can estimate these quantities from samples of an unknown distribution D, can successfully determine to which family D belongs.

A note on the *noise parameter*. Note that the distinguishing feature between $\mathcal{F}_0$ and $\mathcal{F}_1$ is the number of *light* elements that appear twice, and the *heavy* elements mask this signal. We thus refer to the heavy elements as *noise*. In the example above, we defined the heavy elements to be of probability $\frac{1}{2N^{0.6}}$, however, we could have picked even heavier elements, with for example probability $\frac{1}{2N^{0.55}}$ each. In the small-sample regime, this creates more heavy collisions and makes the light-collision signal harder to detect. In Section 4 we parameterize the problem also according to this *noise* parameter, and take it to be $\frac{1}{N^k}$ for $k \in (0.5, 1)$. Note that this parameter controls the collision probability of a distribution taken from either family. That is, for every $D \in \mathcal{F}_0 \cup \mathcal{F}_1$,

$$\|D\|_2^2 = \Theta\left(\frac{1}{N^k}\right).$$

2.1.2 Communication Lower Bound for NCP Verification

A proof system for NCP (as defined in Section 1.1) satisfies the following completeness and soundness conditions:

- **Completeness.** There exists a prover strategy P, such that if $D \in \mathcal{F}_0$, the verifier accepts with high probability.
- **Soundness.** If $D \in \mathcal{F}_1$, no matter what cheating strategy a prover might employ, the verifier rejects with high probability.

One protocol for this problem that uses $\widetilde{O}(\sqrt{N})$ communication and verifier samples is the protocol for general label-invariant properties of Herman and Rothblum [HR23]. In this protocol, the verifier sends all its samples (alongside some additional randomness) to the prover, who responds with the probability of each element, effectively allowing the verifier to know which elements are light and which are heavy. However, is this optimal? How many bits are necessary to solve this problem?

To show the lower bound, we first need to understand what information the verifier has access to. We introduce useful notation for capturing all the information in the verifier's sample. Let s be the sample complexity of the verifier, and assume it satisfies $s = \Omega(\sqrt{N})$ and $s = o(N^{0.6})$[4]. For each $i \in [s]$, let $V_i \subseteq [N]$ be the set of elements that appear exactly i times in the sample. Thus, the random variable $(V_i)_{i\in[s]}$ represents any possible verifier sample. Note that since the problem is label-invariant and each case is composed only of permutations of a single distribution, $(|V_i|)_i$ is distributed identically among all distributions inside each case.

Moreover, with high probability, V_i for $i \geq 3$ contains only heavy elements, and thus is distributed identically in both cases. $|V_1|$ in both cases (no matter if D is from $\mathcal{F}_0$ or $\mathcal{F}_1$) is composed, in expectation, in nearly half of its elements light and the other half heavy (with a very slight difference between the two cases).

The informative set that differs significantly between the two cases is V_2, the set of two-way collisions. Here lies the signal: in the YES case we expect about one third as many *light elements* as in the NO case (while the distribution of heavy elements that appear twice is the same between the cases). We choose sample size parameter $s = \widetilde{\Omega}(\sqrt{N})$ such that with high probability V_2 contains in the YES case about $\log N$ light elements, while in the NO case it contains about $3 \log N$. At the same time, regardless of case, with high probability

$$|V_2| = \Theta\left(\binom{s}{2} \cdot \frac{1}{N^{0.6}}\right) = \widetilde{\Theta}\left(N^{0.4}\right) = \widetilde{\Theta}\left(N \cdot \|D\|_2^2\right),$$

and this set is dominated by heavy elements. To solve NCP, the verifier must learn how many elements in V_2 are light. It is like finding a needle in a haystack where the needle looks like hay. This is the communication bottleneck: The verifier knows V_2, while the prover knows the set $S_{\text{light}} \subseteq [N]$ of light elements in the support. Solving NCP amounts to estimating

$$|V_2 \cap S_{\text{light}}|$$

up to a gap of order $\log N$. This is precisely the kind of task captured by set disjointness.

We want to show that any such proof system with verifier sample complexity s requires at least $\Omega\left(N \cdot \|D\|_2^2\right) = \Omega\left(N^{0.4}\right)$ bits of communication.

2.1.3 Reducing NCP to GapDisj

We show that if a protocol for NCP exists it is possible to construct a protocol for the communication problem of Gap Set Disjointness that breaks known lower bounds for this problem. The main insight is the observation made above: only a small part of the verifier's sample is informative - can we make the intersection between Alice's and Bob's sets correspond to this informative part? If so, then we can use a protocol for NCP to solve GapDisj.

We provide here the minimal tools from communication complexity we need for the sake of our proof, see Section 3.2 for more detail and formal statements.

Definition 2.1 (Gap Set Disjointness)**.** *Let $N, g \in \mathbb{N}$ with $g \leq N$. In the* Gap Disjointness *problem, denoted $\mathtt{GapDisj}_{N,g}$, Alice holds a set $A \subseteq [N]$ and Bob holds a set $B \subseteq [N]$, with the following promise:*

[4]For general $K = N^k$, the proof requires $s = o(\min\{K, N/\sqrt{K}\})$. This includes $s = \widetilde{O}(\sqrt{N})$ for every fixed $k \in (1/2, 1)$; see Section 4.

- ***YES case:*** $A \cap B = \emptyset$*; or*
- ***NO case:*** $|A \cap B| \geq g$*.*

The goal is for Alice and Bob to distinguish between the two cases: they must output 1 *(accept) with high probability in the YES case, and output* 0 *(reject) with high probability in the NO case.*

Theorem 2.2 (Gap disjointness lower bound - informal, [She14, RY20])**.** *In the parameter regime used below,*[5] *the communication complexity required for* `GapDisj`*, even when Alice and Bob share unlimited randomness is* $\Omega\left(\frac{\min\{|A|,|B|\}}{g}\right)$*.*

The reduction. We start with an instance of `GapDisj` with carefully chosen parameters: Alice has a sparse set $A \subseteq [N]$ such that $|A| = \Theta(N^{0.4} \log N)$, and Bob has a dense set $B \in \{0,1\}^N$ such that $|B| = N - N^{0.6}$. They share randomness, and it is guaranteed that one of the following conditions holds:

$$|A \cap B| = 0 \qquad \text{or} \qquad |A \cap B| = 2 \log N$$

We show a setup procedure at the end of which:

- **If** $|\mathbf{A} \cap \mathbf{B}| = \mathbf{0}$**:** Alice will have obtained $(V_i)_{i \in [s]}$ such that for all i, $V_i \in \{0,1\}^{2N}$, is distributed like the verifier's sample drawn from some $D \in \mathcal{F}_0(2N)$; Bob will have obtained $S_{\text{heavy}}, S_{\text{light}} \in \{0,1\}^{2N}$ indicators of heavy and light support elements for the same distribution D.
- **If** $|\mathbf{A} \cap \mathbf{B}| = \mathbf{2} \log \mathbf{N}$**:** Alice will have obtained $(V_i)_{i \in [s]}$ such that for all i, $V_i \in \{0,1\}^{2N}$, is distributed similarly to the verifier's sample drawn from a *randomly chosen* $D \in \mathcal{F}_1(2N)$, while Bob will have obtained $S_{\text{heavy}}, S_{\text{light}} \in \{0,1\}^{2N}$ consistent with a distribution $D' \in \mathcal{F}_0$.

This suffices: given such a process, Alice and Bob can simulate a run of any protocol for $\texttt{NCP}(2N)$, using as many bits of communication as required by this protocol, where Alice simulates the verifier and Bob simulates the honest prover strategy with the distribution he has. If $|A \cap B| = 0$, by the completeness of the protocol, the simulated verifier should accept. Otherwise, Alice obtained a sample from a distribution in $\mathcal{F}_1(2N)$, yet the simulated prover simply enacts one possible cheating prover strategy (answering according to a fixed distribution in $\mathcal{F}_0(2N)$), and so by the soundness of the protocol, the simulated verifier rejects with high probability. In either case, Alice's decision at the end of the protocol will solve `GapDisj`, with the communication required for $\texttt{NCP}(2N)$. Note that the setup procedure does not require any communication between the sides. We proceed to outline the setup procedure, concluding this section.

The setup procedure. **Alice** needs to construct the verifier's entire sample without knowing B, while **Bob** needs to construct the prover's distribution without knowing A. We divide $[2N]$ into the input labels $[N]$ and the additional labels $\{N+1, \ldots, 2N\}$. Among the additional labels, we fix heavy and light, *that are known to both parties*[6]. **Bob** considers the elements of B light and those

[5] The size conditions are stated in Theorem 3.14; our dense-input parameters satisfy them, as recorded in Corollary 3.15.

[6] The size of the heavy set on a domain of size $2N$ should be $(2N)^{0.6}$, and so, we can consider the heavy set in the new part of the domain to contain $(2N)^{0.6} - N^{0.6}$ many elements, while the rest of $\{N+1, \ldots, 2N\}$ is *light*. In the technical section we adopt a slightly different approach, parameterizing `NCP` also by the size of the heavy set

of $[N] \setminus B$ heavy, and completes these sets using the additional labels (making the elements in the intersection of A and B by definition, *light*). Putting the light and heavy parts of both sides of the new domain together, **Bob** obtains $S_{\text{light}}, S_{\text{heavy}} \subseteq [2N]$ of the sizes required for a distribution in $\mathcal{F}_0(2N)$. Bob can construct this distribution using only B, and it is a YES distribution regardless of A.

To compose her sample, **Alice** first draws a sample from a known canonical YES distribution and records, for every i, how many heavy and light elements appear exactly i times. She keeps these counts and discards the sampled labels. For each $i \neq 2$, she constructs V_i using the prescribed numbers, and labels from $\{N+1, \ldots, 2N\}$ - recall that by construction, in this part of the domain, **Alice** knows which elements are light and which are heavy, and so she can populate the sets with heavy and light elements as prescribed by the sample she drew.

We now consider V_2. We consider A to be related to the number of 2-way collisions in the sample. **Alice** takes the number of heavy elements that appear twice in the canonical sample she drew, and subtracts $|A|$ from it. Then, using this altered heavy-count, **Alice** populates V_2 with heavy and light elements from the new label $\{N+1, \ldots, 2N\}$, and to it, she adds A. This ensures the following:

- If $A \cap B = \emptyset$, all elements of A are heavy according to **Bob's** distribution, and so, adding A restores the heavy count to the canonical YES count.
- If $|A \cap B| = 2 \log N$, exactly $2 \log N$ of the elements of A are light according to **Bob's** distribution, and so, adding A increases the light count by $2 \log N$ and decreases the heavy count by $2 \log N$. Since the heavy count is of magnitude $\widetilde{\Theta}(N^{0.4})$, it is not disturbed by this change, but the light count now has about three times as many light elements as expected in a YES sample. The technical proof bounds the distance of this sample from the NO distribution.

Finally, **Alice** and **Bob** use their shared randomness to choose a uniform permutation π of $[2N]$, which ensures that the labels are distributed uniformly, and thus completeness and soundness is inherited from the `NCP` protocol. Alice applies it to every V_i, and Bob applies it to S_{light} and S_{heavy}. This preserves which sampled elements are heavy or light and makes their labels uniformly distributed within each type (a similar approach was made in [HR25b]). Section 4 bounds the distance of the resulting setup from the desired YES and NO distributions, including the small probability that the construction aborts. Neither party needs to learn the intersection, and the setup exchanges no bits.

2.2 The Upper Bound: A Short Overview

We provide a protocol for general label-invariant distribution properties, as stated in Theorem 1.2, showing that our communication lower bound is optimal up to polynomial dependence on the proximity gap and polylogarithmic factors in N. Our new protocol is a variant of the protocol of Herman and Rothblum [HR23].

2.2.1 The [HR23] Protocol

In their protocol, the verifier sets

$$s = \sqrt{N} \cdot \text{poly}(\sigma^{-1}, \log N).$$

It draws $2s$ independent elements $(z_i)_{i\in[2s]}$, choosing each from D or $U_{[N]}$ with equal probability. Its first message contains all these elements, while the bits indicating their origins remain secret. The prover responds with an alleged probability π_{z_i} for each element. The verifier groups similar alleged probabilities into geometric buckets[7] and performs two tests over each bucket.

Fix a bucket with alleged probability p, and let α_p be the fraction of D-samples assigned to it. The *consistency test* compares the bucket's alleged empirical size with its alleged empirical mass. If the prover is honest, and its empirical *mass* according to D is α_p, then we can deduce that the set of all domain elements with probability p also has mass approximately α_p. This in turn implies that the domain contains roughly α_p/p such elements. A uniform sample therefore should land in it with probability approximately $\alpha_p/(Np)$. Thus, the verifier checks that the ratio of D-sample occurrences to uniform occurrences in the bucket is approximately Np, this is the first test.

The *collisions test* draws s fresh samples from D and counts their matches with the first sample's D-occurrences in the bucket. For an honest prover, each of the approximately $s\alpha_p$ occurrences in the bucket has probability approximately p, so the expected number of matching pairs is approximately $s^2\alpha_p p$, and so the verifier checks for that too.

Herman and Rothblum show that substantially incorrect alleged probabilities cause at least one of the two tests to reject with high probability for at least one of the buckets; see [HR23] for the analysis.

2.2.2 Reducing the Communication Complexity

We show how to reduce the communication complexity of the protocol. Our key observation is that both tests can be supported with communication proportional to $N\|D\|_2^2$, up to logarithmic and accuracy factors.

Remark 2.3 (Max probability assumption.)**.** *We first assume that the maximum probability of D is at most $\sigma^2/\sqrt{N}$. Any element with higher probability can be explicitly learned using $\widetilde{O}(\sqrt{N}\sigma^{-4})$ samples, up to accuracy σ. This also implies that $\|D\|_2^2$ is at most $\sigma^2/\sqrt{N}$. This assumption is also used in [HR23].*

The verifier first draws a sample from D of expected size $s = \sqrt{N}\cdot\mathrm{poly}(\sigma^{-1}, \log N)$ and estimates $\|D\|_2^2$ from its number of equal pairs. From a fresh, independent sample of the same expected size, it retains the elements appearing at least twice as a set C. Since each element of C contributes at least one equal pair,

$$\mathbb{E}|C| \le \frac{s^2}{2}\|D\|_2^2 = N\|D\|_2^2 \cdot \mathrm{poly}(\sigma^{-1}, \log N).$$

Using the preliminary number of equal pairs, the verifier sets m to that number times a suitable $\mathrm{poly}(\log N)$ factor, giving

$$m = \widetilde{\Theta}\big(s^2\|D\|_2^2\big) = \widetilde{\Theta}\big(N\|D\|_2^2\sigma^{-6}\big)\,.$$

As in the previous protocol, it then draws a sample of size $2m$, choosing each element from D or $U_{[N]}$ with equal probability, so that each origin contributes m samples in expectation. The verifier randomly mixes these elements with C and sends their union to the prover, each label once, keeping origins and multiplicities secret. Observe that the number of elements sent is at most $|C| + 2m$, making C a small fraction of the total. The prover returns one alleged probability per label. The

[7] Each bucket groups alleged probabilities within a multiplicative factor $1 + O(\sigma^2)$, using representatives $\sigma(1 + O(\sigma^2))^j/N$. There are $O(\sigma^{-2}\log N)$ buckets; rounding and normalization incur $O(\sigma^2)$ TV error.

verifier uses the D- and uniform samples for the consistency test and equal pairs in the sample defining C for the collisions test.

As in [HR23], the prover sees only the labels, without being told whether they came from D, $U_{[N]}$, or C, and must give one alleged probability for all occurrences of each label. Fix a bucket with alleged probability p, and denote by α_p the fraction of D-samples assigned to it. The consistency test checks that the ratio of D- to uniform occurrences is approximately Np, while the collisions test checks for approximately $(s^2/2)\alpha_p p$ equal pairs whose common label lies in C and is assigned to this bucket.

For completeness, the verifier still has the large sample of expected size s for the collisions test: reducing communication does not discard any of its equal pairs. For the consistency test, an honest bucket of mass approximately α_p contains approximately α_p/p domain elements. Its expected numbers of D and uniform occurrences are therefore $m\alpha_p$ and $m\alpha_p/(Np)$. These elements contribute approximately $\alpha_p p$ to $\|D\|_2^2$, so

$$m\frac{\alpha_p}{Np} = \frac{\alpha_p\|D\|_2^2}{p}\cdot \mathrm{poly}(\sigma^{-1}, \log N) \gtrsim \alpha_p^2 \cdot \mathrm{poly}(\sigma^{-1}, \log N).$$

Thus, a bucket of sufficient mass receives enough uniform samples even if it contains few elements; concentrating mass on those elements increases $\|D\|_2^2$ and hence m. The D-count is also large, since $N\|D\|_2^2 \geq 1$. The tests are applied only when their counts are sufficiently large; the effect of the remaining buckets is bounded in Appendix A.

For soundness, C is masked by combining its labels with the D and uniform samples, sending each label once and hiding all origins and multiplicities. Labels with the same true probability have identical conditional distributions of their hidden source counts, so the prover cannot select among them according to their actual origins. More generally, conditioning on the message changes each label's expected D-count, uniform count, and equal-pair count by the same factor. This preserves the comparison underlying the soundness argument of [HR23], despite the addition of C. We refer the reader to Appendix A for the full conditional analysis and concentration bounds.

Remark 2.4 (Dependence on the accuracy parameter)**.** *Separately, we improve the dependence on σ by dividing buckets into tiers and testing them at different accuracy levels. This refinement yields verifier sample complexity $\widetilde{O}(\sqrt{N}\sigma^{-4})$ and is distinct from the reduction in communication using $\|D\|_2^2$; its full analysis also appears in Appendix A. The optimal accuracy dependence for this information-theoretic verification task remains open: the known sample lower bound is $\Omega(\sqrt{N}\sigma^{-2})$ [CG18, HR22]. Closing the gap between the exponents 2 and 4 is left for future work.*

Our goal is to minimize communication and verifier sample complexity while obtaining a tagged sample. If instead we want to tag a larger sample with accuracy σ, we can retain the tier analysis and follow the large-sample approach of [HR23], with only a σ^{-4} dependence in the required sample size; see Remark 5.2. This larger-sample variant could also improve the bounds for general distribution properties in [HR24]. Taking $c = 4$ in their tradeoff in Section 2.3 would reduce the verifier's sample complexity from $\widetilde{O}(N^{18/19})$ to $\widetilde{O}(N^{9/10})$, and the communication complexity from $\widetilde{O}(N^{37/38})$ to $\widetilde{O}(N^{19/20})$, for constant proximity gap. For circuits of depth d and size S, communication additionally includes $d\mathrm{polylog}S$. The adaptation to previously used samples is not established here.

3 Preliminaries

3.1 Distribution Testing and Verification

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:*

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

Definition 3.2 (Distribution property)**.** *We say that $\mathcal{P} = (\mathcal{P}_N)_{N \in \mathbb{N}}$ is a distribution property if $\mathcal{P}_N \subseteq \Delta_N$, where Δ_N is the set of all distributions over domain $[N]$.*

Definition 3.3 (Distribution tester for property $\mathcal{P}$)**.** *Let δ be some distance measure between distributions, $\mathcal{P}$ 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 the domain $[N]$, 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 $\delta_{TV}(D, \mathcal{P}_N) > \varepsilon$, 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.

Definition 3.4 (Tolerant proof system for distribution property, [GMR85, CG18])**.** *A proof system for a distribution property $\mathcal{P} = (\mathcal{P}_N)_{N \in \mathbb{N}}$ with parameter $\varepsilon \in (0, 1)$ is a two-party game, between a verifier executing a probabilistic polynomial time strategy V, and a prover. Both sides are given black-box sample access to distribution D over the domain $[N]$, domain size parameter N, and proximity parameters $\delta_c < \delta_f \in [0, 1)$. At the end of the interaction, the following conditions hold:*

- ***Completeness:*** *There exists a prover strategy P, such that if $\delta_{TV}(D, \mathcal{P}_N) \leq \delta_c$, the verifier V, after interacting with the prover P, accepts with probability at least $2/3$.*
- ***Soundness:*** *If $\delta_{TV}(D, \mathcal{P}_N) \geq \delta_f$, for every cheating prover strategy P^* (including unbounded ones), 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 well as the honest prover (strategy P), the communication complexity, the runtime of both agents, and the round complexity (how many messages were exchanged).

