Two Party Distribution Testing: Communication and Security

Alexandr Andoni
Columbia University, New York City, NY, USA

Tal Malkin
Columbia University, New York City, NY, USA

Negev Shekel Nosatzki
Columbia University, New York City, NY, USA

Abstract

We study the problem of discrete distribution testing in the *two-party setting*. For example, in the standard closeness testing problem, Alice and Bob each have t samples from, respectively, distributions a and b over $[n]$, and they need to test whether $a = b$ or a, b are ϵ-far (in the ℓ_1 distance). This is in contrast to the well-studied one-party case, where the tester has unrestricted access to samples of both distributions. Despite being a natural constraint in applications, the two-party setting has previously evaded attention.

We address two fundamental aspects of the two-party setting: 1) what is the communication complexity, and 2) can it be accomplished securely, without Alice and Bob learning extra information about each other's input. Besides closeness testing, we also study the independence testing problem, where Alice and Bob have t samples from distributions a and b respectively, which may be correlated; the question is whether a, b are independent or ϵ-far from being independent. Our contribution is three-fold: 1) We show how to gain **communication efficiency** given more samples, beyond the information-theoretic bound on t. The gain is polynomially better than what one would obtain via adapting one-party algorithms. 2) We prove *tightness* of our trade-off for the closeness testing, as well as that the independence testing requires tight $\Omega(\sqrt{m})$ communication for unbounded number of samples. These **lower bounds** are of independent interest as, to the best of our knowledge, these are the first 2-party communication lower bounds for testing problems, where the inputs are a set of *i.i.d. samples*. 3) We define the concept of **secure distribution testing**, and provide secure versions of the above protocols with an overhead that is only polynomial in the security parameter.

2012 ACM Subject Classification Mathematics of computing → Hypothesis testing and confidence interval computation

Keywords and phrases distribution testing, communication complexity, security

Digital Object Identifier 10.4230/LIPIcs.ICALP.2019.15

Category Track A: Algorithms, Complexity and Games

Related Version A full version of the paper is available at `https://arxiv.org/abs/1811.04065`.

Acknowledgements We thank Devanshi Nishit Vyas for her contribution to some of the initial work which led to this paper. We thank Clement Canonne for invaluable comments on an early draft of the manuscript. We thank Yuval Ishai for helpful discussions. Work supported in part by Simons Foundation (#491119), NSF grants CCF-1617955 and CCF-1740833.

1 Introduction

Distribution property testing is a sub-area of statistical hypothesis testing, which has enjoyed continuously growing interest in the theoretical computer science community, especially since the 2000 papers [36, 12]. One of the most basic problems is closeness testing, also known as the *homogeneity testing* ; see [37, 55, 58]. Here, given two distributions a, b and t

46th International Colloquium on Automata, Languages, and Programming (ICALP 2019).
Editors: Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi;
Article No. 15; pp. 15:1–15:16
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

samples from each of them, distinguish between the cases where $a = b$ versus a and b are ϵ-far, which usually means $\|a - b\|_1 > \epsilon$.[1] For this specific problem, the extensive research led to algorithms with optimal sample complexity [12, 59, 13, 24, 31, 29], including when the number of samples from the two distributions is unequal [5, 16, 31]. Further research directions of interest include obtaining instance-optimal algorithms, which depend on further properties of the distributions a, b [3, 4, 31], quantum algorithms [20], as well as algorithms whose output is differentially-private [30, 21, 6, 8]. An even larger body of work studied numerous other related problems; see, e.g., surveys [35, 22, 53, 52].

Focusing on testing two distributions, such as in the closeness problem, a very natural aspect has, surprisingly, evaded attention so far: such a task would often be run by two players, each with access to their own distribution. Specifically, Alice has samples from distribution a, Bob has samples from distribution b, and they need to jointly solve a distribution testing problem on (a, b). This setting models many of the envisioned usage scenarios of distribution testing, where different parties wish to jointly perform a statistical hypothesis testing task on their distributions. For example, [55] describes the scenario where two distinct sensors need to test whether they sample from the same distribution ("noise") or not.