Definition 3.5 (Permutation of distribution)**.** *For two distributions D and Q over domain $\mathcal{X}$, and permutation π over $\mathcal{X}$, we say $Q = \pi(D)$ if for every $x \in \mathcal{X}$,*

$$Q(x) = D(\pi^{-1}(x))$$

Definition 3.6 (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 $\sigma(D) \in \mathcal{P}_N$.*

3.1.1 Poissonization and Other Statistical Claims

We use Poissonization to replace a fixed sample size by $s' \sim \mathrm{Poi}(s)$. The following claims describe the distributions of the resulting counts and justify this replacement for interactive proofs, including adaptive sample requests. See also [Val11, Can22].

Claim 3.7 (Poissonized counts)**.** *Fix a distribution D over $[N]$. Draw $s' \sim Poi(s)$ independently, and then draw s' independent samples from D. The occurrence counts X_x, $x \in [N]$, are independent, with $X_x \sim Poi(sD(x))$.*

If n labels each have probability p, and h_i counts those occurring exactly i times, then $(h_i)_{i\geq 0}$ has the multinomial distribution with n trials and probabilities $q_i := e^{-sp}(sp)^i/i!$. In particular, $h_i \sim Bin(n, q_i)$, and, for $i \geq 1$,

$$h_i \mid (h_j)_{j\geq 1,\ j\neq i} \sim Bin\left(n - \sum_{j\geq 1,\ j\neq i} h_j, \frac{q_i}{q_0 + q_i}\right).$$

Count vectors for disjoint sets of labels are independent. Moreover, for every positive integer r,

$$\Pr[s' < s/2] \leq e^{-s/8}, \qquad \Pr[s' > 2s] \leq e^{-s/3}, \qquad \mathbb{E}\binom{X_x}{r} = \frac{(sD(x))^r}{r!} \quad (r \geq 1).$$

Proof. For nonnegative integers k_x with $m := \sum_x k_x$, the joint probability of these occurrence counts is

$$e^{-s}\frac{s^m}{m!}\frac{m!}{\prod_x k_x!}\prod_x D(x)^{k_x} = \prod_x e^{-sD(x)}\frac{(sD(x))^{k_x}}{k_x!}.$$

This proves independence and the stated Poisson distributions. Categorizing the independent label counts by their multiplicity gives the multinomial distribution; conditioning leaves only categories 0 and i, giving the conditional binomial distribution. The factorial-moment identity follows by shifting the index in the Poisson sum. Finally, the moment-generating function $\mathbb{E}[e^{ts'}] = \exp(s(e^t - 1))$ and Chernoff's method give the two tail bounds. □

Claim 3.8 (Poissonization of interactive proofs)**.** *A private Poissonized sample budget means that the verifier draws $s' \sim Poi(s)$ independently before the interaction and may use at most s' samples. For any interactive proof with completeness and soundness errors $\varepsilon_c, \varepsilon_s$:*

- *A verifier using at most q samples can be simulated with a private $Poi(s)$ sample pool, for any $s \geq 2q$, increasing each error by at most $e^{-s/8}$.*
- *A verifier with a private $Poi(s)$ sample budget can be simulated using at most $\lceil 2s \rceil$ samples, increasing each error by at most $e^{-s/3}$.*

These simulations preserve the communication, round, and honest-prover sample bounds, and allow adaptive verifier sample requests.

Proof. For the first direction, reject if $s' < q$; otherwise serve successive sample requests from the pool in their original order. For the second, draw the original private budget s' and reject if $s' > \lceil 2s \rceil$; otherwise use the first s' samples of a fixed pool of size $\lceil 2s \rceil$. Keep the pool size and unused samples private, except for any information revealed by the original protocol.

Couple each simulation to the original protocol using the same infinite independent sample stream and the same protocol coins. For every prover strategy, including one depending on the entire previous transcript, the interactions coincide unless the budget test fails. Thus each error changes by at most the corresponding tail probability in Claim 3.7. No additional messages or prover sample requests are needed. □

Claim 3.9 (Sample reconstruction and interaction)**.** *An independent sample represented by its multiplicity indicators $(v_i)_i$ can be recovered in distribution by taking each label in v_i exactly i times and privately ordering the resulting multiset uniformly. This preserves the joint distribution with the prover's initial information whenever that information and the verifier's sample are independent conditional on D.*

If two joint distributions of the prover's retained information and the verifier's multiplicity indicators are within η in TV, then applying the same interactive strategies to them changes the transcript and verdict distributions by at most η, provided the strategies use only their respective retained information, exchanged messages, and fresh independent coins. In particular, the acceptance probabilities differ by at most η, for any fixed cheating-prover strategy.

Proof. Conditional on the multiplicities, all orders of an independent sample have equal probability, even after fixing D and the prover's independent initial information. Fresh uniform ordering therefore recovers the original sample distribution and can serve adaptive requests. The ordering and subsequent interaction form a common randomized map of the retained joint setup. Total variation contracts under this map. □

Claim 3.10 (Sequential amplification)**.** *If an interactive proof has both errors at most $\varepsilon < 1/2$, then t sequential repetitions, for odd t, with fresh samples and coins and a majority verdict, have both errors at most*

$$\exp\left(-2t(1/2-\varepsilon)^2\right).$$

Sample, communication, and round bounds increase by at most a factor t. Combining independent Poissonized sample pools of mean s gives a total sample size distributed as $Poi(ts)$.

Proof. Restart the honest prover independently for completeness. For soundness, conditional on any previous transcript, the next verifier uses fresh randomness and samples, so any cheating prover's conditional probability of acceptance is at most ε. The number of unsuccessful repetitions is therefore stochastically dominated by $\mathrm{Bin}(t,\varepsilon)$, in either case. Its Chernoff bound gives the assertion. Independent Poisson budgets add to a Poisson budget of the summed mean. □

Claim 3.11 (Discretized normal approximation, [Fan14])**.** *Let $X \sim Bin(n,p)$, where $0 < p < 1$, and $Z' = \lceil Z \rceil$ for $Z \sim \mathcal{N}(np, np(1-p))$. There is an absolute constant C such that*

$$\delta_{TV}(X,Z') \le C \min\left\{1, \frac{1}{\sqrt{np(1-p)}}\right\}.$$

Proof. The discretized normal bound for sums of independent Bernoulli variables gives $\delta_{\mathrm{TV}}(X, \lfloor Z + 1/2 \rfloor) \le C/\sqrt{np(1-p)}$; see [Fan14, Theorem 1.3 and Remark 1.4]. Changing from nearest-integer rounding to ceiling changes the approximating normal mean by $1/2$. Total variation contracts under rounding, and this change costs $O(1/\sqrt{np(1-p)})$. For variance below one, use $\delta_{\mathrm{TV}} \le 1$. □

3.2 Communication Complexity

Set disjointness has been a central problem in communication complexity since the inception of the area [Yao79, Raz92, HW07, She14, RY20]. In this work we show that it can also be used for communication lower bounds for the new area of proofs for distribution properties.

Definition 3.12 (Set Disjointness)**.** *In the* set disjointness *problem* DISJ_N*, Alice holds a set* $A \subseteq [N]$ *and Bob holds a set* $B \subseteq [N]$*. Their goal is to determine whether* $A \cap B = \emptyset$ *with probability at least* $2/3$*, communicating over a public channel shared with a randomness source.*

In this work, we require the following variant of set disjointness:

Definition 3.13 (Gap Set Disjointness)**.** *Let* $N, g \in \mathbb{N}$ *with* $g \leq N$*. In the* Gap Disjointness *problem, denoted* $\texttt{GapDisj}_{N,g}$*, Alice holds a set* $A \subseteq [N]$ *and Bob holds a set* $B \subseteq [N]$*, with the following promise:*

- ***YES case:*** $A \cap B = \emptyset$*; or*
- ***NO case:*** $|A \cap B| \geq g$*.*

The goal is for Alice and Bob to distinguish between the two cases: they must output 1 *(accept) with high probability in the YES case, and output* 0 *(reject) with high probability in the NO case.*

If the no case is guaranteed to satisfy $|A \cap B| = g$*, we call the problem* exact gap disjointness*. Moreover, for sparsity parameters* $|A| = k_a$ *and* $|B| = k_b$*, we denote* $\texttt{GapDisj}_{N,g,k_a,k_b}$*.*

In particular, we use the following standard consequence of Razborov's unique-disjointness lower bound [Raz92]; see also the survey [She14, Section 4].

Theorem 3.14 (Communication complexity of exact gap disjointness)**.** *Let* $N, k_a, k_b, g \in \mathbb{N}$ *satisfy* $1 \leq g \leq k_a \leq k_b$ *and* $3k_a + k_b \leq N$*. The communication complexity of* exact $\texttt{GapDisj}_{N,g,k_a,k_b}$ *in the public-coin model, with error at most* $1/3$*, is* $\Omega\left(\frac{k_a}{g}\right)$*.*

Proof. Put $r := \lfloor k_a/g \rfloor$. Razborov's lower bound [Raz92] gives communication $\Omega(r)$ for sets of size r in a universe of size $4r$, with intersection promised to have size 0 or 1. Replace each universe element by a block of g elements. The resulting sets each have size rg, and their intersection has size 0 or g.

Add disjoint fixed sets of $k_a - rg$ Alice-only elements and $k_b - rg$ Bob-only elements. The total universe size is then

$$4rg + (k_a - rg) + (k_b - rg) = k_a + k_b + 2rg \leq 3k_a + k_b \leq N.$$

Pad the remaining positions with elements belonging to neither set. This preserves the intersection and gives the required input sizes, so the communication lower bound is $\Omega(r) = \Omega(k_a/g)$. □

Corollary 3.15 (Dense-input exact gap disjointness)**.** *Let* N, K, k_a, g *be positive integers with* $K = o(N)$*,* $k_a = o(K)$*, and* $1 \leq g \leq k_a$*. For all sufficiently large* N*, the public-coin communication complexity of exact* $\texttt{GapDisj}_{N,g,k_a,N-K}$*, with error at most* $1/3$*, is* $\Omega(k_a/g)$*.*

Proof. Apply Theorem 3.14 with $k_b = N - K$: the assumptions give $3k_a \leq K$ and $k_a \leq N - K$ for all sufficiently large N. □

Claim 3.16 (Total variation under conditioning and randomized maps)**.** *A common randomized map cannot increase total variation. Moreover,*

$$\delta_{TV}\left(\bigotimes_i P_i, \bigotimes_i Q_i\right) \le \sum_i \delta_{TV}(P_i, Q_i).$$

Two joint distributions with the same marginal on U satisfy

$$\delta_{TV}((U,X),(U,Y)) = \mathbb{E}_U\left[\delta_{TV}(X \mid U, Y \mid U)\right].$$

Proof. Sum the absolute differences of the probability masses, conditioning on the input for the first assertion and on U for the last. The product bound follows by replacing one factor at a time and applying the triangle inequality. □

Claim 3.17 (Integer shifts of a binomial variable)**.** *If $X \sim Bin(n,p)$, $0 < p < 1$, and $r \in \mathbb{Z}$, then*

$$\delta_{TV}(X+r, X) \le \min\left\{1, \frac{C|r|}{\sqrt{np(1-p)}}\right\}$$

for an absolute constant C.

Proof. Unimodality gives $\delta_{\mathrm{TV}}(X+1, X) = \max_k \Pr[X=k]$. Fourier inversion bounds this maximum by

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-2np(1-p)\sin^2(t/2)}\,dt \le \frac{C}{\sqrt{np(1-p)}}.$$

Here $\sin(|t|/2) \ge |t|/\pi$ on $[-\pi,\pi]$. Apply the triangle inequality $|r|$ times and use $\delta_{\mathrm{TV}} \le 1$. □

Claim 3.18 (Comparison and rounding of normal variables)**.** *Let $Z_b \sim \mathcal{N}(\mu_b, \sigma_b^2)$, where $\sigma_b > 0$. Then*

$$\delta_{TV}(Z_0, Z_1) \le C\frac{|\mu_0-\mu_1| + |\sigma_0-\sigma_1|}{\min\{\sigma_0,\sigma_1\}}.$$

For every integer r,

$$\delta_{TV}(\lceil Z_0\rceil + r, \lceil Z_1\rceil) \le \delta_{TV}(Z_0+r, Z_1).$$

Proof. Integrate the absolute derivatives of the normal density with respect to its mean and standard deviation, and apply the triangle inequality. The second assertion follows from $\lceil z+r\rceil = \lceil z\rceil + r$ for integer r and Claim 3.16. □

Claim 3.19 (Joint singleton and doubleton counts)**.** *Let $X_1,\dots,X_n$ be independent $Poi(\lambda)$ variables and $h_i = |\{x : X_x = i\}|$. Set $q_i = e^{-\lambda}\lambda^i/i!$. For independent $U \sim Bin(n,q_1)$ and $V \sim Bin(n,q_2)$,*

$$\delta_{TV}\left((h_1,h_2,h_3,\dots),(U,V,0,\dots)\right) \le \min\left\{1, \frac{7}{6}n\lambda^3\right\}.$$

Proof. For one label, the exclusive indicators of $X_x = 1$ and $X_x = 2$ have TV distance $2q_1q_2 \le \lambda^3$ from independent Bernoulli indicators with these same marginals. Replacing these pairs independently over the labels costs at most $n\lambda^3$ by Claim 3.16. Discarding multiplicities at least three costs at most $n\Pr[X_x \ge 3] \le n\mathbb{E}\binom{X_x}{3} = n\lambda^3/6$. Sum the indicators and apply the triangle inequality. □

4 Proof of the Communication Complexity Lower Bound

We follow the general framework described in Section 2.1, and provide the explicit construction of the setup protocol. We formally define the distributions $D_0(H)$ and $D_1(H, L)$ over domain $[N]$. Fix a domain size parameter $N \in 6\mathbb{N}$[8], and noise parameter $K := K(N) = o(N)$ with $0 < K < N/3$:

- For a heavy set $H \subseteq [N]$ of size K, the distribution $D_0(H)$ assigns each element in H a probability of $\frac{1}{2K}$, and each of the remaining $N - K$ elements in $[N] \setminus H$ the probability $\frac{1}{2(N-K)}$.
- For disjoint heavy and light sets $H, L \subseteq [N]$ of sizes K and $N/3 - K$, respectively, the distribution $D_1(H, L)$ assigns each element in H a probability of $\frac{1}{2K}$, and each element in L the probability $\frac{1}{2(N/3-K)}$, while the other $2N/3$ elements in $[N] \setminus (H \cup L)$ are outside the support.

Define the *canonical* distributions to be

$$D_0 := D_0\left(\{1, \ldots, K\}\right),$$
$$D_1 := D_1\left(\{1, \ldots, K\}, \{K+1, \ldots, N/3\}\right).$$

Denote $k := \log_N(K)$. The families $\mathcal{F}_0(N, K)$ and $\mathcal{F}_1(N, K)$ are defined as the closure under permutation of D_0 and D_1 respectively.

Definition 4.1 (Permutation of distribution)**.** *For two distributions P and Q over domain $\mathcal{X}$, we say that Q is a* permutation *of P and write $Q = \pi(P)$ if there exists a permutation π of $\mathcal{X}$ such that for every $x \in \mathcal{X}$, $Q(x) = P\left(\pi^{-1}(x)\right)$.*

Definition 4.2 (The `Noisy-Collisions-Problem (NCP)`)**.** *Fix parameters $N \in 6\mathbb{N}$ and $K = N^k$ such that $k \in (0.5, 1)$. For $b \in \{0, 1\}$, define:*

$$\mathcal{F}_b(N, K) = \{\pi\left(D_b\right) \mid \pi : [N] \to [N] \text{ is a permutation}\}$$

Define the noisy-collisions-problem *(NCP) to be the following promise problem:*

$$\texttt{NCP}(N, K) = (\mathcal{F}_0(N, K), \mathcal{F}_1(N, K))$$

Given parameters N and K, and samples from a distribution D promised to belong to either $\mathcal{F}_0(N, K)$ or $\mathcal{F}_1(N, K)$, decide to which family D belongs.

Claim 4.3 (Sample complexity of testing `NCP`)**.** *For $K = o(N)$, there exists a tester for $\texttt{NCP}(N, K)$ with error at most $1/3$ and sample complexity:*

$$O\left(\min\left\{\sqrt{N} + K \log N, \frac{N}{\sqrt{K}}\right\}\right)$$

samples, while $\Omega(\sqrt{N})$ samples are necessary.

[8] Divisibility is only for convenience; rounding the support sizes extends the construction to other domain sizes without changing the asymptotic bounds.

Proof. Identify the heavy set by frequency thresholds using $O(K \log N)$ samples, then collision-test the light part using $O(\sqrt{N})$ fresh samples. Alternatively, when the second bound is smaller, the Poissonized doubleton count has mean separation $\Theta(s^2/N)$ and variance $O(s^2/K)$; Chebyshev's inequality gives the bound at $s = CN/\sqrt{K}$ for a sufficiently large constant C (Claim 3.8 fixes the sample budget). For the lower bound, even revealing the heavy set leaves the two light-support models indistinguishable before a light collision, which requires the birthday scale $\Omega(\sqrt{N})$. □

Remark 4.4 (Other regimes for k)**.** *For fixed $0 \le k \le 1/2$, testing uses $\widetilde{\Theta}(\sqrt{N})$ samples, and the protocol offers no speedup. At $k = 1$, interpret the linear regime as $K = \alpha N$, for fixed $0 < \alpha < 1/3$. Collision testing then uses $\Theta(\sqrt{N})$.*

We show that every protocol for NCP with verifier sample complexity $s(N, K)$ and communication complexity $c(N, K)$ can be used to solve an instance of $\mathtt{GapDisj}_{N,g,k_a,k_b}$ (see Definition 3.13) with communication complexity $c(N, K)$, where $g := g(N, K, s(N, K)) = \Theta\left(\frac{s^2}{N}\right)$, $k_a = |a| := \left\lfloor \frac{s^2}{64K} e^{-s/(4K)} \right\rfloor$, and $k_b = N - K$. Bob's input indicates light elements. The corresponding dense-input source lower bound is Corollary 3.15; its communication scale for our parameters is $\Omega\left(\frac{k_a}{g}\right) = \Omega\left(\frac{N}{K}\right)$.

Thus the reduction below establishes the following communication lower bound for NCP in this sample regime:

$$c_{\mathtt{NCP}}(N, K) = \Omega\left(\frac{N}{K}\right). \tag{1}$$

We introduce some notations that will assist us with the reduction. The following definition aims to capture the information that each side in an NCP instance requires from the distribution D to run the protocol. We think of this information as available before any communication takes place, and it defines the view of the protocol (disregarding coin-flips). We call this information the *setup* of the protocol.

Definition 4.5 (The *setup* of an interactive proof for NCP)**.** *Fix an instance of $\mathtt{NCP}(N, K)$ with input distribution D, and verifier sample complexity $s(N, K)$. The setup of a protocol for this instance is composed of two components:*

- ***Verifier's sample*** *is a random variable that depends on the distribution D, and is represented by $v_1, \ldots, v_s \in \{0,1\}^N$, where v_i is an indicator of the elements that appeared i times in the sample.*
- ***Prover's information,*** *denoted as $p_{light}, p_{heavy} \in \{0,1\}^N$, indicating respectively which elements in the support of D are* light *(with probability $O\left(\frac{1}{N}\right)$) and which are* heavy *(with probability $\frac{1}{2K}$).*

Given this setup, each side can disregard the distribution D and run the protocol only with this information, using optional additional randomness (Claim 3.9).

Definition 4.6 (A random setup for an NCP protocol)**.** *For verifier sample complexity $s := s(N, K)$ and for $b \in \{0,1\}$, consider the joint random variables parameterized by N, K, s:*

$$\left(\left(\mathbf{P}^b_{light}, \mathbf{P}^b_{heavy}\right), \left(\mathbf{V}^b_1, \mathbf{V}^b_2, \ldots, \mathbf{V}^b_s\right)\right)(N, K, s)$$

defined to be the setup of a protocol for NCP with verifier sample complexity s, where the input distribution D was taken uniformly from $\mathcal{F}_b(N,K)$. This is a joint random variable with randomness over the choice of the distribution in the family (i.e. the heavy and light sets), as well as the verifier's sample.