This 2-party setting raises the following standard theoretical challenges, neither of which has been previously studied in the context of distribution testing:

- What is the *communication complexity* of the testing problem? In particular, can we do better than the straightforward approach, where Alice sends her samples to Bob who then runs an offline algorithm? Can we prove matching lower bounds?
  This aspect parallels the quest for low memory or communication usage for hypothesis testing on a *single distribution*, initiated in the statistics community [25, 41] and [7, 40, 10]. In fact, very recent, independent work has considered this aspect for binary sources [54, 38].
- Is it possible to design a distribution testing protocol that is *secure*, i.e., where Alice and Bob learn nothing about each other's samples, besides testing result? This question is highly relevant in today's push for doing statistics in a privacy-respecting manner.

1.1 Our Contributions

In this paper, we initiate the study of testing problems in the two-party model, and design protocols which are both communication-efficient and secure. We do so for two basic problems on pairs of distributions (i.e., where the two-party setting is natural): 1) closeness testing, and 2) independence testing.

Our main finding is that, once the number of samples exceeds the information-theoretic minimum, we can obtain protocols with polynomially smaller communication than the naïve adaptation of existing algorithms. We complement our protocols with lower bounds on the communication complexity of such problems that are near-optimal for closeness testing, as well as for independence for an unbounded number of samples. Our upper and lower bounds on communication are novel even without any security considerations.

To argue security, we also put forth a definition for secure distribution testing in the multi-party model. Our definition differs from the standard secure computation setting due to two unique features of the considered setting. First, this is "testing" (a promise problem) and not "computing"; second, the function of interest is defined with respect to distributions, while the parties' inputs are samples. These features do not come into play if the distributions satisfy the promise (e.g., they are either identical or ϵ-far), in which case

[1] This is equivalent to saying that the total variation distance is more than $\epsilon/2$.

the security guarantee matches the standard cryptographic one (no information is leaked beyond the output). However, when the promise is not satisfied, we need to allow for some information on the parties' samples to be leaked by the protocol. Our definition permits leakage of at most one bit in this case, and leaks nothing when the promise is satisfied.

▶ **Definition 1** (Security Definition). *Let D be the set of input distributions over $\bigtimes_{i=1}^{d}[n_i]$, and let $g : D \to \{0,1\}$ be a partial boolean function, defined on $P \subseteq D$.*

*Let π be a d-party protocol, and let k be a security parameter. We say that π is a t-*sample secure distribution testing protocol *(for the testing task defined by g), if there exists a boolean function $f : \{\bigtimes_{i=1}^{d}[n_i]\}^t \to \{0,1\}$ such that the following holds:*

Correctness: *for any $p \in P$, $\Pr_{\zeta_1 \ldots \zeta_t \sim_{i.i.d}\, p}[f(\zeta) = g(p)] = 1 - neg(k)$*

Security: *For any $\zeta \in \{\bigtimes_{i=1}^{d}[n_i]\}^t$, if we give each player $i \in [d]$ the input $\left(1^k, \zeta_1(i), \ldots, \zeta_t(i)\right)$, then protocol π is a secure computation of the function $f(\zeta)$.*

We provide a detailed discussion of the above definition in the full version of this paper.

Closeness Testing. In the *2-party closeness testing* problem $2\text{PCT}_{n,t,\epsilon}$, Alice and Bob each have access to t samples from some distributions respectively a, b over alphabet $[n]$. Their goal is to distinguish between $a = b$ and $\|a - b\|_1 \geq \epsilon$ with probability $\geq 2/3$.

We first give a non-secure near-optimal communication protocol, and then show how to make it secure with only a small overhead (polynomial in the security parameter). Our secure version is based on the existence of a PRG that stretches from polylog(m) bits to m bits, and of an OT protocol with polylog communication. Overall, we prove the following.

▶ **Theorem** (Closeness, Secure). *Fix a security parameter $k > 1$. Fix $n > 1$ and $\epsilon \in (0, 2)$, and let t be such that $t \geq C \cdot k \cdot \max\left(n^{2/3} \cdot \epsilon^{-4/3}, \sqrt{n} \cdot \epsilon^{-2}\right)$ for some (universal) constant $C > 0$. Then, assuming PRG and OT as above, there exists a* secure distribution testing protocol *for $2\text{PCT}_{n,t,\epsilon}$ which uses $\tilde{O}_k\left(\frac{n^2}{t^2\epsilon^4} + 1\right)$ communication.*

To contrast the communication bounds of our protocol to the classic 1-party setting, consider what happens in the extreme settings of the parameters s, t, for a fixed ϵ. When $t \approx \Theta(n^{2/3})$, the communication is $\tilde{O}(n^{2/3})$ as well, i.e., Alice may as well just send all the samples over to Bob. However the communication decreases as the players have more samples. This may not be surprising given the testing results with unequal number of samples [16, 31]: indeed, Alice can send $\approx \max\{n/\sqrt{t}, \sqrt{n}\}$ samples to Bob, and Bob can run the tester. In contrast, our protocol obtains a *polynomially smaller* complexity, $\approx n^2/t^2$, whenever $t \gg n^{2/3}$. Intuitively, considering the extreme of $t \gg n$, we can obtain near-constant communication: with so many samples, we can *learn* the distribution, and then use *sketching* tools [9, 42].

We prove a *near-tight* lower bound on the above trade-off (even without security considerations) in Section 4. We note that our lower bound differs from the common communication complexity lower bounds as the players' inputs are *i.i.d. samples* and not worst-case.

▶ **Theorem** (Closeness lower bound). *Any two-way communication protocol for $2\text{PCT}_{n,t,1/2}$ requires $\tilde{\Omega}\left(n^2/t^2\right)$ communication.*

Independence Testing. Our second problem is the independence testing problem in the 2-party model, denoted $2\text{PIT}_{n,m,t,\epsilon}$. Let $p = (a, b)$ be some joint distribution over $[n] \times [m]$, where $n \geq m$, and for $i \in [t]$, let ζ_i be a sample drawn from p. Now we provide Alice with the first coordinates of ζ_i's and Bob with the second coordinates. Alice and Bob's goal is to test whether distributions a and b are independent (p is a product distribution) or p is ϵ-far from any product distribution. We prove the following:

▶ **Theorem** (Independence, Secure). *Fix a security parameter $k > 1$. Fix $\epsilon \in (0,2)$, $1 \leq m \leq n$, and let t be such that $t \geq C \cdot k \cdot \left(n^{2/3}m^{1/3}\epsilon^{-4/3} + \sqrt{nm}/\epsilon^2\right)$, for some (universal) constant C, and assuming OT, there is a secure distribution testing protocol for $2\mathrm{PIT}_{n,m,t,\epsilon}$ using $\tilde{O}_k\left(\frac{n^2 \cdot m}{t^2\epsilon^4} + \frac{n \cdot m}{t\epsilon^4} + \frac{\sqrt{m}}{\epsilon^3}\right)$ bits of communication.*

We note that the lower bound on t from the above theorem is necessary as it is the information-theoretic bound, as proven in [31]. An important qualitative aspect of the communication complexity for 2PIT is that, when the number of samples $t \to \infty$, the protocol uses $\tilde{\Theta}_\epsilon(\sqrt{m})$ bits of communication. This is in contrast to 2PCT, where the communication becomes $\tilde{O}(1)$ for $t \to \infty$. Indeed, we show that $\Omega(\sqrt{m})$ is necessary for one-way protocols for 2PIT. Since our protocol can easily be converted to a one-way (non-secure) protocol, this lower bound is tight for one-way protocols. We conjecture that the bound from the above theorem is near-tight in n, m, t for two-way communication protocols, even without security.

▶ **Theorem** (Independence lower bound). *For $n, t \in \mathbf{N}$, any one-way protocol for $2\mathrm{PIT}_{n,n,t,1}$ requires $\Omega(\sqrt{n})$ bits of communication.*

1.2 Related work

Our work bridges three separate areas and models: distribution testing, streaming/sketching, and secure computation. There's a large body of work in each of these areas. We mention work most relevant to us.

Testing and learning with memory or space constraints. Two-party communication model is tightly connected to the streaming and distributed models, which have received lots of focus in the context of testing and learning questions. As early as in 1960s, [25, 41] considered the hypothesis testing (of one distribution) in the streaming model, where samples are streamed over while keeping small extra space. More recently, much attention has been drawn to streaming (memory) lower bounds for learning problems, such as parity learning [50, 51, 46, 48, 33]. Another direction was to consider *stochastic streaming* problems [26], where the input is generated from a distribution. All these results apply to samples from *one distribution*, and show a (tight) trade-off between number of samples and space complexity.