For every $i \in [s]$ we also define $\mathbf{H}_i^b \in \{0,1\}^N$ and $\mathbf{L}_i^b \in \{0,1\}^N$ to be the random variables indicating the heavy *and* light *elements respectively that appeared i times in the verifier's sample. Note that for every i the random variables $\mathbf{H}_i^b, \mathbf{L}_i^b$ are disjoint and satisfy:*

$$\mathbf{H}_i^b + \mathbf{L}_i^b = \mathbf{V}_i^b$$

Poissonized protocols. Before showing the reduction from Disjointness to NCP, we first note that for the sake of the analysis we consider protocols for NCP where the verifier draws $s' \sim \mathrm{Poi}(s(N,K))$ samples. This *trick*, known as *Poissonization*, makes the analysis considerably easier. By Claim 3.8, a fixed verifier sample bound s can be replaced by a private $\mathrm{Poi}(2s)$ sample pool, preserving communication and increasing each error by at most $e^{-s/4}$. Thus communication lower bounds for Poissonized protocols transfer to standard protocols up to constant factors, with amplification justified by Claim 3.10.

4.1 A Reduction from Gap Disjointness to NCP

Proposition 4.7. *[Reducing from disjointness to NCP] For any parameter $N \in 3\mathbb{N}$, K such that $K = N^k$ for $k \in (0.5, 1)$, and sample parameter $s \geq C_1\sqrt{N}$ for a sufficiently large absolute constant C_1, with $s = o(\min\{K, N/\sqrt{K}\})$, assume Alice and Bob have inputs $a \in \{0,1\}^N$ and $b \in \{0,1\}^N$ respectively, such that $|b| = N - K$, $|a| = \left\lfloor \frac{s^2}{64K} e^{-s/(4K)} \right\rfloor$, and that they share (unlimited) public randomness. Bob's vector indicates the light elements in the first block $[N]$. Let $s' \sim Poi(s)$. There exists a 0-communication setup procedure for Alice and Bob, such that on non-abort Alice obtains vectors $u_1, u_2, \ldots, u_{s'} \in \{0,1\}^{2N}$, and Bob obtains vectors $p_{heavy}, p_{light} \in \{0,1\}^{2N}$ such that the following holds. In the TV comparisons, abort is a separate outcome.*

- *If $\langle a, b \rangle = 0$:*

$$\delta_{TV}\Big(\left((p_{light}, p_{heavy}), (u_1, \ldots, u_{s'})\right), \\ \left(\left(\mathbf{P}^0_{light}, \mathbf{P}^0_{heavy}\right), \left(\mathbf{V}^0_1, \ldots, \mathbf{V}^0_{s'}\right)\right)(2N, 2K, s')\Big) = o(1).$$

  Where the randomness on the left-hand side is over the shared randomness of Alice and Bob as well as the choice of s'.

- *If $\langle a, b \rangle = g(N,K,s) := \left\lfloor \frac{s^2}{4} \left(\frac{e^{-\frac{s}{2(2N/3-2K)}}}{2(2N/3-2K)} - \frac{e^{-\frac{s}{2(2N-2K)}}}{2(2N-2K)} \right) + \frac{1}{2} \right\rfloor = \Theta\left(\frac{s^2}{N}\right)$ then the distribution described by (p_{heavy}, p_{light}) belongs to $\mathcal{F}_0(2N, 2K)$, and:*

$$\delta_{TV}\left((u_1, \ldots, u_{s'}), \left(\mathbf{V}^1_1, \mathbf{V}^1_2, \ldots, \mathbf{V}^1_{s'}\right)(2N, 2K, s')\right) \leq 0.27$$

  Where the randomness of $(u_1, \ldots, u_{s'})$ is taken over Alice's setup process as well as the shared randomness r. The same bound holds jointly with p_{heavy} on the left and $\mathbf{P}^1_{heavy}$ on the right.

Before we prove it, we show that this is all we need to show a lower bound for NCP:

Corollary 4.8 (Lower bound for NCP)**.** *Let $(V,P)(N,K)$ be an interactive proof for $\mathtt{NCP}(N,K)$ with communication complexity $c(N,K)$, verifier sample count $Poi(s(N,K))$, and completeness and soundness errors at most $1/3$. For source parameters N,K, let s be the actual Poisson mean of the protocol at $(2N,2K)$ after a fixed constant amplification to both errors at most 0.01; thus $s=\Theta(s(2N,2K))$. Assume $\log_N(K)\in(0.5,1)$, $s\geq C_1\sqrt{N}$, and $s=o(\min\{K,N/\sqrt{K}\})$, where C_1 is the constant in Proposition 4.7. Then exact $\mathtt{GapDisj}_{N,g,k_a,k_b}$, with $k_a=|a|=\left\lfloor\frac{s^2}{64K}e^{-s/(4K)}\right\rfloor$, $k_b=N-K$, and $g:=g(N,K,s)$ as defined in that proposition, can be solved with error at most $1/3$ and communication $O(c(2N,2K))$, for all sufficiently large N.*

Proof. First amplify the protocol at $(2N,2K)$ by Claim 3.10, fixing its Poisson mean s before defining k_a and g. Have Alice and Bob run the setup procedure of Proposition 4.7 with this same s, and then simulate the amplified protocol: Alice plays the verifier using her sample, and Bob plays the YES prover for $D_0(\mathrm{Supp}(p_{\mathrm{heavy}}))$. Both disregard their original inputs and setup randomness thereafter, using only their setup information, exchanged messages, and fresh coins.

If the inputs are disjoint, completeness and the $o(1)$ setup error give acceptance probability at least $0.99-o(1)$. If their intersection has size $g(N,K,s)$, the setup is within 0.27 of an ideal NO setup jointly with Bob's heavy indicator. Bob's strategy is a possible cheating-prover strategy in that ideal interaction, so soundness and Claim 3.9 bound acceptance by $0.01+0.27=0.28$. Thus both errors are at most $1/3$ for sufficiently large N. The setup uses no communication, and the fixed constant amplification costs $O(c(2N,2K))$ bits. □

For these parameters, $k_a=\Theta(s^2/K)=o(K)$ and $g=\Theta(s^2/N)$, so $1\leq g\leq k_a$ for all sufficiently large N. Corollary 3.15 therefore gives a source communication lower bound $\Omega(k_a/g)=\Omega(N/K)$, and hence $c(2N,2K)=\Omega(N/K)$.[9]

Corollary 4.9 (Applications)**.** *Fix $K=\Theta(N^k)$ for $k\in(1/2,1)$ and verifier sample complexity $s=o(\min\{K,N/\sqrt{K}\})$. Estimating entropy to additive error $1/3$, distance from uniformity to additive error $1/20$, or support size to additive error $N/8$ requires communication $\Omega(N/K)$ in the worst case over $\mathcal{F}_0\cup\mathcal{F}_1$. The same lower bound holds for verifying $\mathtt{Ent}(N,1+\frac{1}{2}\log_2 K+\frac{1}{2}\log_2(N-K))$ with $\delta_c=0$ and $\delta_f=1/(4\log_2 N)$, and for tolerant uniformity verification with $\delta_c=1/2$ and $\delta_f=7/12$.*

Proof. For $D_b\in\mathcal{F}_b$, direct calculation gives

$$\begin{aligned} H(D_0)&=1+\tfrac{1}{2}\log_2 K+\tfrac{1}{2}\log_2(N-K), & H(D_1)&=1+\tfrac{1}{2}\log_2 K+\tfrac{1}{2}\log_2(N/3-K),\\ \delta_{\mathtt{TV}}(D_0,U_{[N]})&=\tfrac{1}{2}-K/N, & \delta_{\mathtt{TV}}(D_1,U_{[N]})&=\tfrac{2}{3},\\ |\mathrm{Supp}(D_0)|&=N, & |\mathrm{Supp}(D_1)|&=N/3. \end{aligned}$$

The entropy gap is $\frac{1}{2}\log_2((N-K)/(N/3-K))=\frac{1}{2}\log_2 3+o(1)$, so each estimator distinguishes the NCP families. For entropy verification, the continuity bound $|H(P)-H(Q)|\leq\delta\log_2(N-1)+h_2(\delta)$ [Aud07], where $\delta=\delta_{\mathtt{TV}}(P,Q)$ and h_2 is binary entropy, shows that every D_1 is at least $1/(4\log_2 N)$ from the displayed entropy level for sufficiently large N. Applying the NCP lower bound proves the claims; both families satisfy $\|D_b\|_2^2=\Theta(1/K)$. □

We now proceed to prove Proposition 4.7.

Figure 4.9.1: Setup Procedure

Input: parameters $N \in 3\mathbb{N}$, $k \in (0.5, 1)$, $K := N^k$, $s \geq C_1\sqrt{N}$ for a sufficiently large absolute constant C_1, and $s = o(\min\{K, N/\sqrt{K}\})$. Alice has vector $a \in \{0,1\}^N$ such that

$$|a| = \left\lfloor \frac{s^2}{64K} e^{-s/(4K)} \right\rfloor.$$

Bob has vector $b \in \{0,1\}^N$ such that $|b| = N - K$. Alice and Bob have unlimited shared randomness r.

1. **Parameter setting.** Define $p \in \{0,1\}^{2N}$ to be such that $p_i = 1$ iff $i \in \{N+1, N+2, \ldots, N+K\}$. Let $\pi : [2N] \to [2N]$ be a uniformly random permutation, defined by the shared randomness r.
2. **Bob.** Define $b' = \left((\mathbf{1}_N - b) \mid\mid \vec{\mathbf{0}}_N\right)$ (concatenating the complement of Bob's light indicator with the 0 vector of dimension N). Define $p_{\text{heavy}} = \pi(b' + p)$ and $p_{\text{light}} = \vec{\mathbf{1}}_{2N} - p_{\text{heavy}}$, and output $(p_{\text{light}}, p_{\text{heavy}})$.
3. **Alice.**
   (a) Set $s' \sim \text{Poi}(s)$. Draw a sample of size s' from the distribution $D_0([2K])$ over $[2N]$. For every $i \geq 1$, denote by $\ell_i, h_i \in \{0, \ldots, s'\}$ the numbers of *light* and *heavy* elements, respectively, that appeared i times in the sample. These counts are zero for $i > s'$. If $\ell_j > 0$ for some $j \geq 3$, or $h_2 < |a|$, abort[a]. Set $h_2 \leftarrow h_2 - |a|$.
   (b) **Choosing the *heavy* elements.** Independently of π, choose vectors $v_1^h, \ldots, v_{s'}^h \in \{0,1\}^{2N}$ to be pairwise disjoint vectors such that $\left|v_i^h\right| = h_i$ and $\text{Supp}(v_i^h) \subseteq \text{Supp}(p)$. If this is impossible, abort.
   (c) **Choosing the *light* elements.** Independently of π, choose vectors $v_1^\ell, \ldots, v_{s'}^\ell \in \{0,1\}^{2N}$ to be pairwise disjoint vectors such that $\left|v_i^\ell\right| = \ell_i$ and $\text{Supp}(v_i^\ell) \subseteq \{N+K+1, \ldots, 2N\}$. If this is impossible, abort.
   (d) **Producing Alice's setup output.** For all $i \in \{1\} \cup \{3, 4, \ldots, s'\}$, set $v_i \leftarrow v_i^h + v_i^\ell$, and set $v_2 \leftarrow v_2^h + v_2^\ell + a$. Finally, set $u_i \leftarrow \pi(v_i)$ and output $(u_1, \ldots, u_{s'})$.

[a] Note that since Alice knows this canonical distribution, she also knows which samples are heavy and which are light.

4.2 Proving Proposition 4.7

We show that the procedure in Figure 4.9.1 satisfies the conditions of Proposition 4.7.

The main insight behind the proof is that if we disregard the labels, a sample from $\mathcal{F}_0$ and a sample from $\mathcal{F}_1$ are almost identical, except for the number of light elements that appear exactly twice in the sample. Therefore, if we consider the intersection of Alice and Bob's inputs to be a collection of light elements that appeared exactly twice in the sample, when it is 0, the setup procedure produces a sample from $\mathcal{F}_0$, and when it is $g(N, K, s)$, the setup procedure produces a sample that is close to a sample from $\mathcal{F}_1$.

In order to show this, we first prove that indeed, disregarding the light elements that appear

[9] A verifier using fewer than $C_1\sqrt{N}$ samples can draw and ignore additional samples to meet this lower threshold, preserving communication and the stated upper sample regime.

twice, a sample from $\mathcal{F}_0$ and a sample from $\mathcal{F}_1$ are almost identical. Then, we show that the difference between the two cases in the number of light elements that appear twice in the sample is $g(N, K, s)$.

The simulated prover in the setup protocol always obtains a description of a distribution from $\mathcal{F}_0$, regardless of the inputs. For completeness, when the strings are disjoint, the verifier's sample is consistent with the distribution obtained by the prover.

First, we show that the random variables $\left(\mathbf{H}_i^b\right)$ as well as $\left(\mathbf{L}_i^b\right)$ for $i \neq 2$ are distributed identically (or almost identically) across $b \in \{0,1\}^N$. This will imply that the difference between the case that $b = 0$ and the case that $b = 1$ in NCP is only in the distribution of $\mathbf{L}_2^b$.

Claim 4.10. *For every choice of N, K, s, such that $s = \Omega(\sqrt{N})$, and $K, s = o(N)$ the following hold:*

- $\delta_{TV}\left(\left(\mathbf{H}_1^1, \ldots, \mathbf{H}_s^1\right), \left(\mathbf{H}_1^0, \ldots, \mathbf{H}_s^0\right)\right) = 0.$
- $\delta_{TV}\left(\left(\mathbf{L}_3^1, \ldots, \mathbf{L}_s^1\right), \left(\mathbf{L}_3^0, \ldots, \mathbf{L}_s^0\right)\right) = O\left(\frac{s^3}{N^2}\right)$
- $\delta_{TV}\left(\mathbf{L}_1^1, \mathbf{L}_1^0\right) = O\left(\frac{s^{3/2}}{N}\right)$

Proof. The first item is immediate, since by definition, $\mathbf{P}^1_{\text{heavy}} \equiv \mathbf{P}^0_{\text{heavy}}$, and every distribution in either case assigns 0.5 mass to the heavy elements in its support.

Moving on to the second item, write $n_0 := N - K$ and $n_1 := N/3 - K$. Fix $b \in \{0,1\}$ and condition on a distribution $D \in \mathcal{F}_b(N, K)$ with light set L, where $|L| = n_b$. For each $x \in L$, let X_x be the number of occurrences of x in the Poissonized sample. Then by Claim 3.7 $X_x \sim \text{Poi}\left(s \cdot \frac{1/2}{n_b}\right)$. For every such x:

$$\Pr\left[X_x \geq 3\right] = 1 - \sum_{i=0}^{2} e^{-s/(2n_b)} \frac{(s/(2n_b))^i}{i!} = O\left(\frac{s^3}{n_b^3}\right) = O\left(\frac{s^3}{N^3}\right).$$

Taking a union bound over all $n_b = O(N)$ light elements, we get that the probability that any light element appears at least three times in the sample is $O\left(\frac{s^3}{N^2}\right)$. This implies that with probability $1 - O\left(\frac{s^3}{N^2}\right)$, $\mathbf{L}_i^b = \mathbf{0}_N$ for all $i \geq 3$, which establishes the second item.

Finally, let's look at $\delta_{\text{TV}}\left(\mathbf{L}_1^1, \mathbf{L}_1^0\right)$, i.e. how different the distribution of *light elements that appear once in the verifier's sample* is if the input distribution is chosen uniformly from $\mathcal{F}_1$ or from $\mathcal{F}_0$. By Poissonization (Claim 3.7), we have that:

$$\left|\mathbf{L}_1^0\right| = \text{Bin}\left(N - K, \left(s \cdot \frac{1}{2(N-K)}\right) \cdot e^{-s \cdot \frac{1}{2(N-K)}}\right)$$

$$\left|\mathbf{L}_1^1\right| = \text{Bin}\left(N/3 - K, \left(s \cdot \frac{1}{2(N/3-K)}\right) \cdot e^{-s \cdot \frac{1}{2(N/3-K)}}\right)$$

To estimate the distance between these two distributions, denote $\mu_0' = \frac{s}{2} \cdot e^{-s/(2(N-K))}$, and $\mu_1' = \frac{s}{2} \cdot e^{-s/(2(N/3-K))}$. We get that:

$$\sigma_0' = \sqrt{\frac{s}{2} \cdot e^{-s/(2(N-K))}\left(1 - \frac{s}{2(N-K)} \cdot e^{-s/(2(N-K))}\right)}$$

$$\sigma'_1 = \sqrt{\frac{s}{2} \cdot e^{-s/(2(N/3-K))} \left(1 - \frac{s}{2(N/3-K)} \cdot e^{-s/(2(N/3-K))}\right)}$$

Note that the distance in bias is $O\left(\frac{s^2}{N}\right)$, while the standard deviation of both distributions is $\Theta\left(\sqrt{s}\right)$, and the difference in standard deviations is $O\left(\frac{s^{3/2}}{N}\right)$. Via an approximation to the normal distribution (see Claim 3.11), we get that:

$$\begin{aligned}
\delta_{\mathtt{TV}}\left(\left|\mathbf{L}_1^0\right|, \left|\mathbf{L}_1^1\right|\right) &\leq \delta_{\mathtt{TV}}\left(\mathcal{N}\left(\mu'_0, \sigma'^2_0\right), \mathcal{N}\left(\mu'_1, \sigma'^2_1\right)\right) + O\left(\frac{1}{\sigma'_0} + \frac{1}{\sigma'_1}\right) \\
&= O\left(\frac{|\mu'_0 - \mu'_1|}{\sigma'_0} + \frac{|\sigma'_0 - \sigma'_1|}{\sigma'_0} + \frac{1}{\sigma'_0}\right) \\
&= O\left(\frac{\frac{s^2}{N}}{\sqrt{s}} + \frac{\frac{s^{3/2}}{N}}{\sqrt{s}} + \frac{1}{\sqrt{s}}\right) \\
&= O\left(\frac{s^{3/2}}{N}\right)
\end{aligned}$$

We showed that the number of elements appearing several times (save for light elements appearing twice) is almost the same across both cases. Since both families are closed under permutation, we obtain the desired result on the full vectors (accounting for labels). □

We now consider $\mathbf{L}_2^0$ and $\mathbf{L}_2^1$. We show that these vectors carry the significant signal that distinguishes between the two cases of NCP.

Claim 4.11. *The following conditions hold:*

- $\left|\mathbf{L}_2^0\right| \equiv Bin\left(N-K, \frac{1}{2}\left(\frac{s}{2(N-K)}\right)^2 \cdot e^{-\frac{s}{2(N-K)}}\right)$.
- $\left|\mathbf{L}_2^1\right| \equiv Bin\left(N/3-K, \frac{1}{2}\left(\frac{s}{2(N/3-K)}\right)^2 \cdot e^{-\frac{s}{2(N/3-K)}}\right)$.

Also:

$$\mathbb{E}\left[\left|\mathbf{L}_2^1\right|\right] - \mathbb{E}\left[\left|\mathbf{L}_2^0\right|\right] = \frac{s^2}{4}\left(\frac{e^{-\frac{s}{2(N/3-K)}}}{2(N/3-K)} - \frac{e^{-\frac{s}{2(N-K)}}}{2(N-K)}\right)$$

$$\frac{Var\left[\left|\mathbf{L}_2^1\right|\right]}{Var\left[\left|\mathbf{L}_2^0\right|\right]} \leq 3.01$$

Proof. As in Claim 4.10, Poissonization makes the number of occurrences of each light element independent (Claim 3.7). Write $n_0 = N - K$ and $n_1 = N/3 - K$ for the sizes of the light sets. In case $b \in \{0, 1\}$, each light element appears exactly twice with probability

$$q_b := \Pr\left(\mathrm{Poi}\left(\frac{s}{2n_b}\right) = 2\right) = \frac{1}{2}\left(\frac{s}{2n_b}\right)^2 e^{-\frac{s}{2n_b}}.$$

Thus $\left|\mathbf{L}_2^b\right| \equiv \mathrm{Bin}(n_b, q_b)$, giving the two binomial distributions in the statement.