Another recent avenue is to study such problems in the distributed model, where there are many symmetric players, each with a number of i.i.d. samples from the same distribution. For learning problems (e.g., parameter or density estimation), see, e.g., [19, 28, 27]. We note that since learning is a much harder problem, typically proving lower bounds is easier (e.g., as shown in [28], merely communicating the output requires $\Omega(n)$ communication). In contrast, for testing problems, the output is just one bit. For testing problems (of one distribution), see also the recent (independent) manuscript [2].

None of the lower bounds from the above papers are relevant here as they become vacuous for a 2-party setting. Indeed, when two players have two sets of samples from the *same* distribution, then purely doubling the sample set of a player trivializes the question (she can solve it without communication).

Finally, a very recent, independent works of [54, 38] consider a problem very similar to 2PIT: estimating the correlation of two binary sources ($n = m = 2$) in the two-party model.

Secure approximations. Our results on secure distribution testing can also be seen in the context of the area of secure computation of approximations. This is a framework introduced by [32], allowing to combine the benefits of approximation algorithms and secure computation.

This was considered in different settings [32, 39, 14, 43, 15, 44, 45], but the most relevant to us is private approximation of distance between two input vectors. In particular, for ℓ_2 distance, Alice and Bob each have a vector $a, b \in \mathbb{R}^n$ and want to estimate $\|a-b\|_2$, without revealing any information that does not follow from the ℓ_2 distance itself. For this problem, [43] show that secure protocols are possible with only poly-logarithmic communication complexity.We use some of their techniques in our secure protocols.

Approximation and testing have a similar flavor in that they both trade accuracy for efficiency, in different ways. The security goals are also similar (prevent leakage beyond the intended output). One important difference is that the intended output in secure testing is just the single bit of whether or not the test passed. Thus, for example, when approximating a distance function, even a secure protocol can leak any information that follows from the distance. In contrast, when testing for closeness, if the inputs are either identical or far, the protocol may only reveal this fact, but no other information about what the distance is.

Security and privacy of testing. While we are not aware of any work on secure testing, several recent papers address *differentially-private distribution testing* [30, 21, 6, 8]. Here the privacy guarantee relates to the value of the output after the computation is concluded, requiring it to be differentially-private with respect to the inputs. Our notion of security for distribution testing is different, in the same way that secure computation is different from differentially private computation. While differential privacy (DP) is concerned with what the intended output may leak about the inputs (even if the input came from a single party or the computation is done by a trusted curator), secure 2-party computation is concerned with how to compute an intended output without leaking any information beyond the output itself. The difference in goals is also reflected in the privacy guarantees, which are typically statistical in nature (for DP testing) and provide a non-negligible adversarial advantage. Secure testing protocols rely on cryptographic assumptions and provide negligible advantage.

Even more recently, a stringent model of Locally Differentially Private Testing was proposed [56, 1]. This model provides a stronger notion of differential privacy, where users send noisy samples to an untrusted curator, and the goal is to allow the curator to test the distribution of user inputs (for some property) without learning "too much" about the individual samples. For LDP, the main goal is to optimize the sample complexity as a function of the privacy guarantees. While this notion of privacy also incorporates some privacy of the individual inputs, it is much closer to DP than to our security notion. In addition, both DP and LDP do not provide sub-linear communication (in the sample size, as we achieve here). In fact, their goal is to allow $O(1)$ communication per sample, with minimal sample overhead. In contrast, our protocols provide security "nearly for free" while allowing for faster communication with more samples. Finally, in the case of independence testing, our work assumes samples are *distributed* between the parties who need to test the joint distribution, while in the above work, each data point contains full sample information.

1.3 Our Techniques

We now outline the techniques used to establish our main results. Since our overall contribution is painting a big picture of the 2-party complexity of distribution testing problems, we appeal to a number of diverse tools. First, we design communication-efficient protocols. Second, we argue optimality of our protocols by proving communication complexity lower bounds on the considered problems, which are near-tight in some of the parameter regimes. Third, we show how to transform our protocols into secure protocols, under standard cryptographic assumptions, without further loss in efficiency. All three of these contributions are independently first-of-a-kind, to the best of our knowledge.

Communication-efficient protocols. We start by reducing the testing problem under the ℓ_1 distance to the same problem under the ℓ_2 distance, using now-standard methods of [31, 24]. Here, our main challenge is actually testing under the ℓ_2 distance.