Denote their means and variances by μ_b and σ_b^2, respectively. The binomial formulas give

$$\mu_b = n_b \cdot q_b = \frac{s^2}{8n_b} e^{-\frac{s}{2n_b}}, \qquad \sigma_b^2 = \mu_b \cdot (1 - q_b).$$

In particular,

$$\mu_1 - \mu_0 = \frac{s^2}{4}\left(\frac{e^{-\frac{s}{2(N/3-K)}}}{2(N/3-K)} - \frac{e^{-\frac{s}{2(N-K)}}}{2(N-K)}\right).$$

Finally, since $K = o(N)$ and $s = o(N^{2/3})$, we have $q_0, q_1 = o(1)$, and hence

$$\frac{\sigma_1^2}{\sigma_0^2} = \frac{n_0}{n_1} \cdot \frac{1-q_1}{1-q_0} \cdot e^{\frac{s}{2n_0} - \frac{s}{2n_1}} = 3 + o(1) \le 3.01$$

for all sufficiently large N. □

Claim 4.12. *There exists an absolute constant C_0 such that, if $s \ge C_0\sqrt{N}$ and $K, s = o(N)$, the following holds. Denote*

$$g := \left\lfloor \mathbb{E}\left[\left|\mathbf{L}_2^1\right|\right] - \mathbb{E}\left[\left|\mathbf{L}_2^0\right|\right] + \frac{1}{2} \right\rfloor.$$

For all sufficiently large N:

$$\delta_{TV}\left(\left|\mathbf{L}_2^0\right| + g, \left|\mathbf{L}_2^1\right|\right) \le 0.269.$$

Proof. Write $X_b := |\mathbf{L}_2^b|$, $\mu_b := \mathbb{E}[X_b]$, and $\sigma_b^2 := \mathrm{Var}(X_b)$. With the light-set sizes $n_0 = N - K$ and $n_1 = N/3 - K$, Claim 3.7 gives

$$X_b \sim \mathrm{Bin}(n_b, q_b), \qquad q_b := \frac{1}{2}\left(\frac{s}{2n_b}\right)^2 e^{-s/(2n_b)}.$$

Thus $\mu_b = s^2 e^{-s/(2n_b)}/(8n_b)$ and $\sigma_b^2 = \mu_b(1-q_b)$. Since $K, s = o(N)$, for all sufficiently large N we have

$$\sigma_b \ge \frac{s}{8\sqrt{N}}, \qquad r := \frac{\sigma_1^2}{\sigma_0^2} = 3 + o(1) \in [1, 3.01].$$

Let $A \ge 1$ be an absolute constant in the upper bound of Claim 3.11, so that, for $Z_b \sim \mathcal{N}(\mu_b, \sigma_b^2)$ and $\widehat{Z}_b := \lceil Z_b \rceil$,

$$\delta_{\mathrm{TV}}(X_b, \widehat{Z}_b) \le \frac{A}{\sigma_b}.$$

We will take $C_0 := 10000A$; in particular, the variances are bounded below by an absolute constant large enough for this estimate.

$$\delta_{\mathrm{TV}}(X_0 + g, X_1) \le \delta_{\mathrm{TV}}(Z_0 + g, Z_1) + A\left(\frac{1}{\sigma_0} + \frac{1}{\sigma_1}\right).$$

By the definition of g, $|\mu_0+g-\mu_1| \le 1/2$. The total variation distance between normal distributions with a common standard deviation σ and means differing by d is at most $|d|/\sigma$. Consequently,

$$\begin{aligned}\delta_{\mathrm{TV}}(X_0+g, X_1) \le \delta_{\mathrm{TV}}\left(\mathcal{N}(\mu_1,\sigma_0^2),\mathcal{N}(\mu_1,\sigma_1^2)\right) \\ + A\left(\frac{1}{\sigma_0}+\frac{1}{\sigma_1}\right)+\frac{1}{2\sigma_0}.\end{aligned}$$

To bound the remaining normal distance, let Φ be the standard normal cumulative distribution function. Integrating the densities between their two crossing points gives, for $r > 1$,

$$\begin{aligned}&\delta_{\mathrm{TV}}\left(\mathcal{N}(\mu_1,\sigma_0^2),\mathcal{N}(\mu_1,\sigma_1^2)\right) \\ &\quad = 2\left[\Phi\left(\sqrt{\frac{r\ln r}{r-1}}\right)-\Phi\left(\sqrt{\frac{\ln r}{r-1}}\right)\right].\end{aligned}$$

This expression is increasing in r: with $a := \sqrt{\ln r/(r-1)}$, its derivative is $a\Phi'(a)/r > 0$. At $r = 3.01$ its value is approximately $0.26011 < 0.261$; at $r = 1$ the distance is zero. Hence the normal distance is at most 0.261 for our variance ratio.

Using $\sigma_b \ge s/(8\sqrt{N})$ and $A \ge 1$, we conclude that

$$\begin{aligned}\delta_{\mathrm{TV}}(X_0+g, X_1) &\le 0.261 + (16A+4)\frac{\sqrt{N}}{s} \\ &\le 0.261 + 20A\frac{\sqrt{N}}{s} \\ &\le 0.261 + \frac{20A}{C_0} = 0.263 \le 0.269.\end{aligned}$$

□

Claim 4.13. *Assume the parameters of Proposition 4.7, with $s \ge C_1\sqrt{N}$ for a sufficiently large absolute constant C_1. The TV comparisons below are unconditioned; the final counts refer to executions that do not abort:*

- *If $\langle a, b\rangle = 0$, then the final heavy and light doubleton counts are h_2 and ℓ_2, respectively, and:*
$$\delta_{TV}\left((h_2,\ell_2),\left(|\mathbf{H}_2^0|,|\mathbf{L}_2^0|\right)\right) = 0.$$
- *If $\langle a, b\rangle = g$, then the final heavy and light doubleton counts are $h_2 - g$ and $\ell_2 + g$, respectively, and*
$$\begin{aligned}\delta_{TV}\left(h_2-g,|\mathbf{H}_2^1|\right) &= O\left(\frac{g\sqrt{K}}{s}\right) = o(1),\\ \delta_{TV}\left(\ell_2+g,|\mathbf{L}_2^1|\right) &\le 0.269.\end{aligned}$$

These bounds also hold jointly over all multiplicities:

$$\begin{aligned}\delta_{TV}\left((h_i - g\mathbf{1}_{\{i=2\}})_{i\ge1},(|\mathbf{H}_i^1|)_{i\ge1}\right) &= o(1),\\ \delta_{TV}\left((\ell_i + g\mathbf{1}_{\{i=2\}})_{i\ge1},(|\mathbf{L}_i^1|)_{i\ge1}\right) &\le 0.269 + o(1).\end{aligned}$$

Here h_i, ℓ_i are the unconditioned initial counts, all ideal vectors use parameters $(2N, 2K)$, and $g := \lfloor \mathbb{E}[|\mathbf{L}_2^1|] - \mathbb{E}[|\mathbf{L}_2^0|] + 1/2\rfloor$.

Proof. Write $r := \langle a, b\rangle \in \{0, g\}$. Of the $|a|$ planted elements, precisely r are light and $|a| - r$ are heavy. Consequently, on non-abort, the final heavy and light doubleton counts are

$$(h_2 - |a|) + (|a| - r) = h_2 - r, \qquad \ell_2 + r.$$

The canonical sample is drawn from $D_0([2K])$ over $[2N]$, so its initial numbers of elements appearing twice in the sample are distributed as $(|\mathbf{H}_2^0|, |\mathbf{L}_2^0|)$. This proves the disjoint case where $\langle a, b\rangle = 0$.

For the case where $\langle a, b\rangle = g$, by Poissonization (Claim 3.7), before subtraction, the number of heavy elements appearing twice in the sample is distributed as

$$h_2 \sim \mathrm{Bin}(2K, q_{\mathrm{h}}), \qquad q_{\mathrm{h}} := \frac{1}{2}\left(\frac{s}{4K}\right)^2 e^{-s/(4K)}, \qquad \sigma_{\mathrm{h}}^2 := \mathrm{Var}(h_2) \sim \frac{s^2}{16K} \longrightarrow \infty.$$

The ideal heavy count has this same distribution in both cases. Subtracting the integer gap g changes the bias (mean) by g and leaves the variance unchanged. Since $g = \Theta(s^2/N)$, whereas $\sigma_{\mathrm{h}} = \Theta(s/\sqrt{K})$, the difference in bias is asymptotically smaller than the standard deviation:

$$\frac{|\mathbb{E}[h_2 - g] - \mathbb{E}[h_2]|}{\sigma_{\mathrm{h}}} = \frac{g}{\sigma_{\mathrm{h}}} = \Theta\left(\frac{s\sqrt{K}}{N}\right) = o(1),$$

Thus, we get that:

$$\delta_{\mathtt{TV}}(h_2 - g, h_2) = o(1).$$

For the light count, let $X_b := |\mathbf{L}_2^b|$, with light-set sizes $n_0 := 2N - 2K$ and $n_1 := 2N/3 - 2K$. Then

$$X_b \sim \mathrm{Bin}(n_b, q_b), \qquad q_b := \frac{1}{2}\left(\frac{s}{2n_b}\right)^2 e^{-s/(2n_b)}, \qquad \mathbb{E}[X_b] = \mu_{\ell,b}, \qquad \sigma_b^2 := \mathrm{Var}(X_b) = \mu_{\ell,b}(1 - q_b).$$

Since $K, s = o(N)$, for all sufficiently large N,

$$\sigma_b \geq \frac{s}{8\sqrt{N}}, \qquad \frac{\sigma_1^2}{\sigma_0^2} \in [1, 3.01].$$

Let $A \geq 1$ be the absolute constant in the upper bound of Claim 3.11. For $Z_b \sim \mathcal{N}(\mu_{\ell,b}, \sigma_b^2)$, that claim gives $\delta_{\mathtt{TV}}(X_b, \lceil Z_b \rceil) \leq A/\sigma_b$. Because g is an integer, $\lceil Z_0 \rceil + g = \lceil Z_0 + g \rceil$. Applying the triangle inequality and data processing under rounding therefore gives

$$\begin{aligned}
\delta_{\mathtt{TV}}(X_0 + g, X_1) &\leq \delta_{\mathtt{TV}}(Z_0 + g, Z_1) + A\left(\frac{1}{\sigma_0} + \frac{1}{\sigma_1}\right) \\
&\leq 0.261 + \frac{1}{2\sigma_0} + A\left(\frac{1}{\sigma_0} + \frac{1}{\sigma_1}\right) \\
&\leq 0.261 + 20A\frac{\sqrt{N}}{s}.
\end{aligned}$$

The second inequality uses $|\mu_{\ell,0} + g - \mu_{\ell,1}| \leq 1/2$ and the equal-mean normal variance comparison computed in the preceding proof. Taking $C_1 := 10000A$ makes this at most 0.263, and hence at most 0.269.

For heavy counts, condition on $(h_i)_{i\neq 2}$; then h_2 is binomial by Claim 3.7. Except with probability $O(s/K)$, at least K heavy labels remain unassigned, so its conditional variance is $\Theta(s^2/K)$; the same mean-shift argument applies. For each light label, replacing its exclusive singleton/doubleton indicators by independent ones costs $O((s/n_b)^3)$ in TV. Summing over labels and discarding multiplicities at least three costs $O(s^3/N^2) = o(1)$. Hence the light counts are $o(1)$-close to the product of their singleton and doubleton marginals, with higher counts zero; combining the singleton bound in Claim 4.10 with the doubleton bound above proves the joint light estimate. □

Claim 4.14 (Abort probability)**.** *For the parameters of Proposition 4.7, let*

$$\mu_{\mathrm{h}} := \frac{s^2}{16K} e^{-s/(4K)}, \qquad |a| = \left\lfloor \frac{\mu_{\mathrm{h}}}{4} \right\rfloor.$$

The setup in Figure 4.9.1 satisfies, uniformly over the promised inputs a, b,

$$\Pr[\text{abort}] \le e^{-9\mu_{\mathrm{h}}/32} + \frac{s}{2K} + \frac{s}{2(N-K)} + \frac{s^3}{192(N-K)^2} = o(1).$$

Moreover, $|a| = \Theta(s^2/K)$.

Proof. Throughout this proof, h_i, ℓ_i denote the counts in Alice's initial canonical sample, before subtracting $|a|$ from h_2. This sample is drawn from $D_0([2K])$ over $[2N]$, so Claim 3.7 gives

$$h_2 \sim \mathrm{Bin}\left(2K, \frac{1}{2}\left(\frac{s}{4K}\right)^2 e^{-s/(4K)}\right), \qquad \mathbb{E}[h_2] = \mu_{\mathrm{h}}.$$

Since $|a| \le \mu_{\mathrm{h}}/4$, a multiplicative Chernoff bound gives

$$\Pr[h_2 < |a|] \le e^{-9\mu_{\mathrm{h}}/32}.$$

Let $Z_{\mathrm{h}} := \sum_{i\ge 1} h_i$ and $Z_\ell := \sum_{i\ge 1} \ell_i$ be the numbers of distinct sampled heavy and light elements. Each is at most the number of observations of its type, so $\mathbb{E}[Z_{\mathrm{h}}], \mathbb{E}[Z_\ell] \le s/2$. After subtracting $|a|$, the heavy allocation is feasible whenever $Z_{\mathrm{h}} - |a| \le K$, because its reserve has size K. The light allocation is feasible whenever $Z_\ell \le N - K$. Markov's inequality therefore yields

$$\Pr[Z_{\mathrm{h}} - |a| > K] \le \frac{s}{2(K + |a|)} \le \frac{s}{2K}, \qquad \Pr[Z_\ell > N - K] \le \frac{s}{2(N-K)}.$$

These capacity conditions suffice to choose the pairwise disjoint allocation vectors.

Finally, put $n_\ell := 2N - 2K$. Each light element has an independent $\mathrm{Poi}(s/(2n_\ell))$ occurrence count. For $X \sim \mathrm{Poi}(\lambda)$, the third factorial moment gives $\Pr[X \ge 3] \le \mathbb{E}[\binom{X}{3}] = \lambda^3/6$. A union bound over the n_ℓ light elements gives

$$\Pr[\exists i \ge 3 : \ell_i > 0] \le n_\ell \cdot \frac{1}{6}\left(\frac{s}{2n_\ell}\right)^3 = \frac{s^3}{192(N-K)^2}.$$

Taking a union bound over the four abort events proves the displayed estimate. Under the stated sample assumptions, $s = o(K)$ and $s = o(N^{2/3})$, while

$$\mu_{\mathrm{h}} \sim \frac{s^2}{16K} \longrightarrow \infty$$

because $s = \Omega(\sqrt{N})$ and $K = o(N)$. Hence the total abort probability is $o(1)$, and rounding $\mu_{\mathrm{h}}/4$ down preserves $|a| = \Theta(s^2/K)$. □

We are now set to prove Proposition 4.7.

Proof of Proposition 4.7. By Claim 4.14, the setup aborts with probability $o(1)$. We account for aborts as an additional $o(1)$ total variation error, keeping the initial histogram unconditioned.

Case I. We start with the case that $\langle a, b\rangle = 0$. Looking at **Bob**'s output at the end of the setup procedure, by construction $|p_{\text{heavy}}| = (N - |b|) + |p| = 2K$, and $|p_{\text{light}}| = 2N - 2K$, corresponding to some distribution in $\mathcal{F}_0(2N, 2K)$.

We move to **Alice**. In the disjoint case, every element of $\text{Supp}(a)$ is heavy. Subtracting $|a|$ heavy doubletons and then planting a therefore restores the entire initial heavy and light histogram. Conditional on this histogram and Bob's heavy set, the uniform permutation allocates all heavy multiplicity classes uniformly within p_{heavy} and all light multiplicity classes uniformly within p_{light}.

Apply this uniform allocation rule to the unconditioned canonical histogram to define an ideal YES setup. On every successful execution, the actual setup has exactly this conditional allocation distribution, so the two setups can be coupled to agree whenever the actual procedure does not abort. Their joint total variation distance is therefore at most $\Pr[\text{abort}] = o(1)$, proving the first part of the proposition.

Case II. Assume $\langle a, b\rangle = g(N, K, s)$. First, as before, **Bob** holds a description of a distribution in $\mathcal{F}_0(2N, 2K)$. Writing $A = \text{Supp}(a)$ and $B = \text{Supp}(b)$, every planted element in $\pi(A \cap B)$ is light, while every planted element in $\pi(A \setminus B)$ is heavy.

Moving on to **Alice.** The final heavy and light multiplicity counts are $h_i - g\mathbf{1}_{\{i=2\}}$ and $\ell_i + g\mathbf{1}_{\{i=2\}}$, respectively. By Claim 4.13 and independence of the heavy and light counts, the total variation inequality for product distributions gives

$$\delta_{\text{TV}}\left(\left(h_i - g\mathbf{1}_{\{i=2\}}, \ell_i + g\mathbf{1}_{\{i=2\}}\right)_{i\geq 1}, \left(|\mathbf{H}_i^1|, |\mathbf{L}_i^1|\right)_{i\geq 1}\right) \leq 0.269 + o(1).$$

As in Case I, conditional on these counts and Bob's heavy set, the permutation allocates heavy labels uniformly within $\text{Supp}(p_{\text{heavy}})$ and light labels uniformly within $\text{Supp}(p_{\text{light}})$. The ideal NO sample has the same conditional allocation distribution after averaging over its light support. Thus data processing and the $o(1)$ abort bound give

$$\begin{aligned}\delta_{\text{TV}}\Big(&(p_{\text{heavy}}, (u_i)_{i=1}^{s'}),\\ &(\mathbf{P}_{\text{heavy}}^1, (\mathbf{V}_i^1)_{i=1}^{s'})(2N, 2K, s')\Big) \leq 0.269 + o(1) \leq 0.27\end{aligned}$$

for all sufficiently large N. Bob simulates the YES prover for $D_0(\text{Supp}(p_{\text{heavy}}))$ using only his heavy indicator, protocol messages, and fresh coins; Alice uses only her sample and fresh protocol coins. Both disregard their original inputs and setup randomness, as required by Claim 3.9. This proves the second part of the proposition. □

Remark 4.15 (Computational soundness)**.** *The reduction also applies to computationally sound proof systems with a poly(N)-time honest prover, as in [HR25a]. On a NO instance $D_1(H, L)$,*

a cheating prover can recover the heavy set H with probability $1 - o(1)$ by frequency thresholding $O(K \log N)$ independent samples, since $K = o(N)$, and then efficiently simulate the honest prover using fresh samples from the YES distribution $D_0(H)$. The setup guarantee in Proposition 4.7 controls the joint law of H and the verifier's sample, so computational soundness applies up to the stated total-variation bound and the $o(1)$ recovery error. An `NCP` *argument with smaller communication would therefore yield a correct* `GapDisj` *protocol with comparably small communication, contradicting the information-theoretic lower bound [Raz92, She14], which holds even with unlimited local computation.*

4.3 Lower Bound for Tolerant Identity Verification

We reduce `NCP` to tolerant identity verification: a protocol for the latter using $c(N)$ bits yields an NCP protocol using $O(c(N) + \log N)$ bits.

We start with an instance of $\texttt{NCP}(N, K)$ for $K = \Theta(N^k)$ and fixed $k \in (1/2, 1)$. Let D be the input distribution, promised to belong to either $\mathcal{F}_0(N, K)$ or $\mathcal{F}_1(N, K)$. For $j \in \{0, \ldots, N-1\}$, let $\rho_j(x) := 1 + ((x + j - 1) \bmod N)$ be the cyclic shift by j. The verifier chooses j uniformly at random, independently of D, and sets

$$Q := \rho_j(D_0), \qquad Q' := \frac{1}{8}D + \frac{7}{8}Q.$$

The distribution Q has an explicit description determined by j, while samples from Q' can be generated using samples from D and the known distribution Q.