Closeness testing (2pCT) is technically the simpler problem, but it already illustrates some phenomena, how to leverage a larger number of samples to improve communication. To estimate the ℓ_2 distance between the 2 unknown distributions, we compute the ℓ_2 distance approximation between the given *samples* of these distributions. In order to approximate the latter in the 2-party setting, we use the ℓ_2 *sketching* tools [9]. The crux is to show that we can tolerate a cruder ℓ_2 approximation if we are given a larger sample size. Since the complexity of $(1+\alpha)$-approximating the ℓ_2 distance is $\Theta(1/\alpha^2)$, we obtain an improvement in communication that is quadratic in the number of samples.

Independence Testing (2pIT) is more challenging since any distance approximation would need to be established based on the distribution(s) implicitly defined via the *joint samples*, split between Alice and Bob, and hence our approximation techniques above are not sufficient. Instead, we develop a reduction from a large, $[n] \times [m]$, alphabet problem, to a smaller alphabet problem, which can be efficiently solved by communicating fewer samples. This is accomplished by sampling a rectangle of the joint alphabet, and showing that such a process, when combined with the *split-set* technique from [31], generates *sub-distributions* (defined later) which satisfy some nice properties. We then show one can test the original distribution $p = (a, b)$ over a "large" domain of size $[n] \times [m]$ for independence by distinguishing closeness of 2 simulated distributions $\hat{p}, \hat{q}$, defined on a smaller domain of size $[l] \times [m]$, where $l = \tilde{\Theta}_\epsilon(n^3 m/t^3 + n^2 m/t^2 + 1)$. We show it is possible for Alice and Bob to simulate joint samples from $\hat{p}$ and $\hat{q}$ using $O(1)$ communication per sample, after they have down-sampled letters from one of the marginals.

The trade-off on communication–vs–samples emerges from two compounding effects: 1) balancing the size of the target rectangle with the expected number of available samples over such rectangle; and 2) the additional advantage from a tighter bound on the ℓ_2 norm of $\hat{q}$. Each of the above independently generates linear improvement in communication with more samples. The latter advantage, however, is helpful only while $t = O(n)$, and therefore we benefit from quadratic improvement in that regime, and linear improvement thereafter.

Lower bounds on communication. We note that the lower bounds on communication of testing problems present a particular technical challenge: for testing problems, the inputs are i.i.d. samples from some distributions. This is more akin to the average-case complexity setup, as opposed to "worst case" complexity as is standard for communication lower bounds.

We manage to prove such testing lower bounds for the *Closeness Testing* problem (2pCT). While our lower bound is, at its core, a reduction from some "hard 2-party communication problem", our main contribution is dealing with the above challenge. One may observe that a "hard 2-party communication problem" is hard under a certain input distribution (by Yao's minimax theorem), and hence a reduction algorithm would also produce a hard distribution on the inputs to our problem. However, a priori, it is hard to ensure that the resulting input distribution resembles anything like a set of samples from distributions a, b. For example, the inputs may have statistical quirks that actually depend on whether it is a "close" or "ϵ-far" instance, which a reduction is not able to generate without knowing the output.

At a high level, the role of the "hard problem" is played by a variant of the well-known two-way Gap Hamming Distance (GHD) problem [23, 57]. The known GHD lower-bound variants are insufficient for us precisely because of the above challenge – we need a better control over the actual hard distribution.Therefore, we study the following *Exact GHD* variant: given $x, y \in \{0,1\}^n$, with $\|x\|_1 = \|y\|_1 = n/2$, distinguish between $\|x-y\|_1 = n/2$ versus $\|x-y\|_1 \in [n/2+\beta, n/2+2\beta]$. We show there exists some $\beta \in [\Omega(\sqrt{n}), O(\sqrt{n \log n})]$ for which communication complexity must be $\tilde{\Omega}(n)$, by adapting the proof of [57].

Using one instance of *Exact GHD*, our reduction performs a careful embedding of this hard instance into the samples from distributions a, b, while patching the set of samples to look like i.i.d. samples from the two distributions. While we don't manage to get the output of the reduction to look precisely like i.i.d. samples from a, b, our reduction produces two sets of size $\mathrm{Poi}(t)$ whose distribution is within a small statistical distance from the distribution of two set of samples that would be drawn from two distributions a and b which are either "equal" (when $\|x-y\|_1 = n/2$) or "far" (when $\|x-y\|_1 \in [n/2+\beta, n/2+2\beta]$).