Claim 4.16. *If $D \in \mathcal{F}_0(N, K)$, then, for every $Q \in \mathcal{F}_0(N, K)$:*

$$\delta_{TV}(Q', Q) \leq \frac{1}{16}$$

Proof. Write $D = D_0(H_D)$ and $Q = D_0(H_Q)$, where H_D and H_Q are their respective heavy sets. These distributions differ only on the symmetric difference of their heavy sets, so, for sufficiently large N,

$$\delta_{\texttt{TV}}(D, Q) = (K - |H_D \cap H_Q|)\left(\frac{1}{2K} - \frac{1}{2(N-K)}\right) \leq \frac{1}{2}.$$

Consequently, $\delta_{\texttt{TV}}(Q', Q) = \frac{1}{8}\delta_{\texttt{TV}}(D, Q) \leq 1/16$. □

Claim 4.17. *If $D \in \mathcal{F}_1(N, K)$, then, with high probability over the choice of Q:*

$$\delta_{TV}(Q', Q) \geq \frac{5}{48} - o(1).$$

In particular, for all sufficiently large N, this distance is at least $\frac{5}{64}$.

Proof. Write $D = D_1(H_D, L_D)$ and $Q = D_0(H)$, and let $Z := [N] \setminus (H_D \cup L_D)$, so that $|Z| = 2N/3$. Under a uniformly random cyclic shift, each element belongs to H with probability K/N, and hence

$$\mathbb{E}_Q[D(H)] = \sum_{x \in [N]} D(x) \Pr_Q[x \in H] = \frac{K}{N}.$$

By Markov's inequality, $D(H) \le \sqrt{K/N} = o(1)$ with probability at least $1 - \sqrt{K/N} = 1 - o(1)$. On this event, using $Q(H) = 1/2$, $D(Z) = 0$, and $|Z \setminus H| \ge 2N/3 - K$, we obtain

$$\begin{aligned}
\delta_{\mathtt{TV}}\left(Q', Q\right) &= \frac{1}{8}\delta_{\mathtt{TV}}(D, Q) \\
&\ge \frac{1}{8}\left(Q(H \cup Z) - D(H \cup Z)\right) \\
&= \frac{1}{8}\left(\frac{1}{2} + \frac{|Z \setminus H|}{2(N-K)} - D(H)\right) \\
&\ge \frac{1}{8}\left(\frac{1}{2} + \frac{2N/3 - K}{2(N-K)} - o(1)\right) \\
&= \frac{5}{48} - o(1),
\end{aligned}$$

□

Proof of Theorem 1.3. Fix $k \in (1/2, 1)$ and $K = \Theta(N^k)$. Assume that $(V_{\mathrm{id}}, P_{\mathrm{id}})$ is a tolerant identity verification protocol: given an explicit reference distribution Q, the parameters N, δ_c, δ_f, and sample access to an unknown distribution R, it accepts when $\delta_{\mathtt{TV}}(R, Q) \le \delta_c$ and rejects against every prover when $\delta_{\mathtt{TV}}(R, Q) \ge \delta_f$. Let its verifier sample complexity be $s(N) = o\left(\min\{K, N/\sqrt{K}\}\right)$, and let $c(N)$ be its maximum communication over the mixture instances above. Amplify the protocol to error at most 0.01 by a constant number of repetitions.

We construct a protocol for $\mathtt{NCP}(N, K)$ with sample access to D. The verifier chooses j and forms Q and Q' as above, and sends j to the prover using $\lceil \log_2 N \rceil$ bits. Both parties can then reconstruct the explicit reference distribution Q. They run the assumed identity verification protocol with the following inputs and sampling oracles:

$$\left(V_{\mathrm{id}}^{Q'}, P_{\mathrm{id}}^{Q'}\right)(N, Q, \delta_c = 1/16, \delta_f = 5/64).$$

Here the superscript Q' denotes sample access. Thus Q is the explicit distribution supplied as input, and $Q' = \frac{1}{8}D + \frac{7}{8}Q$ is the unknown distribution whose distance from Q is being verified. To answer each sample request to Q', the verifier independently draws from D with probability $1/8$, and otherwise draws from the explicitly known Q. The honest prover simulates its sample access to Q' in the same way, using its own samples from D. The NCP verifier outputs the verdict of V_{id}. Each simulated sample uses at most one sample from D, so the resulting verifier uses $O(s(N))$ samples.

If $D \in \mathcal{F}_0(N, K)$, Claim 4.16 gives $\delta_{\mathtt{TV}}(Q', Q) \le 1/16$ for every choice of j. Completeness of the identity protocol therefore gives an honest NCP prover for which the verifier accepts with probability at least 0.99. If $D \in \mathcal{F}_1(N, K)$, Claim 4.17 gives $\delta_{\mathtt{TV}}(Q', Q) \ge 5/64$ with probability $1 - o(1)$ over j. For each such j, soundness of the identity protocol bounds the acceptance probability by 0.01 against every prover strategy. This also covers a cheating NCP prover, since for fixed D and Q its behavior is one possible identity-prover strategy. Thus the NCP soundness error is at most $0.01 + o(1)$.

We have obtained a constant-error protocol for $\mathtt{NCP}(N, K)$ with communication complexity $O(c(N) + \log N)$ and verifier sample complexity $O(s(N))$. The NCP communication lower bound (1) implies

$$c(N) + \log N = \Omega(N/K).$$

Since $K = \Theta(N^k)$ with fixed $k < 1$, we have $\log N = o(N/K)$, and hence $c(N) = \Omega(N/K)$.

It remains to relate K to the collision probability of the identity input Q'. For either NCP family, and for every $Q \in \mathcal{F}_0(N, K)$,

$$\|D\|_2^2 = \frac{1}{4K} + O\left(\frac{1}{N}\right), \qquad \|Q\|_2^2 = \frac{1}{4K} + \frac{1}{4(N-K)}.$$

Since $Q' \geq (7/8)Q$ pointwise, and the squared norm is convex, we obtain

$$\frac{49}{256K} \leq \left(\frac{7}{8}\right)^2 \|Q\|_2^2 \leq \|Q'\|_2^2 \leq \frac{1}{8}\|D\|_2^2 + \frac{7}{8}\|Q\|_2^2 = \frac{1}{4K} + O\left(\frac{1}{N}\right).$$

Thus both $\|Q\|_2^2$ and $\|Q'\|_2^2$ are $\Theta(1/K)$. Equivalently,

$$K = \Theta\left(1/\|Q\|_2^2\right) = \Theta\left(1/\|Q'\|_2^2\right).$$

Consequently,

$$c(N) = \Omega\left(N \cdot \|Q'\|_2^2\right).$$

Here Q' is the unknown distribution denoted by D in the theorem statement. Moreover,

$$\min\{K, N/\sqrt{K}\} = \Theta\left(\min\left\{\frac{1}{\|Q'\|_2^2}, N\|Q'\|_2\right\}\right),$$

giving the sample condition stated in the theorem. □

4.4 Lower Bound for Tolerant Monotonicity Verification on the Boolean Cube

Write $M = 2^n$, $U = U_{\{0,1\}^n}$, and $d_{\mathsf{MON}}(P) = \inf_{Q \in \mathsf{MON}_n} \delta_{\mathsf{TV}}(P, Q)$, where MON_n consists of the coordinatewise nonincreasing distributions.

Background. Batu, Kumar, and Rubinfeld [BKR04] studied distribution monotonicity testing on total orders and grids; Bhattacharyya et al. [BFRV11] extended the study to general posets. Work on the cube includes testing under a monotonicity promise [RS05], sample lower bounds [AGP+19], and learning and testing algorithms [RV20]. Chakrabarty et al. [CCR+25] study the stronger subcube conditional sampling model. Here we bound communication for tolerant verification with ordinary samples.

Theorem 4.18 (Communication lower bound for tolerant cube monotonicity)**.** *For every fixed $\beta \in (1/2, 1)$, tolerant verification of MON_n with thresholds $\delta_c = 1/16$ and $\delta_f = 5/64$ requires worst-case communication $\Omega(M\|D\|_2^2)$ over input distributions with $\|D\|_2^2 = \Theta(M^{-\beta})$, whenever the verifier uses*

$$s(M) = o\left(\min\left\{\frac{1}{\|D\|_2^2}, M\|D\|_2\right\}\right)$$

samples, including $\widetilde{O}(\sqrt{M})$. The bound also holds for computationally sound arguments when both the honest prover and the nonuniform cheating prover run in time poly(M), with security parameter polynomially related to M.

Reduction idea. We randomly relabel an NCP instance onto the cube and mix it with uniformity. YES instances remain close to uniformity and hence to monotonicity. Farness from uniformity alone is insufficient for NO instances: a sparse distribution can itself be monotone. Instead, we show that after relabeling, almost all NO probability lies on a set occupying only about a third of every chain prefix in a fixed symmetric chain decomposition. A monotone distribution cannot put much mass on such a set, which forces a distance gap. Random affine bijections suffice, so describing the relabeling costs only $O(n^2)$ bits.

Lemma 4.19 (Sparse distributions after relabeling)**.** *Let P_0 have support size at most $M/3$. Relabel its probabilities by a uniformly random affine bijection $\pi(x) = Ax + b$ of $\mathbb{F}_2^n$, obtaining $P = \pi(P_0)$. Then $d_{\mathsf{MON}}(P) \geq 5/8$ with probability $1 - o(1)$, uniformly over P_0.*

Proof. Fix a symmetric chain decomposition with $L = \binom{n}{\lfloor n/2 \rfloor} = O(M/\sqrt{n})$ chains [dBvETK51], and put $t = \lceil n^{1/4} \rceil$. Call a chain bad if some prefix of length at least t has support density above $17/48$. Let E contain all bad chains and the first t vertices of every chain (or the whole chain if shorter).

Affine bijections are uniform on ordered triples of distinct points. Thus, even conditional on any fixed source atom x mapping to y, a prefix support count X_j satisfies $\mathbb{E}X_j \leq j/3 + 1$ and $\mathrm{Var}(X_j) \leq j/4$. Chebyshev's inequality on a geometric grid of prefix lengths with ratio $1 + 1/192$, together with monotonicity of prefix counts, bounds the conditional probability that a fixed chain is bad by $O(1/t)$. Averaging over source atoms with weights $P_0(x)$ therefore gives

$$\mathbb{E}_\pi[P(E)] \leq \frac{tL}{M} + O(1/t) = O(t/\sqrt{n} + 1/t) = o(1).$$

This conditioning handles heavy atoms as well. Hence $P(E) \leq 1/48$ with probability $1 - o(1)$.

On this event, $T = \mathrm{Supp}(P) \setminus E$ has density at most $17/48$ in every chain prefix. The restriction of any monotone Q to a chain is a nonnegative combination of uniform distributions on prefixes, so $Q(T) \leq 17/48$. Thus, simultaneously for every monotone Q,

$$\delta_{\mathsf{TV}}(P, Q) \geq P(T) - Q(T) \geq 47/48 - 17/48 = 5/8.$$

□

Lemma 4.20 (Mixing with uniformity)**.** *For every distribution P and $\lambda \in (0, 1]$, $d_{\mathsf{MON}}((1-\lambda)U + \lambda P) = \lambda d_{\mathsf{MON}}(P)$.*

Proof. Mixing a monotone approximation proves the upper bound. Conversely, put $Y = (1 - \lambda)U + \lambda P$ and $a = (1 - \lambda)/M$. For monotone Q, set $\overline{Q} = \max\{Q, a\}$ and $u = \sum_x (\overline{Q}(x) - Q(x))$. Clipping decreases $\|Y - Q\|_1$ by u. The monotone distribution $Q' = a + \lambda(\overline{Q} - a)/(\lambda + u)$ satisfies $\|\overline{Q} - Q'\|_1 = u$, so $\delta_{\mathsf{TV}}(Y, Q') \leq \delta_{\mathsf{TV}}(Y, Q)$. Since $(Q' - a)/\lambda$ is a monotone distribution, $\delta_{\mathsf{TV}}(Y, Q') \geq \lambda d_{\mathsf{MON}}(P)$. □

Proof of Theorem 4.18. Fix a monotonicity protocol with sample complexity $s(M)$ and worst-case communication $c(M)$ at the stated norm scale. Set $N = 12\lfloor M/12 \rfloor$ and $K = 2\lfloor M^\beta/2 \rfloor$. Given an $\mathtt{NCP}(N, K)$ instance P_0, pad it with zeros to the cube, apply a uniformly random affine bijection π, and simulate the protocol on

$$P = \pi(\overline{P}_0), \qquad D = \tfrac{7}{8}U + \tfrac{1}{8}P.$$

The verifier sends the $n^2 + n$ bits describing π; each simulated sample uses at most one source sample. Since both NCP families have $\|P_0\|_2^2 = 1/(4K) + O(1/N)$,

$$\|D\|_2^2 = \tfrac{1}{64}\|P_0\|_2^2 + \tfrac{63}{64M} = \Theta(1/K) = \Theta(M^{-\beta}).$$

Consequently, the assumed sample bound is $o(\min\{K, N/\sqrt{K}\})$.

For YES instances, $\delta_{\mathsf{TV}}(P, U) = 1/2 - K/M \le 1/2$, so $d_{\mathsf{MON}}(D) \le 1/16$. For NO instances, the support has size $N/3 \le M/3$; Lemmas 4.19 and 4.20 give $d_{\mathsf{MON}}(D) \ge 5/64$ with probability $1 - o(1)$.

Constant amplification, Poissonization, and padding to $\Theta(\sqrt{N})$ samples if needed yield an NCP protocol with $O(s(M) + \sqrt{N})$ samples, still within the required regime, and $O(c(M) + n^2)$ communication. Corollary 4.8, applied at source parameters $(N/2, K/2)$, implies $c(M) + n^2 = \Omega(N/K)$. Since $n^2 = o(N/K)$, this gives $c(M) = \Omega(N/K) = \Omega(M\|D\|_2^2)$.

Polynomial honest-prover time is preserved. For computational soundness, a nonuniform target cheating prover can hardwire P_0 and π, whose descriptions have polynomial size, to simulate source samples [HR25a]. □

5 Low Communication Protocol for Label-Invariant Distribution Properties

A *tagged sample* consists of independent draws $z_i \sim D$ paired with their probabilities $D(z_i)$, approximated by π_{z_i} in our protocol. With sufficiently accurate tags, bucket frequencies estimate bucket masses; dividing by representative probabilities estimates bucket sizes and hence the probability histogram [HR23, Remark 2.2]. This histogram suffices for estimating distance to the label-invariant properties covered by Theorem 1.2, under the computational assumptions of [HR22, Section 4] and [HR23, Section 6.3].[10] We focus on $D(x) \le \sigma^2/\sqrt{N}$ because there are at most $\sqrt{N}\sigma^{-2}$ heavier elements, whose contribution the verifier can learn to accuracy $O(\sigma)$ using $\widetilde{O}(\sqrt{N}\sigma^{-4})$ samples without the prover. The heavy-light reduction of [HR23, Section 6.2] combines this learned part with the verified lighter part. Thus, it suffices to construct the tagged-sample protocol of Proposition 5.1.

Proposition 5.1 (New result: low communication tagged-sample protocol)**.** *For sufficiently large N, every distance parameter $\sigma \in [N^{-1/10}, 1/100]$, and every distribution D over $[N]$ satisfying $D(x) \le \sigma^2/N^{1/2}$ for all x, there is a 2-message interactive protocol with the following guarantees. The verifier has black-box sample access to D; for completeness, the honest prover knows D exactly.*

- ***Completeness.*** *With probability at least 0.99, the verifier accepts and returns a tagged D-sample $((z_i, \pi_{z_i}))_{i\in[n]}$ of random size*

$$n = \widetilde{\Theta}\big(N\|D\|_2^2\sigma^{-6}\big)\,,$$

such that

$$\forall i \in [n], \qquad |\pi_{z_i} - D(z_i)| = O(\sigma^2/\log N)\pi_{z_i}.$$

[10] For additive entropy error ε, use histogram accuracy $O(\varepsilon/\log N)$ and grid parameter $O(\varepsilon)$; see [HR22, Claim 6.9]. For support-size error εN, impose minimum nonzero probability η/N on D and the reconstructed histogram and use accuracy $O(\eta\varepsilon)$; see [HR22, Claim 6.10]. Our hard families allow $\eta = 1/2$; internal accuracies must stay in the stated range.

- ***Soundness.*** *For any cheating prover, with probability at least* 0.99 *the verifier either rejects or its output satisfies*

$$\frac{1}{n}\sum_{i\in[n]}\left(1-\min\left\{\frac{D(z_i)}{\pi_{z_i}},\frac{\pi_{z_i}}{D(z_i)}\right\}\right)<\sigma,$$

where a zero tag on a D-sample contributes one.

The verifier's sample complexity and runtime are $\widetilde{O}(N^{1/2}\sigma^{-4})$, and the communication complexity is $\widetilde{O}(N\|D\|_2^2\sigma^{-6})$, in expectation and with probability at least 0.99.

With sample access alone, the honest prover can instead use $\widetilde{O}(N\sigma^{-4})$ independent samples. With probability at least 0.99, all tests pass and the output has average error below σ; the stronger per-tag completeness guarantee above is reserved for exact knowledge of D. Section A.4, under "Accuracy of the honest prover," proves this implementation through all tests.

Accuracy and resource bounds. The bounds in Theorem 1.2 hold in expectation and with constant high probability, for $CN^{-1/10}\le\delta_f-\delta_c\le c$ with sufficiently large C and sufficiently small c; see Proposition 5.1 and Remark A.15. Property-specific decision and witness costs, including any finer histogram accuracy they require, are additional; the displayed bounds describe the complete protocol when these costs fit within them. The entropy conversion incurs only polylogarithmic losses.

Remark 5.2 (Tagging larger samples)**.** *The goal here is to minimize communication and verifier sample complexity while obtaining a tagged sample. If instead we want to tag a larger sample with accuracy σ, we can retain the tier analysis and follow the large-sample approach of [HR23]: use Protocol A.0.1 with a fixed expected sample size $m'\ge C\sqrt{N}\sigma^{-4}\log^{40}N$ from each of D and $U_{[N]}$, for a sufficiently large constant C. Omit the first sample used to calculate m' and send all mixed-sample and collision labels, each only once. The fresh collision sample and the remaining tests are as in that protocol. Under the probability bound of Proposition 5.1, increasing m' strengthens mixed-count concentration, and the collision estimates use only normalized bucket sizes. This gives a tagged sample of size $\Theta(m')$ with high probability, with verifier sample complexity and runtime, and communication $\widetilde{O}(m')$, hence only a σ^{-4} dependence in the required sample size. With the compatibility needed for samples already used in the interaction of [HR24], this would improve their sample and communication bounds to $\widetilde{O}(N^{9/10})$ and $\widetilde{O}(N^{19/20})$, respectively, for constant proximity gap, with the circuit-depth communication term unchanged.*

Protocol A.0.1 and its full proof appear in Appendix A. The proof retains the consistency and collision tests and divides buckets into tiers, treating each error scale separately. Section A.4 derives the σ^{-6} communication bound and discusses the accuracy sufficient for a sample-based prover to pass the tests. The stronger completeness guarantee for every output tag stated above uses exact knowledge of D.

References

[AGP+19] Maryam Aliakbarpour, Themis Gouleakis, John Peebles, Ronitt Rubinfeld, and Anak Yodpinyanee. Towards testing monotonicity of distributions over general posets. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second*

Conference on Learning Theory, volume 99 of *Proceedings of Machine Learning Research*, pages 34–82. PMLR, 2019.

[Aud07] Koenraad M. R. Audenaert. A sharp continuity estimate for the von neumann entropy. *Journal of Physics A: Mathematical and Theoretical*, 40(28):8127–8136, 2007.

[BBCS26] Ari Biswas, Mark Bun, Clément L. Canonne, and Satchit Sivakumar. Interactive proofs for distribution testing with conditional oracles. In Shubhangi Saraf, editor, *17th Innovations in Theoretical Computer Science Conference, ITCS 2026, Bocconi University, Milan, Italy, January 27-30, 2026*, LIPIcs, pages 18:1–18:13. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2026.

[BCG17] Eric Blais, Clément L. Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. In *32nd Computational Complexity Conference (CCC 2017)*, volume 79 of *Leibniz International Proceedings in Informatics*, pages 28:1–28:40. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017.

[BFR+00] Tugkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *41st Annual Symposium on Foundations of Computer Science, FOCS 2000, 12-14 November 2000, Redondo Beach, California, USA*, pages 259–269. IEEE Computer Society, 2000.

[BFRV11] Arnab Bhattacharyya, Eldar Fischer, Ronitt Rubinfeld, and Paul Valiant. Testing monotonicity of distributions over general partial orders. In Bernard Chazelle, editor, *Proceedings of the 2nd Symposium on Innovations in Computer Science*, pages 239–252. Tsinghua University Press, 2011.

[BKR04] Tugkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In László Babai, editor, *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, June 13-16, 2004*, pages 381–390. ACM, 2004.

[Can22] Clément L. Canonne. Topics and techniques in distribution testing: A biased but representative sample. *Found. Trends Commun. Inf. Theory*, 19(6):1032–1198, 2022.

[CCR+25] Deeparnab Chakrabarty, Xi Chen, Simeon Ristic, C. Seshadhri, and Erik Waingarten. Monotonicity testing of high-dimensional distributions with subcube conditioning. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 1019–1030. ACM, 2025.

[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.

[dBvETK51] N. G. de Bruijn, C. van Ebbenhorst Tengbergen, and D. Kruyswijk. On the set of divisors of a number. *Nieuw Archief voor Wiskunde, Series 2*, 23:191–193, 1951.

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

[Fan14] Xiao Fang. Discretized normal approximation by Stein's method. *Bernoulli*, 20(3):1404–1431, 2014.

[GGR98] Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *J. ACM*, 45(4):653–750, 1998.

[GHR25] Oded Goldreich, Tal Herman, and Guy N. Rothblum. Interactive proof systems for FARNESS. *Electron. Colloquium Comput. Complex.*, TR25, 2025.

[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.

[GMR+25] Parikshit Gopalan, Raghu Meka, Prasad Raghavendra, Mihir Singhal, and Avi Wigderson. The communication complexity of distributed estimation. *CoRR*, abs/2511.21015, 2025.

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Electron. Colloquium Comput. Complex.*, (20), 2000.

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

[GV99] Oded Goldreich and Salil P. Vadhan. Comparing entropies in statistical zero knowledge with applications to the structure of SZK. In *Proceedings of the 14th Annual IEEE Conference on Computational Complexity, Atlanta, Georgia, USA, May 4-6, 1999*, page 54. IEEE Computer Society, 1999.

[Her24] Tal Herman. Public coin interactive proofs for label-invariant distribution properties. In Amit Kumar and Noga Ron-Zewi, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2024, August 28-30, 2024, London School of Economics, London, UK*, volume 317 of *LIPIcs*, pages 72:1–72:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024.

[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. In *65th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2024, Chicago, IL, USA, October 27-30, 2024*, pages 528–538. IEEE, 2024.

[HR25a] Tal Herman and Guy N. Rothblum. How to verify any (reasonable) distribution property: Computationally sound argument systems for distributions. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[HR25b] Tal Herman and Guy N. Rothblum. Proving natural distribution properties is harder than testing them. In *66th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2025, Sydney, Australia, December 14-17, 2025*, pages 2003–2016. IEEE, 2025.

[HW07] Johan Håstad and Avi Wigderson. The randomized communication complexity of set disjointness. *Theory Comput.*, 3(1):211–219, 2007.

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

[Raz92] Alexander A. Razborov. On the distributional complexity of disjointness. *Theor. Comput. Sci.*, 106(2):385–390, 1992.

[RS05] Ronitt Rubinfeld and Rocco A. Servedio. Testing monotone high-dimensional distributions. In Harold N. Gabow and Ronald Fagin, editors, *Proceedings of the 37th Annual ACM Symposium on Theory of Computing, Baltimore, MD, USA, May 22-24, 2005*, pages 147–156. ACM, 2005.

[RV20] Ronitt Rubinfeld and Arsen Vasilyan. Monotone probability distributions over the boolean cube can be learned with sublinear samples. In Thomas Vidick, editor, *11th Innovations in Theoretical Computer Science Conference, ITCS 2020, January 12-14, 2020, Seattle, Washington, USA*, volume 151 of *LIPIcs*, pages 28:1–28:34. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020.

[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.

[RY20] Anup Rao and Amir Yehudayoff. *Communication Complexity and Applications*. Cambridge University Press, 2020.

[She14] Alexander A. Sherstov. Communication complexity theory: Thirty-five years of set disjointness. In Erzsébet Csuhaj-Varjú, Martin Dietzfelbinger, and Zoltán Ésik, editors, *Mathematical Foundations of Computer Science 2014 - 39th International Symposium, MFCS 2014, Budapest, Hungary, August 25-29, 2014. Proceedings, Part I*, Lecture Notes in Computer Science, pages 24–43. Springer, 2014.

[Vad99] Salil Vadhan. *A Study of Statistical Zero-Knowledge Proofs*. PhD thesis, USA, 1999.

[Val11] Paul Valiant. Testing symmetric properties of distributions. *SIAM J. Comput.*, 40(6):1927–1968, 2011.

[VV10] Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. *Electron. Colloquium Comput. Complex.*, 17:179, 2010.

[Yao79] Andrew Chi-Chih Yao. Some complexity questions related to distributive computing (preliminary report). In Michael J. Fischer, Richard A. DeMillo, Nancy A. Lynch, Walter A. Burkhard, and Alfred V. Aho, editors, *Proceedings of the 11h Annual ACM Symposium on Theory of Computing, April 30 - May 2, 1979, Atlanta, Georgia, USA*, pages 209–213. ACM, 1979.

A Low Communication Protocol - Full Description

We present the protocol for tagging samples with low communication underlying Theorem 1.2, with formal guarantees stated in Proposition 5.1. The construction builds on ideas from [HR23]; we include the full protocol and analysis for completeness. For further details on extending the protocol to general distributions and using tagged samples to estimate label-invariant distribution properties, we refer the reader to [HR23].

We use the assumption $D(x) \le \sigma^2/\sqrt{N}$ of Proposition 5.1. The permitted tags the prover uses in its message are zero and

$$\frac{e^{j\tau}}{N}, \qquad \left\lfloor \frac{-19 \log N}{\tau} \right\rfloor \le j \le \left\lfloor \frac{\log N}{\tau} \right\rfloor .$$

Thus there are at most $21 \log N/\tau$ positive buckets for sufficiently large N, and a tag is encoded by its grid index using $O(\log N)$ bits. We first analyze an honest prover with exact knowledge of D.

We follow the consistency and collision arguments of [HR23]. Sample sizes are Poisson, as in Section 3.1.1; the first sample sets m' independently of the samples used in the tests. We keep the verifier's source choices, repeated occurrences, and equal pairs private. In particular, sending each label once does not discard multiplicities in the tests. The Poissonized output has random size $\Theta(m')$ with high probability; resource bounds below are stated in expectation and with high probability.

The basic tests first exclude large errors on individual labels and bound the mass assigned very small tags. The consistency test compares each bucket's size with its mass, while the collisions test compares its equal pairs with the number predicted by its tags. We show that honest tags pass both tests and that a bucket with large average error cannot pass both with high probability. Section A.4 then uses the tiers to detect large total error while testing only a small fraction of the buckets.

Throughout this appendix, the first sample determines m and m'. Conditional on the first sample, the fresh collision sample $(y_i)_i$ and the mixed sample $(z_i, b_i)_i$ are independent and have sizes $\mathrm{Poi}(s)$ and $\mathrm{Poi}(2m')$, respectively. Indices range over the samples actually drawn. We analyze an honest prover that knows D exactly and rounds its probabilities to the tag grid.

A.1 Component I: Verifier Basic Tests

These tests detect large discrepancies before the verifier runs the two bucket tests. They check that:

- the prover assigns one valid alleged probability π_x to each sent label x, even if x appears several times in the verifier's samples;
- for every sent label with $\pi_x \ge \sigma^4/\sqrt{N}$, its frequency in an independent fresh sample from D agrees with π_x up to a constant factor. Labels with smaller tags have frequency at most a constant multiple of $\sigma^4/\sqrt{N}$;
- for each accuracy σ_k, only a small fraction of the mixed samples drawn from D receive tags below $\sigma/(10^6 N \sigma_k \log N)$.

Figure A.0.1: *Low Communication* Verified Tagged Sample Protocol

Input: domain size parameter $N \in \mathbb{N}$, distance parameter $\sigma \in \left[N^{-1/10}, 1/100\right]$, as well as sample access to distribution D over domain $[N]$. As in Proposition 5.1, assume $D(x) \le \sigma^2/N^{1/2}$ for all x. Set $\tau = \sigma^2/(100 \log N)$ and $s = \lceil cN^{1/2}\sigma^{-3}\log^{10} N \rceil$, for a sufficiently large universal constant c.

1. **V:**
   (a) **Sample size.** Draw a $\mathrm{Poi}(s)$ sample from D and let m be its number of equal pairs. Set $m' = \lceil c \max\{m, 1\} \log^{20} N \rceil$.
   (b) **Collisions set.** Independently draw a fresh $\mathrm{Poi}(s)$ sample from D, denoted by $(y_i)_i$. Let C be the set of *elements* that appeared at least twice. Keep all equal pairs, including those from elements appearing more than twice.
   (c) **Mixed sample.** Draw a fresh $\mathrm{Poi}(2m')$ sample size. For each of these draws, toss a balanced coin to obtain $b_i \in \{0,1\}$. If $b_i = 0$, draw $z_i \leftarrow D$; otherwise, draw $z_i \leftarrow U_{[N]}$.

   Send $\{z_i\}_i \cup C$ to the prover randomly mixed, with each label appearing once.
2. **P:** For every label x sent, round $D(x)$ down to the tag grid defined above if $D(x) \ge N^{-20}$, and otherwise return $\pi_x = 0$. Send back $(x, \pi_x)_x$.

Verifier Tests

3. **Basic Tests.** Run the tests in Figure A.0.2.
4. **Advanced Tests.** Divide the samples into buckets according to their alleged probability and the distribution from which they were sampled. All indices below range over the samples actually drawn. For every $j \in \mathbb{Z}$ define:

$$V_j = \left\{ i \;\middle|\; b_i = 0,\ \pi_{z_i} \in \left[\frac{e^{j\tau}}{N}, \frac{e^{(j+1)\tau}}{N}\right) \right\}$$

$$W_j = \left\{ i \;\middle|\; b_i = 1,\ \pi_{z_i} \in \left[\frac{e^{j\tau}}{N}, \frac{e^{(j+1)\tau}}{N}\right) \right\}$$

$$C_j = \left\{ (i,i') \;\middle|\; i < i',\ y_i = y_{i'},\ \pi_{y_i} \in \left[\frac{e^{j\tau}}{N}, \frac{e^{(j+1)\tau}}{N}\right) \right\}$$

   Further divide the buckets into tiers according to their *size*. For every integer $0 \le t \le \left\lceil \log_2\left(\frac{2\cdot 10^6 \log N}{\tau\sigma}\right)\right\rceil$, set

$$\mathrm{tier}_t = \left\{ j \in \mathbb{Z} \;\middle|\; |V_j| \in m' \cdot \left[\frac{\tau\sigma}{10^6 \log N} \cdot 2^t, \frac{\tau\sigma}{10^6 \log N} \cdot 2^{t+1}\right) \right\}.$$

   For each nonempty tier t and $1 \le k \le \min\{t+1, \lceil \log_2(4/\sigma)\rceil\}$, set $\sigma_k = 2^{-k}$. Independently choose $\lceil 10^6(\sigma_k/\sigma)\log^3 N\rceil$ uniform bucket indices from the tier, with replacement; test all buckets if this exceeds the tier's size. Skip a chosen bucket j if

$$|W_j| < c\sigma_k^{-4}\log^4 N \quad \text{or} \quad \frac{e^{j\tau}}{N} < \frac{\sigma}{10^6 N \sigma_k \log N}.$$

   For every other chosen bucket:

   (a) **Consistency Test.** Reject unless:

$$\left| |W_j| - |V_j| \cdot e^{-j\tau} \right| \le \frac{\sigma_k^2}{20} \cdot |V_j| \cdot e^{-j\tau}$$

   (b) **Collisions Test.** Reject unless:

$$\left| |C_j| - \frac{s^2}{2m'} |V_j| \frac{e^{j\tau}}{N} \right| \le \frac{\sigma_k^2}{5} \left(\frac{s^2}{2m'} |V_j| \frac{e^{j\tau}}{N} \right).$$

Output: If all tests pass, output $(z_i, \pi_{z_i})_{i:b_i=0}$ in a fresh private uniform order.

Figure A.0.2: Basic Tests

Input: the labels sent in Protocol A.0.1, the returned tags, and the verifier's private mixed sample $(z_i, b_i)_i$.

1. **Elementary coherence.** Reject missing, conflicting, or invalid tags. A valid tag is zero or one of the grid values specified in Protocol A.0.1.
2. **Basic probability checks.** Draw $s' = \lceil c\sqrt{N}\sigma^{-4}\log^2 N \rceil$ fresh samples from D, independently of the response. For each label x sent to the prover, let Z_x be its number of occurrences in this fresh sample. Reject unless:
   (a) if $\pi_x \geq \sigma^4/\sqrt{N}$, then $s'\pi_x/2 \leq Z_x \leq 2s'\pi_x$;
   (b) if $\pi_x < \sigma^4/\sqrt{N}$, then $Z_x \leq 2s'\sigma^4/\sqrt{N}$.
3. **Small tags.** For every $k \leq \lceil \log_2(4/\sigma) \rceil$, reject unless
$$\left|\{i : b_i = 0,\ \pi_{z_i} < \sigma/(10^6 N\sigma_k \log N)\}\right| \leq \frac{10m'\sigma}{10^6\sigma_k \log N} + \log^2 N.$$
4. **Sample size.** Reject unless $m'/2 \leq |\{i : b_i = 0\}| \leq 2m'$.

Accept if all checks pass.

The last check prevents the prover from assigning very small probabilities to many D-samples and thereby hiding them in buckets that the verifier skips. The sample-size check ensures that normalizing the output error by m' changes it by at most a constant factor.

Claim A.1. *The basic tests satisfy the following:*

- ***Completeness.*** *The honest prover passes with probability at least* 0.999.
- ***Soundness.*** *Except with probability N^{-20}, either the verifier rejects or every sent label satisfies*
$$D(x) \leq 4\max\{\pi_x, \sigma^4/\sqrt{N}\}.$$
If $\pi_x \geq \sigma^4/\sqrt{N}$, also $D(x) \geq \pi_x/4$. The small-tag and sample-size inequalities hold whenever the tests pass.

Proof. Fix the message and response before drawing the fresh sample. Each Z_x has mean $s'D(x)$. If $D(x) > 4\max\{\pi_x, \sigma^4/\sqrt{N}\}$, passing requires a downward deviation by at least half the mean; if $D(x) < \pi_x/4$ and $\pi_x \geq \sigma^4/\sqrt{N}$, it requires an upward deviation by a factor at least two. Chernoff bounds and $s'\sigma^4/\sqrt{N} \geq c\log^2 N$ give probability N^{-30} for either event. A union bound over $[N]$ proves soundness.

For honest tags, $D(x) \in [\pi_x, e^\tau\pi_x)$ whenever $\pi_x > 0$, while zero tags have $D(x) < N^{-20}$. The same Chernoff bounds prove the probability checks. For each k, the globally defined set of honest tags below $\sigma/(10^6 N\sigma_k \log N)$ has D-mass at most $e^\tau\sigma/(10^6\sigma_k \log N) + N^{-19}$. Its mixed D-count is Poisson with this mass times m' by Claim 3.7. Chernoff, including the additive $\log^2 N$ slack, proves all small-tag checks. The total mixed D-count is distributed as $\mathrm{Poi}(m')$, so the sample-size check also passes with overwhelming probability. □

We also record two facts used below: very large true probabilities carry little mass, and the first sample chooses m' within a constant factor of its intended value.

Claim A.2. *For every distribution D and every $\alpha \in (0,1)$,*

$$D\big(\{x : D(x) \geq \alpha^{-1}\|D\|_2^2\}\big) \leq \alpha.$$

Proof. The sum of $D(x)^2$ over this set is at least $\alpha^{-1}\|D\|_2^2$ times its D-mass and is at most $\|D\|_2^2$. □

Claim A.3. *With probability at least 0.999 over the first sample,*

$$m' = \Theta\big(s^2\|D\|_2^2 \log^{20} N\big) = \Theta\big(N\|D\|_2^2\sigma^{-6}\log^{40} N\big)\,.$$

Proof. Apply Claim A.8 to the first sample and substitute the protocol's definitions of m' and s. □

A.2 Component II: The Consistency Test

For an honest bucket j, each label has probability roughly $e^{j\tau}/N$. Thus the number of mixed D-samples in this bucket should be roughly $e^{j\tau}$ times its number of uniform samples. The consistency test checks this relation. By itself it does not certify the tags: its soundness guarantee below is one inequality on the conditional means, which the collisions test will complement.

We first recall why the hidden origins of the mixed samples can be chosen after their labels. Let $Z \sim (D + U_{[N]})/2$, with

$$\Pr[B = 0 \mid Z = x] = \frac{D(x)}{D(x) + 1/N}, \qquad \Pr[B = 1 \mid Z = x] = \frac{1/N}{D(x) + 1/N}.$$

Claim A.4. *Conditional on the first sample, $Poi(2m')$ independent draws of (Z, B) have exactly the same distribution as the mixed sample in Protocol A.0.1.*

Proof. Both constructions give each draw the distribution

$$\Pr[(Z,B) = (x,0)] = D(x)/2, \qquad \Pr[(Z,B) = (x,1)] = 1/(2N).$$

Conditional on their common sample size the pairs are independent; average over this size, as in Claim 3.7. □

For the rest of the analysis, fix a first sample satisfying Claim A.3. Unless stated otherwise, $\mathbb{E}$ denotes expectation in the mental experiment conditional on this first sample, the complete sent message, and the prover's independent randomness and samples. In particular, the returned tags and their buckets are fixed, while the verifier's private origins and multiplicities remain random. The counts for different labels remain independent by Claim A.9, so we can apply concentration bounds even when the prover chooses its tags dishonestly.

For each label $x \in [N]$, define its sending probability, conditional only on the first sample:

$$\begin{aligned} q_x &:= \Pr[x \text{ is sent} \mid \text{first sample}] \\ &= 1 - (1 + sD(x))e^{-sD(x) - m'(D(x)+1/N)}. \end{aligned} \tag{2}$$

Conditioning on its being sent increases the mean of any count that vanishes when x is not sent by the factor $1/q_x$. In particular,

$$\mathbb{E}[|V_j|] = m' \sum_{x:\pi_x = e^{j\tau}/N} \frac{D(x)}{q_x}, \qquad \mathbb{E}[|W_j|] = \frac{m'}{N} \sum_{x:\pi_x = e^{j\tau}/N} \frac{1}{q_x},$$

where sums are over sent labels. We use the parameter $\beta = \tau\sigma/(10^6 \log N)$ from the protocol.

A.2.1 Concentration of the Mixed Counts

We need two guarantees. Buckets with too small a conditional D-count are not eligible for testing. For the remaining buckets, the observed D-count and average error are close to their conditional means; the same holds for the uniform count whenever it exceeds the protocol's threshold.

Claim A.5. *Except with probability N^{-20}, the following hold simultaneously for every bucket j and scale σ_k:*

- *If $\mathbb{E}[|V_j|] < m'\beta/(10\sigma_k)$, then $|V_j| < m'\beta/(2\sigma_k)$. Otherwise,*

$$\big||V_j| - \mathbb{E}[|V_j|]\big| \le \frac{\sigma_k^2}{1000}\mathbb{E}[|V_j|],$$

  and

$$\left|\frac{\sum_{i\in V_j}(1-\min\{D(z_i)/\pi_{z_i}, \pi_{z_i}/D(z_i)\})}{|V_j|} - \frac{\mathbb{E}\Big[\sum_{i\in V_j}(1-\min\{D(z_i)/\pi_{z_i}, \pi_{z_i}/D(z_i)\})\Big]}{\mathbb{E}[|V_j|]}\right| \le \frac{\sigma_k}{100}.$$

- *If $|W_j| \ge c\sigma_k^{-4}\log^4 N$, then*

$$\mathbb{E}[|W_j|] \ge \frac{c}{2}\sigma_k^{-4}\log^4 N, \qquad \big||W_j| - \mathbb{E}[|W_j|]\big| \le \frac{\sigma_k^2}{1000}\mathbb{E}[|W_j|].$$

  If $\mathbb{E}[|W_j|] \ge 2c\sigma_k^{-4}\log^4 N$, then $|W_j| \ge c\sigma_k^{-4}\log^4 N$.

Proof. For each sent label, its mixed D-count before conditioning is $\mathrm{Poi}(m'D(x))$. Its conditional moment generating function is

$$1 + \frac{\exp(m'D(x)(e^u-1)) - 1}{q_x} \le \exp\left(\frac{m'D(x)}{q_x}(e^u-1)\right).$$

Indeed, $1 + r(v-1) \le v^r$ for $r \ge 1$; the expression on the left is nonnegative here. The same holds for the uniform count. Thus independent labels give the usual Poisson Chernoff bounds, also for sums of D-counts with weights in $[0,1]$.

Apply these bounds to $|V_j|$ and to its error sum. In the significant case the exponent for relative precision $\sigma_k^2/1000$ is at least

$$c\sigma_k^4\mathbb{E}[|V_j|] \ge cm'\beta\sigma_k^3 \ge c\log^{38} N,$$

since $N\|D\|_2^2 \ge 1$ and $\sigma_k > \sigma/8$. The weighted error sum needs additive precision $\sigma_k\mathbb{E}[|V_j|]/1000$ and has a larger exponent. The same upper-tail bound proves the small-mean implication. For W_j, the guard gives exponent at least $c\log^4 N$; an upper tail also excludes passing the guard with mean below half its threshold. There are $O(\log N/\tau)$ buckets and $O(\log N)$ scales, so a union bound proves the claim. □

A.2.2 The Consistency Test is Complete and Sound

Claim A.6 (Single Bucket Consistency Test)**.** *On the event of Claim A.5, every eligible bucket above the uniform-count guard satisfies:*

- ***Completeness.*** *Honest tags pass the consistency test.*
- ***Soundness.*** *If the consistency test passes, then*

$$\frac{\mathbb{E}[|W_j|]}{\mathbb{E}[|V_j|]} \le \left(1 + 0.08\sigma_k^2\right) e^{-j\tau}. \tag{3}$$

Consequently, the conditional mean of the reciprocal-probability sum satisfies

$$\frac{\mathbb{E}\left[\sum_{i\in V_j} 1/D(z_i)\right]}{\mathbb{E}[|V_j|]} \le \left(1 + 0.08\sigma_k^2\right) N e^{-j\tau}. \tag{4}$$

Proof. Eligibility gives $|V_j| \ge m'\beta/(2\sigma_k)$, hence the significant-mean conclusions of Claim A.5 apply.

Completeness. For honest tags, $D(x)/\pi_x \in [1, e^\tau)$ and therefore

$$e^{-\tau} \le \frac{\mathbb{E}[|W_j|]}{e^{-j\tau}\mathbb{E}[|V_j|]} \le 1.$$

Thus the conditional means satisfy the required relation up to the rounding of the tags. The concentration errors and $\tau = o(\sigma_k^2)$ put the observed ratio within the test tolerance $\sigma_k^2/20$, so the verifier accepts.

Soundness. Passing the test bounds the observed uniform count in terms of the observed D-count. Both counts are close to their conditional means by Claim A.5. Substituting these bounds into the acceptance inequality gives (3), with the slightly larger tolerance $0.08\sigma_k^2$. Moreover,

$$\mathbb{E}\left[\sum_{i\in V_j} \frac{1}{D(z_i)}\right] = m' \sum_{\substack{x:\pi_x = e^{j\tau}/N \\ D(x)>0}} \frac{1}{q_x} \le N\mathbb{E}[|W_j|],$$

which proves (4). The inequality allows uniform samples from labels with $D(x) = 0$; such labels never occur among the D-samples in V_j. □

A.2.3 Buckets Below the Uniform-Count Guard

The verifier skips the consistency and collisions tests when there are too few uniform samples in the bucket to guarantee the required precision. We must therefore show that such buckets cannot account for much of the output error. The next claim bounds their combined error at each scale σ_k, rather than requiring every skipped bucket to have small error.

Claim A.7. *Except with probability* 0.001*, simultaneously for every k, eligible buckets whose empirical average error lies in $(\sigma_k, 2\sigma_k]$ and whose conditional uniform-count mean is at most $2c\sigma_k^{-4}\log^4 N$ contribute at most*

$$O\left(\frac{\sigma^2}{\sigma_k}\log^{-17} N\right) m'$$

to the output error. Summed over all k, this is less than $\sigma m'/1000$ for sufficiently large N.

Proof. Conditional only on the first sample,

$$\mathbb{E}\left[\sum_{x \text{ sent}} \frac{D(x)^2}{q_x}\right] = \|D\|_2^2.$$

Markov's inequality bounds this sum by $10^4\|D\|_2^2$, except with probability 10^{-4}. On this event, denote the buckets in the claim by BAD. Cauchy–Schwarz and the at most $21 \log N/\tau$ grid buckets give

$$\sum_{j \in \texttt{BAD}} \frac{\mathbb{E}[|V_j|]}{m'} \le \sqrt{\frac{21 \log N}{\tau} \frac{2Nc\sigma_k^{-4} \log^4 N}{m'} 10^4 \|D\|_2^2}.$$

By Claim A.5, the actual error in these buckets is at most $4\sigma_k \sum_{j \in \texttt{BAD}} \mathbb{E}[|V_j|]$. Substitute Claim A.3 and $\tau = \sigma^2/(100 \log N)$ into the displayed bound to obtain the stated $O(\sigma^2 \sigma_k^{-1} \log^{-17} N)m'$. The Markov event is common to all scales, and $\sum_k \sigma_k^{-1} = O(\sigma^{-1})$, proving the final assertion. □

This bounds the error that can be missed because of the uniform-count guard. For the remaining eligible buckets, the consistency guarantee above applies. We next show how the collisions test turns that guarantee into rejection of incorrect tags.

A.3 Component III: the Collisions Test

We use the same buckets and tiers as in the consistency test. The basic tests ensure, except with probability N^{-20}, that every sent label satisfies

$$D(x) \le 4 \max\{\pi_x, \sigma^4/\sqrt{N}\}.$$

Unlike the set C sent to the prover, $|\widetilde{C}_j|$ counts equal pairs, including all pairs contributed by a label occurring more than twice.

Claim A.8. *For a sufficiently large choice of the absolute constant in s, the pair count in the first sample satisfies, with probability at least .999,*

$$\frac{s^2}{4}\|D\|_2^2 \le m \le s^2\|D\|_2^2, \qquad m' = \Theta\big(N\|D\|_2^2 \sigma^{-6} \log^{40} N\big).$$

Proof. Poisson factorial moments and independence across labels (Claim 3.7) give

$$\mathbb{E}m = \frac{s^2}{2}\|D\|_2^2, \qquad \mathrm{Var}(m) = \frac{s^2}{2}\|D\|_2^2 + s^3 \sum_x D(x)^3.$$

Since $\sum_x D(x)^3 \le \|D\|_2^3$ and $\|D\|_2^2 \ge 1/N$, Chebyshev bounds the probability of a relative deviation exceeding $1/2$ by

$$O\left(\frac{N}{s^2} + \frac{\sqrt{N}}{s}\right) < .001.$$

The definition $m' = \lceil c \max\{m, 1\} \log^{20} N \rceil$ gives the second assertion. □

A.3.1 The Mental Experiment

The prover chooses its tags after seeing the entire message, which includes labels from C. Whether a label belongs to C depends on its private occurrences in $(y_i)_i$, so this dependence must be included in the analysis. Fix the first sample and hence m'. The collision sample and the mixed sample are fresh and independent of the first sample. The experiment below generates the message first and then generates the verifier's private counts conditional on it. Fixing the message and the prover's randomness fixes all tags and buckets; the remaining counts are independent across labels, as the next claim shows.

For the experiment below, fix a domain size N, a distribution D over $[N]$, and the distance parameter σ of Protocol A.0.1. Set

$$s = \left\lceil c\sqrt{N}\sigma^{-3}\log^{10} N \right\rceil, \qquad m' = \left\lceil c\max\{m,1\}\log^{20} N \right\rceil.$$

Here m is the number of equal pairs in the independent first sample. Fix this first sample before running the experiment. The fresh collision and mixed samples have expected sizes s and $2m'$. Neither the first sample nor m' is given to the prover, whose tag π_x is its claimed value of $D(x)$.

For each label $x \in [N]$, the three counts below are independent before conditioning, by Claim 3.7 in Section 3.1.1:

count	sample	distribution
$\lvert\{i : y_i = x\}\rvert$	fresh collision sample from D	$\mathrm{Poi}(sD(x))$
$\lvert\{i : z_i = x,\ b_i = 0\}\rvert$	mixed sample, from D	$\mathrm{Poi}(m'D(x))$
$\lvert\{i : z_i = x,\ b_i = 1\}\rvert$	mixed sample, from $U_{[N]}$	$\mathrm{Poi}(m'/N)$

The first count is the number of occurrences of x among $(y_i)_i$: $x \in C$ iff this count is at least two, and it contributes $\binom{\lvert\{i:y_i=x\}\rvert}{2}$ pairs to its collision bucket. The second and third counts are the numbers of occurrences of x among the mixed draws from D ($b_i = 0$) and from $U_{[N]}$ ($b_i = 1$), respectively; every repeated occurrence is counted. Counts for different labels are also independent. The last two counts sum to $\lvert\{i : z_i = x\}\rvert$. The message contains x precisely when

$$\lvert\{i : y_i = x\}\rvert \geq 2 \quad \text{or} \quad \lvert\{i : z_i = x\}\rvert > 0.$$

Claim A.9. *Fix a prover strategy. Let S, S' be the joint random variables*

$$\big((y_i)_i, (z_i, b_i)_i, (\pi_x)_x\big)$$

generated by Protocol A.0.1 and Mental Experiment A.8.1, respectively. Conditional on the first sample,

$$\delta_{TV}(S, S') = 0.$$

Conditional also on the sent message and the prover's randomness, the vectors

$$\big(\lvert\{i : y_i = x\}\rvert,\ \lvert\{i : z_i = x, b_i = 0\}\rvert,\ \lvert\{i : z_i = x, b_i = 1\}\rvert\big), \qquad x \in [N],$$

are mutually independent across labels.

Mental Experiment A.8.1: Alternative Production of the View of Protocol A.0.1

1. **V:** Independently for each $x \in [N]$, send x with probability

$$1 - (1 + sD(x))e^{-sD(x)-m'(D(x)+1/N)}.$$

   Order the message uniformly at random.

2. **P:** Return (x, π_x) for every label sent.

3. **V:** Independently for each x, draw its three counts from the table *jointly conditional on whether x was sent*, using the rule above. Privately order $(y_i)_i$ uniformly, and order all mixed pairs $(z_i, b_i)_i$ uniformly (Claim 3.9).

After fixing the message and the prover's randomness, counts for different labels remain independent.

Proof. Claim 3.7 gives the independent Poisson distributions in the table. A label is absent precisely when its first count is at most one and its other counts vanish, giving the displayed sending probability. Fixing the message conditions each label's three counts separately on its presence or absence. This preserves independence across labels. Uniformly ordering the resulting samples recovers their original joint distribution, including for an interactive prover (Claim 3.9). □

In the rest of this subsection, expectations are conditional on the first sample, the sent message, and the prover's randomness, as in the consistency analysis. Recall the sending probability q_x from (2), which is conditional only on the first sample. The labels in bucket j have tag $\pi_x = e^{j\tau}/N$.

The next claim gives the gap used for soundness: if a bucket's average true probability is close to its tag $e^{j\tau}/N$, but its tags have large error, then its average reciprocal probability must be larger than $Ne^{-j\tau}$. This will contradict the upper bound from the consistency test.

Claim A.10 (Conditional means and the gap)**.** *The conditional means satisfy*

$$\mathbb{E}|\widetilde{V}_j| = m' \sum_{x:\pi_x = e^{j\tau}/N} \frac{D(x)}{q_x},$$

$$\mathbb{E}|\widetilde{W}_j| = \frac{m'}{N} \sum_{x:\pi_x = e^{j\tau}/N} \frac{1}{q_x},$$

$$\mathbb{E}|\widetilde{C}_j| = \frac{s^2}{2} \sum_{x:\pi_x = e^{j\tau}/N} \frac{D(x)^2}{q_x}.$$

In particular,

$$\mathbb{E}|\widetilde{C}_j| = \frac{s^2}{2m'} \mathbb{E}\left[\sum_{i \in \widetilde{V}_j} D(z_i)\right],$$

$$Var(|\widetilde{C}_j|) \le \left(1 + 8s \max\left\{\frac{e^{j\tau}}{N}, \frac{\sigma^4}{\sqrt{N}}\right\}\right) \mathbb{E}|\widetilde{C}_j|.$$

The variance bound assumes the displayed basic probability bound for the bucket's labels. Moreover,

when $\mathbb{E}|\widetilde{V}_j| > 0$*, the following gap holds:*

$$Ne^{-j\tau}\frac{\mathbb{E}\left[\sum_{i\in\widetilde{V}_j} D(z_i)\right]}{\mathbb{E}|\widetilde{V}_j|} + \frac{e^{j\tau}}{N}\frac{\mathbb{E}\left[\sum_{i\in\widetilde{V}_j} 1/D(z_i)\right]}{\mathbb{E}|\widetilde{V}_j|}$$
$$\geq 2 + \left(\frac{\mathbb{E}\left[\sum_{i\in\widetilde{V}_j}(1-\min\{ND(z_i)e^{-j\tau}, e^{j\tau}/(ND(z_i))\})\right]}{\mathbb{E}|\widetilde{V}_j|}\right)^2. \tag{5}$$

The expectations use the conditioning stated above; the reciprocal sum is well defined because $D(z_i) > 0$ *for every* D*-sample.*

Proof. Each count is zero when its label is not sent, so conditioning on presence divides its mean by q_x. Poisson factorial moments give unconditional pair mean $s^2D(x)^2/2$ and variance $(1 + 2sD(x))s^2D(x)^2/2$. Conditioning on presence can only reduce the variance divided by q_x; sum over independent labels and apply the basic probability bound.

For the gap, for every $u > 0$,

$$Nue^{-j\tau} + \frac{e^{j\tau}}{Nu} - 2 = \frac{(Nu - e^{j\tau})^2}{Nue^{j\tau}}$$
$$\geq \left(1 - \min\{Nue^{-j\tau}, e^{j\tau}/(Nu)\}\right)^2.$$

Average this inequality with weights $m'D(x)/q_x$ over the sent labels of positive D-probability in the bucket, and divide by $\mathbb{E}|\widetilde{V}_j|$. Cauchy–Schwarz bounds the average squared error below by the square of the average error, proving (5). □

A.3.2 The Collisions Test is Sound

Fix a bucket and an error level σ_k. The consistency test gives an upper bound on the conditional mean of $\sum_{i\in\widetilde{V}_j} 1/D(z_i)$. If the tags have average error at least σ_k and the collisions test passes, the gap in (5) gives a strictly larger lower bound on the same quantity. The only exception is when the observed number of pairs is much smaller than its conditional mean. The next claim bounds this exception and makes the contradiction explicit.

Claim A.11 (The Collisions Test: Single Bucket)**.** *Suppose that the basic probability bound holds for every label in the bucket and that*

$$\frac{\mathbb{E}|\widetilde{V}_j|}{m'} \geq \frac{\beta}{10\sigma_k}, \qquad \frac{e^{j\tau}}{N} \geq \frac{\sigma}{10^6 N\sigma_k \log N}.$$

Except for an event $I_{j,k}$ *of conditional probability* $O((\sigma/\sigma_k)^2 \log^{-6} N)$*, the following holds on the mixed-count event of Claim A.5: if the bucket is above the uniform-count guard, has empirical mean error at least* σ_k*, and passes the collisions test, then*

$$\frac{\mathbb{E}\left[\sum_{i\in\widetilde{V}_j} 1/D(z_i)\right]}{\mathbb{E}|\widetilde{V}_j|} \geq (1 + .48\sigma_k^2)Ne^{-j\tau}. \tag{6}$$

This contradicts the consistency bound (4), so at least one test rejects.

Proof. Let $I_{j,k}$ be empty if $\mathbb{E}|\widetilde{C}_j| < \frac{s^2 e^{j\tau}}{2m'N}\mathbb{E}|\widetilde{V}_j|$. Otherwise, let it be the event that the observed pair count is smaller than its conditional mean by at least a $\sigma_k^2/8$ fraction:

$$I_{j,k} = \left\{ |\widetilde{C}_j| < (1 - \sigma_k^2/8)\mathbb{E}|\widetilde{C}_j| \right\}.$$

Chebyshev and Claim A.10 give, using the two size conditions,

$$\Pr(I_{j,k}) \le O\left(\frac{m'N(1 + s\max\{e^{j\tau}/N, \sigma^4/\sqrt{N}\})}{\sigma_k^4 s^2 e^{j\tau}\mathbb{E}|\widetilde{V}_j|} \right) = O\big((\sigma/\sigma_k)^2 \log^{-6} N\big).$$

Indeed, substitute $s = \Theta(\sqrt{N}\sigma^{-3}\log^{10} N)$ and $\beta = \tau\sigma/(10^6 \log N)$. The constant and $s\sigma^4/\sqrt{N}$ terms give $O((\sigma/\sigma_k)^2 \log^{-7} N)$; the remaining term is $O(1/(\sqrt{N}\,\sigma_k^3 \log^8 N))$, which satisfies the claimed bound since $\sigma_k \ge \sigma/8$ and $\sigma \ge N^{-1/10}$.

If the collisions test passes, mixed-count concentration bounds the observed count by $(1 + \sigma_k^2/4)\frac{s^2 e^{j\tau}}{2m'N}\mathbb{E}|\widetilde{V}_j|$. Outside $I_{j,k}$, either its mean is already below the predicted count or it is at most the observed count divided by $1 - \sigma_k^2/8$. Thus

$$\begin{aligned} Ne^{-j\tau}\frac{\mathbb{E}\left[\sum_{i\in\widetilde{V}_j} D(z_i)\right]}{\mathbb{E}|\widetilde{V}_j|} &= \frac{2m'N\,\mathbb{E}|\widetilde{C}_j|}{s^2 e^{j\tau}\mathbb{E}|\widetilde{V}_j|} \\ &\le \frac{1 + \sigma_k^2/4}{1 - \sigma_k^2/8} \le 1 + \sigma_k^2/2. \end{aligned}$$

Empirical mean error at least σ_k also implies conditional mean error at least $.99\sigma_k$, by Claim A.5. Rearranging (5) therefore gives

$$\frac{e^{j\tau}}{N}\frac{\mathbb{E}\left[\sum_{i\in\widetilde{V}_j} 1/D(z_i)\right]}{\mathbb{E}|\widetilde{V}_j|} \ge 2 + (.99\sigma_k)^2 - (1 + \sigma_k^2/2) \ge 1 + .48\sigma_k^2.$$

This proves (6), contradicting the upper bound $1 + .08\sigma_k^2$ from consistency. □

A.3.3 The Collisions Test is Complete

For honest tags, all probabilities in a bucket are within a factor e^τ of its tag $e^{j\tau}/N$. Thus the conditional collision mean agrees with the value predicted by the test, up to rounding. We control the sampling fluctuations simultaneously for every bucket and error level that the verifier may test.

Claim A.12. *If the prover is honest, then except with probability N^{-20}, every tested collision bucket passes, on the mixed-count event of Claim A.5.*

Proof. Honest positive tags in bucket j satisfy $e^{j\tau}/N \le D(x) < e^{(j+1)\tau}/N$. Consequently

$$1 \le \frac{2m'N\,\mathbb{E}|\widetilde{C}_j|}{s^2 e^{j\tau}\mathbb{E}|\widetilde{V}_j|} \le e^\tau.$$

We show simultaneous relative accuracy $\sigma_k^2/100$ for all eligible pairs (j,k). For each label, the Poisson tail bound in Claim 3.7 gives

$$\big||\{i : y_i = x\}| - sD(x)\big| \le O\big(\sqrt{sD(x)\log N} + \log N\big)$$

except with probability N^{-100}. Conditioning on presence costs at most $1/q_x \le 2N$. For $sD(x) \ge c \log N$, conditioning changes the pair mean by at most N^{-40}, since absence requires the count to be at most one. For smaller $sD(x)$, both pair means are $O(\log^2 N)$. Thus the centered pair counts can be truncated at

$$O\left(\log^2 N + \left(\frac{se^{j\tau}}{N}\right)^{3/2} \sqrt{\log N}\right),$$

with total failure and first two moment errors at most N^{-30}. The truncated variables remain independent; their total variance is at most $O(1 + se^{j\tau}/N)\mathbb{E}|\widetilde{C}_j| + N^{-30}$.

Bernstein's exponent at relative precision $\sigma_k^2/100$ is therefore, up to constants, at least

$$\min\left\{\frac{\sigma_k^4 \mathbb{E}|\widetilde{C}_j|}{1 + se^{j\tau}/N}, \quad \frac{\sigma_k^2 \mathbb{E}|\widetilde{C}_j|}{\log^2 N + (se^{j\tau}/N)^{3/2}\sqrt{\log N}}\right\}.$$

The first entry is $\Omega(\log^6 N)$ by the calculations above. The second is at least a constant times

$$\min\{\sigma^{-2} \log^{15} N, \sqrt{N}\, \sigma^{1/2} \sigma_k \log^{5/2} N\},$$

using $e^{j\tau}/N \le \max_x D(x) \le \sigma^2/\sqrt{N}$. Both exceed any fixed multiple of $\log N$ for sufficiently large N, since $\sigma_k \ge \sigma/8$ and $\sigma \ge N^{-1/10}$. A union bound over the $O(\sigma^{-2} \log^3 N)$ bucket–scale pairs proves the asserted simultaneous accuracy. Together with mixed-count accuracy and $\tau = o(\sigma_k^2)$, this places $|\widetilde{C}_j|$ within the collision tolerance $\sigma_k^2/5$ of $(s^2 e^{j\tau}/(2m'N))|\widetilde{V}_j|$. □

A.4 Putting the Puzzle Together

We combine the preceding claims to prove Proposition 5.1. For completeness, we show that an honest response passes both tests in every bucket that might be chosen. For soundness, we show that if the tags are far from correct, then some tier has large average error. A sufficiently large fraction of its buckets have error at one of the tested scales, so the verifier finds such a bucket with high probability. The single-bucket claims then imply that either its consistency test or its collisions test rejects.

Write $S_D = \{i : b_i = 0\}$ for the indices of the output D-samples. For a nonempty positive-tag bucket, let

$$\delta_j = \frac{1}{|V_j|} \sum_{i \in V_j} \left(1 - \min\left\{\frac{D(z_i)}{\pi_{z_i}}, \frac{\pi_{z_i}}{D(z_i)}\right\}\right).$$

Thus $|V_j|\delta_j$ is this bucket's contribution to the total output error. A zero tag contributes one per occurrence. We use the equivalent mental experiment of Claim A.9, with the conditioning specified in the preceding sections.

A.4.1 Protocol A.0.1 is Complete

Assume that the prover is honest. Every positive tag satisfies $\pi_x \le D(x) < e^{\tau}\pi_x$. The basic tests therefore pass with probability at least 0.999, by Claim A.1.

Next, fix a bucket j that may be tested. Honest tags give the two relations required by the advanced tests:

$$e^{-\tau} \le \frac{e^{j\tau}\mathbb{E}[|W_j|]}{\mathbb{E}[|V_j|]} \le 1, \qquad 1 \le \frac{\mathbb{E}[|C_j|]}{(s^2 e^{j\tau}/(2m'N))\mathbb{E}[|V_j|]} \le e^{\tau}.$$

The first relation says that the uniform and D-counts have the ratio predicted by the tag $e^{j\tau}/N$; the second says that the number of equal pairs has the predicted mean. Since $\tau = o(\sigma_k^2)$ at every tested scale, the rounding error is smaller than either test's tolerance. Claims A.5, A.6, and A.12 show that the actual counts are sufficiently close to these means, simultaneously for every bucket and scale that may be tested. Thus all chosen consistency and collisions tests pass, even though the tiers and buckets are chosen after the samples.

On the first-sample event of Claim A.8, $m' = \widetilde{\Theta}(N\|D\|_2^2\sigma^{-6})$, and the sample-size check gives $|S_D| = \Theta(m')$. The expected output size is $\widetilde{O}(\sqrt{N}\sigma^{-4}) = o(N^2)$. Labels with $0 < D(x) < N^{-20}$ have total mass at most N^{-19}, so the probability that any occurs in the output is at most N^{-17} for sufficiently large N. Otherwise every output tag is positive and satisfies

$$|D(z_i) - \pi_{z_i}| \le (e^{\tau} - 1)\pi_{z_i} = O(\sigma^2/\log N)\pi_{z_i}.$$

A union bound over these events makes the total failure probability less than 0.01. This proves both acceptance and the claimed quality and size of the honest output.

A.4.2 Protocol A.0.1 is Sound

Fix an arbitrary prover. We bound the probability that the verifier accepts while the total output error satisfies

$$\sum_{i \in S_D} \left(1 - \min\left\{\frac{D(z_i)}{\pi_{z_i}}, \frac{\pi_{z_i}}{D(z_i)}\right\}\right) \ge \sigma|S_D|. \tag{7}$$

Consider an execution satisfying (7). If a basic test rejects, there is nothing to prove. Otherwise, except for the failure probabilities already bounded, the basic probability bounds, the sample-size bounds, and the mixed-count concentration bounds hold. We use these implications without conditioning the private counts on these additional events.

Assign each positive-tag bucket with error greater than $2^{-\lceil \log_2(4/\sigma)\rceil}$ to its unique interval

$$\delta_j \in (\sigma_k, 2\sigma_k], \qquad 1 \le k \le \lceil \log_2(4/\sigma)\rceil.$$

We first bound the error in buckets that the verifier cannot reliably test. After removing these contributions, at least half the error in (7) will remain.

Buckets omitted by the protocol. Buckets below the smallest error scale contribute at most $\sigma|S_D|/4$. At each remaining scale k, the basic small-tag check bounds the number of D-samples with tag below $\sigma/(10^6 N\sigma_k \log N)$. Their total error is at most

$$\sum_k 2\sigma_k \left(\frac{10\sigma m'}{10^6\sigma_k \log N} + \log^2 N\right) < \frac{\sigma m'}{100}.$$

This also covers zero tags, whose error is one and which are included at $k = 1$.

Recall $\beta = \tau\sigma/(10^6 \log N)$. There are at most $21 \log N/\tau$ positive buckets, so those with $|V_j| < m'\beta$ contribute at most

$$\frac{21 \log N}{\tau} m'\beta < \frac{\sigma m'}{1000}.$$

Every other bucket belongs to a tier, by the sample-size check. If a bucket belongs to tier t but its error interval has $k > t+1$, then $\delta_j \le 2^{-(t+1)}$ and $|V_j| < m'\beta 2^{t+1}$. Its error is at most $m'\beta$; summing again costs less than $\sigma m'/1000$. This explains why testing only $k \le t+1$ suffices.

For every bucket still considered, $k \le t+1$ gives

$$|V_j| \ge m'\beta 2^t \ge \frac{m'\beta}{2\sigma_k}.$$

Claim A.5 therefore implies $\mathbb{E}[|V_j|] \ge m'\beta/(10\sigma_k)$, as required by the single-bucket claims. If the bucket is skipped because $|W_j| < c\sigma_k^{-4} \log^4 N$, the same concentration claim implies $\mathbb{E}[|W_j|] < 2c\sigma_k^{-4} \log^4 N$. Claim A.7 bounds the total error in all such buckets by $\sigma m'/100$, except with probability 0.001.

Buckets with misleading collision counts. Even for a bucket with large error, the observed pair count might be smaller than its conditional mean. Claim A.11 bounds this exception by the event $I_{j,k}$. We bound the total error of these buckets, rather than requiring this event to fail for every bucket.

Fix a first sample satisfying Claim A.3. Conditional on the sent message and the prover's randomness, the weights $\mathbb{E}[|V_j|]$ are fixed and $\Pr(I_{j,k}) \le O((\sigma/\sigma_k)^2 \log^{-6} N)$. Take $I_{j,k}$ to be empty whenever a hypothesis of that claim fails. Averaging also over the message and response gives

$$\mathbb{E}\left[\sum_{j,k} \frac{4\sigma_k \mathbb{E}[|V_j|]}{m'} \mathbf{1}_{I_{j,k}}\right] \le O(\log^{-6} N) \sum_k \frac{\sigma^2}{\sigma_k} = O(\sigma \log^{-6} N).$$

Here the inner means use the previous conditioning. Their sum over positive buckets has expectation at most m': each sent label contributes $m'D(x)/q_x$ and is sent with probability q_x. Markov's inequality bounds the displayed random sum by $\sigma/100$, except with probability 0.001. On the mixed-count event, a bucket at scale k contributes at most $2\sigma_k|V_j| \le 4\sigma_k \mathbb{E}[|V_j|]$. Hence deleting the buckets for which $I_{j,k}$ occurs removes at most $\sigma m'/100$ error. This argument does not require independence between $I_{j,k}$ and the mixed counts.

A tier with large error. Call a bucket *remaining* if its contribution has not been removed above. Since $m'/2 \le |S_D| \le 2m'$, the removed error is at most

$$\frac{\sigma|S_D|}{4} + \left(\frac{1}{100} + \frac{1}{1000} + \frac{1}{1000} + \frac{1}{100} + \frac{1}{100}\right)\sigma m' < \frac{\sigma|S_D|}{2}.$$

Thus (7) leaves at least $\sigma|S_D|/2$ error in the remaining buckets. In what follows, the size of a tier always means the total size of all its buckets, including any whose error was removed.

Claim A.13. *There is a tier t such that*

$$\sum_{\substack{j \in \text{tier}_t \\ j \text{ remaining}}} |V_j|\delta_j \ge \frac{\sigma}{2} \sum_{j \in \text{tier}_t} |V_j|. \tag{8}$$

Proof. If every tier violated this inequality, summing over the tiers would give remaining error less than $\sigma/2$ times their total size. Their total size is at most $|S_D|$, contradicting the preceding lower bound. □

Claim A.14. *In this tier there is an allowed $k \le t+1$ such that at least a fraction*

$$\frac{\sigma}{8\sigma_k\lceil\log_2(4/\sigma)\rceil}$$

of its buckets remain and have $\delta_j \in (\sigma_k, 2\sigma_k]$.

Proof. There are at most $\lceil\log_2(4/\sigma)\rceil$ error intervals. By (8), one interval contributes at least $\sigma/(2\lceil\log_2(4/\sigma)\rceil)$ times the full size of the tier. Each bucket in this interval contributes at most $2\sigma_k|V_j| < 4\sigma_k m'\beta 2^t$, whereas the full size of the tier is at least $m'\beta 2^t|\mathrm{tier}_t|$. Comparing these bounds gives the stated fraction. All remaining buckets have $k \le t+1$. □

Finding a bucket that rejects. For this tier and scale the verifier chooses $\lceil 10^6(\sigma_k/\sigma)\log^3 N\rceil$ buckets independently. Conditional on all samples and tags, these choices are fresh. The probability of missing every bucket in Claim A.14 is at most

$$\exp\left(-\frac{10^6\log^3 N}{8\lceil\log_2(4/\sigma)\rceil}\right) \le N^{-10}.$$

If the tier is smaller than the prescribed number of choices, all its buckets are tested and none is missed.

Finally, consider a remaining bucket found this way. Its error exceeds σ_k, it satisfies both guards and the conditional-mean size requirement, and $I_{j,k}$ does not occur. If both tests passed, Claims A.6 and A.11 would give conflicting bounds on the same reciprocal-probability sum:

$$\text{Consistency:}\quad \mathbb{E}\left[\sum_{i\in V_j}\frac{1}{D(z_i)}\right] \le (1+.08\sigma_k^2)Ne^{-j\tau}\mathbb{E}[|V_j|],$$

$$\text{Collisions:}\quad \mathbb{E}\left[\sum_{i\in V_j}\frac{1}{D(z_i)}\right] \ge (1+.48\sigma_k^2)Ne^{-j\tau}\mathbb{E}[|V_j|].$$

These are conditional expectations with the message and tags fixed, as in the single-bucket analysis. The lower bound is strictly larger than the upper bound. Hence at least one of the tests rejects this bucket, and the protocol rejects.

Adding the failure probabilities for the preceding estimates, the bounds on omitted error, and the bucket choices gives less than 0.01. Therefore the verifier accepts with output error at least σ with probability less than 0.01, proving soundness.

The error scales are essential here: the omitted error at scale σ_k is $O(\sigma^2/\sigma_k)$ up to logarithmic factors, and

$$\sum_{k\le\lceil\log_2(4/\sigma)\rceil}\frac{\sigma^2}{\sigma_k} = O(\sigma).$$

Keeping this dependence until after summing over the scales gives the required total error bound.

A.4.3 Sample and Communication Complexity

The independent first sample has

$$\mathbb{E}[m] = \frac{s^2}{2}\|D\|_2^2, \qquad \mathbb{E}[m'] = \widetilde{O}(N\|D\|_2^2\sigma^{-6}).$$

Claim A.8 also gives $m' = \widetilde{\Theta}(N\|D\|_2^2\sigma^{-6})$ with probability at least 0.999. The distinct collision labels in the second sample are at most its number of equal pairs, whose expectation is again $s^2\|D\|_2^2/2$. There are $2m'$ expected mixed draws. Sending their distinct union and returning one tag per label therefore costs

$$\widetilde{O}(N\|D\|_2^2\sigma^{-6})$$

bits in expectation and with probability at least 0.99; each label and grid tag uses $O(\log N)$ bits. Equal pairs can be counted from the multiplicities without enumerating them.

The basic tests use $\widetilde{O}(\sqrt{N}\sigma^{-4})$ fresh draws. The assumption of Proposition 5.1 gives

$$\|D\|_2^2 \le \sigma^2/\sqrt{N}.$$

Hence the mixed draws also cost $\widetilde{O}(\sqrt{N}\sigma^{-4})$. Sorting and counting multiplicities, checking the tags, and testing the chosen buckets take the same time up to logarithmic factors. This proves the claimed verifier bound and the σ^{-6} communication dependence.

Accuracy of the honest prover. The prover can instead round the empirical probabilities $\widehat{D}(x)$ from $\widetilde{O}(N\sigma^{-4})$ independent draws. This gives completeness through all tests, with the same average-error guarantee as soundness, although not the stronger relative-accuracy guarantee for every output tag.

Indeed, with sufficiently many logarithmic factors, Chernoff bounds give, simultaneously for every label and scale,

$$\max\{D(x), \pi_x\} \ge \frac{\sigma}{2\cdot 10^6 N\sigma_k \log N} \quad \Longrightarrow \quad |D(x) - \pi_x| \le \frac{\sigma_k^2}{1000}\pi_x.$$

To see that these estimates suffice, the concentration exponent at this probability scale is, up to constants and logarithmic factors,

$$(N\sigma^{-4})\frac{\sigma}{10^6 N\sigma_k \log N}\sigma_k^4 = \frac{(\sigma_k/\sigma)^3}{10^6 \log N};$$

here $\sigma_k > \sigma/8$, and grid rounding contributes $\tau = o(\sigma_k^2)$. An upward-tail bound excludes a tag above the displayed threshold when its true probability is less than half that threshold.

The basic probability checks therefore pass. Moreover, labels tagged below $\sigma/(10^6 N\sigma_k \log N)$ have total D-mass at most $2\sigma/(10^6\sigma_k \log N)$, so the small-tag checks pass by the same Poisson Chernoff bound as before; the sample-size check is unchanged. In every bucket that is actually tested, all its true probabilities differ from its tag by at most $\sigma_k^2/1000$ in relative terms. The conditional-mean relations in the completeness proof are consequently perturbed by only $O(\sigma_k^2/1000)$, within the slack of both tests. The same mixed-count and collision concentration bounds apply. Thus all tests pass with probability at least 0.99, after increasing the universal constants if necessary.

Finally, $\|\widehat{D} - D\|_1 = O(\sigma^2/\log N)$ with high probability; rounding changes this by $O(\tau) + N^{-19}$. Since

$$D(x)\,(1 - \min\{D(x)/\pi_x, \pi_x/D(x)\}) \le |D(x) - \pi_x|,$$

with the usual convention at zero, concentration of the independent mixed D-sample also gives average output error below σ. The constants can make this event and acceptance hold jointly with probability at least 0.99.

A.5 Extension to General Distributions

Remark A.15 (General distributions and parameter accounting)**.** *The heavy-light reduction used here is taken from Herman and Rothblum [HR23, Section 6.2]. We spell out the choice of accuracy parameters needed to preserve our bounds. Assume* $100N^{-1/10} \le \sigma \le 1/100$*. Using independent samples, identify a set* H *satisfying*

$$\{x : D(x) > \sigma^2/(2 \cdot 10^4\sqrt{N})\} \subseteq H \subseteq \{x : D(x) \ge \sigma^2/(8 \cdot 10^4\sqrt{N})\},$$

and learn the probabilities on H *to total* ℓ_1 *error* $\sigma/100$*. Frequency estimates and the usual distribution learner require* $\widetilde{O}(\sqrt{N}\sigma^{-4})$ *samples, since* $|H| = O(\sqrt{N}\sigma^{-2})$*. Let* $w \le 1$ *be an upper estimate of* $D(H^c)$ *with additive error at most* $\sigma/200$*. If* $w \le \sigma/100$*, ignore this light part. Otherwise* $D(H^c) \le w \le 2D(H^c)$*; run Proposition 5.1 on* $D_L = D \mid H^c$*, with*

$$\sigma_L = \frac{\sigma}{100\sqrt{w}}, \qquad \max_x D_L(x) \le \frac{\sigma_L^2}{\sqrt{N}}.$$

The permitted accuracy range holds, and the light contribution to the global error is at most $D(H^c)\sigma_L \le \sigma/100$*. Rejection sampling costs at most* $2/w$ *original draws per light draw. Consequently,*

$$\frac{\sqrt{N}\sigma_L^{-4}}{D(H^c)} = O(\sqrt{N}w\sigma^{-4}), \qquad N\|D_L\|_2^2\sigma_L^{-6} = O\left(Nw\sigma^{-6}\sum_{x\notin H} D(x)^2\right).$$

Thus verifier samples remain $\widetilde{O}(\sqrt{N}\sigma^{-4})$*, prover samples remain* $\widetilde{O}(N\sigma^{-4})$*, and communication remains* $\widetilde{O}(N\|D\|_2^2\sigma^{-6})$*. Include* H *and* w *in the first message so that the prover can sample from* D_L *and compute* σ_L*: their cost is also covered, since* $|H| = O(N\|D\|_2^2\sigma^{-4})$ *and* w *uses* $O(\log N)$ *bits. Combine the learned heavy part with the verified light histogram as in [HR23]; this adds* $O(\sigma)$ *error and no further interaction. These resource bounds hold in expectation and with constant high probability.*[11]

[11] Preprocessing estimates can fail with probability N^{-30}. Capping exceptional sampling loops at N^{10} draws makes their expected cost negligible and changes completeness by $o(1)$.

ECCC ISSN 1433-8092
https://eccc.weizmann.ac.il