Note that our proof recovers the standard lower bound of $\tilde{\Omega}(n^{2/3})$ on samples necessary to solve closeness testing (in the vanilla setting), albeit not a tight bound [24, 31].

For *Independence testing* (2pIT), we focus on the lower bound for unbounded number of samples. We argue such a hardness result under one-way communication only. Our $\Omega(\sqrt{m})$ lower bound uses the Boolean Hidden Hypermatching (BHH) problem [60]. We conjecture our entire trade-off for the Independence problem is tight. The proof of this conjecture would have to overcome the challenge of lower bounds for statistical inputs.

Securing the communication protocols. Once low-communication insecure protocols have been designed, one may try to convert the protocols to secure ones using generic cryptographic techniques. The latter includes various techniques for secure computation ([61] and followup work), fully homomorphic encryption ([34] and followup work), or homomorphic secret sharing ([17, 18]). However, a naïve application of such techniques will blow up the communication to be at least linear in the input size, possibly requiring strong assumptions, a high computation complexity, or not being applicable to arbitrary computations. The constraint of low-overhead, among other considerations, requires design of custom protocols.

Our starting point is a technique that falls into the latter category: secure circuits with ROM [49], a technique that can transform an insecure 2-party protocol to a secure one with a minimal blow-up in communication, and uses a weak assumption only (OT). In order to obtain an efficient protocol, however, it only applies to computations expressible via a very small circuit, whose size is proportional to the target communication, with access to a larger read-only-memory (ROM) table. Thus, the main challenge becomes to design two-party testing protocols that fit this required format.

For *Closeness Testing* (2pCT), we begin with our low-communication non-secure protocol, and adapt it to be secure by designing a small circuit. One of the main difficulties in designing such a circuit is that, in the ℓ_1- to ℓ_2-testing reduction, Alice and Bob need to agree on an alphabet, which depends on their inputs, without compromising the inputs themselves. To bypass this, and other issues, we allow Alice and Bob to perform some off-line work and prepare some polynomial-size inputs (in ROM). First, we devise a method for Alice and Bob to generate a combined split set S (discussed later) by having each of Alice and Bob contribute sampled letters to S. Second, we securely estimate the ℓ_2 distance of Alice and Bob's original, un-splitted samples using techniques from [43]. Finally, we adjust our approximation by accounting for a few letters which differ from the original alphabet or which cannot be estimated efficiently. The main focus of our analysis goes into proving our construction adds only poly-logarithmic communication over the insecure protocol.

We describe the details of our secure protocols, as well as our results on independence testing, in the full version of this paper.

2 Preliminaries

Notation. Throughout this paper we denote distributions in small letters, and distribution samples in capital letters. Unless stated otherwise, any distribution is on alphabet $[n]$, and domain elements of $[n]$ are addressed as letters.

We also denote any multiplicative error arising from approximation as $1+\alpha$, and any error or distance of/between distributions as ϵ. Unless stated otherwise, distance and norms are referring to the Euclidean distance and ℓ_2 norms.

We denote Poisson Random variables with parameter $\lambda > 0$ as $\mathrm{Poi}(\lambda)$.

Split Distributions. We use the concept of split distributions from [31] to essentially reduce testing under ℓ_1 distance to testing under ℓ_2 distance.

▶ **Definition 2.** *Given a probability distribution p on $[n]$ and a multiset S of items from $[n]$, define the* split distribution *p_S on $[n+|S|]$ as follows. For $i \in [n]$, let a_i be equal to 1 plus the number of occurrences of i in S; note that $\sum_{i=1}^{n} a_i = n+|S|$. We associate the elements of $[n+|S|]$ to elements of the set $E = \{(i,j) : i \in [n], 1 \le j \le a_i\}$. Now the distribution p_S has support E and a random draw (i,j) from p_S is sampled by picking i randomly from p and j uniformly at random from $[a_i]$.*

Recall from [31] that split distributions are used to upper bound the ℓ_2 norm of an underlying distribution while maintaining its ℓ_1 distance to other distributions:

▶ **Fact 3** ([31])**.** *Let p and q be probability distributions on $[n]$, and S a given multiset of $[n]$. Then, (1) We can simulate a sample from p_S or q_S by taking a single sample from p or q, respectively; and (2) $\|p_S - q_S\|_1 = \|p-q\|_1$.*

▶ **Lemma 4** ([31])**.** *Let p be a distribution on $[n]$. Then: (i) For any multisets $S \subseteq S'$ of $[n]$, $\|p_{S'}\|_2 \le \|p_S\|_2$, and (ii) If S is obtained by taking $\mathrm{Poi}(m)$ samples from p, then $\mathbb{E}[\|p_S\|_2^2] \le 1/m$.*

3 Closeness Testing: Communication-Efficient Protocol

In this section we consider the closeness testing problem 2pCT, focusing on the 2-party communication complexity only. In the full version of this paper, we show how to modify the protocol to make it secure.

As mentioned in the introduction, one way to obtain a protocol is to use unequal-size closeness testing, where Alice has s samples and Bob has t samples: Alice just sends her s samples to Bob, and Bob invokes a standard algorithm for closeness testing. Using the optimal bounds from, say, [31], we get the following trade-off for fixed ϵ: $s = \tilde{O}(n/\sqrt{t})$, with the condition that $s, t \ge \sqrt{n}$. Here we obtain a *polynomially smaller* communication complexity, $s = \tilde{O}_\epsilon(n^2/t^2)$, whenever t is above the information-theoretic minimum on the number of samples. In Section 4, we show a nearly-matching lower bound.

3.1 Tool: approximation via occurrence vectors

Our protocol uses the framework introduced in [31], allowing us to focus on the ℓ_2 testing problem. For ℓ_2 testing, we show that we can approximate the ℓ_2 distance of two discrete distributions p, q by approximating the ℓ_2 distance of their respective sample occurrence vectors, defined as follows.

▶ **Definition 5.** *Given t samples of distribution p over $[n]$, we define the* occurrence vector *$X \in [t]^n$ such that X_i represent the count of occurrences of element $i \in [n]$ in the sample set.*

The following lemma bounds how well we need to estimate the ℓ_2 distance between occurrence vectors to distinguish between $p = q$ vs. $\|p - q\|_1 \geq \epsilon$. It shows that the more samples we have, the less accurate the ℓ_2 estimation needs to be. Using the framework from [31], for now it is enough to assume that the ℓ_2 norm of both p and q is bounded by $U < 1$.

▶ **Lemma 6.** *Let p, q be distributions over $[n]$ with $\|p\|_2, \|q\|_2 \leq U$ for some $U < 1$. There exists $t = O(U \cdot n \cdot \epsilon^{-2})$, and $\alpha = \Omega(U)$, such that given Δ which is $(1 \pm \alpha)$-factor approximation of $\|X - Y\|_2^2$ where X, Y represent the occurrence vectors of t samples drawn from p, q respectively, then, using Δ, it is possible to distinguish whether $p = q$ versus $\|p - q\|_1 > \epsilon$ with 0.8 probability.*

The actual distinguishing algorithm is simple: for fixed $\alpha = \Omega(U)$, we merely compare Δ to some fixed threshold τ (fixed in the proof below). The intuition is that for a given number of samples, we have some gap between the range of possible distances $\|X - Y\|_2^2$ for each of the cases. If the number of samples is close to the information-theoretic minimum [24], then the gap is minimal and we need to calculate almost exactly the distance, hence estimating the distance between occurrence vectors doesn't help. However, as the number of samples t increases, so does the gap between the ranges, allowing for a looser distance approximation.

Proof of Lemma 6. Given $t = O(U/\epsilon'^2)$ samples from each p, q, according to [24, Proposition 3.1], the estimator $Z = \sqrt{\sum_i (X_i - Y_i)^2 - X_i - Y_i}/t$ is a $\max\{\epsilon', \|p - q\|_2/8\}$ additive approximation of $\|p - q\|_2$ with 0.9 probability. Setting $\epsilon' = \epsilon/(8\sqrt{n})$, we obtain: (1) $\|p - q\|_2 = 0 \Rightarrow \|X - Y\|_2^2 \leq \frac{\epsilon^2 t^2}{4n} + 2t$; and (2) $\|p - q\|_2 > \epsilon/\sqrt{n} \Rightarrow \|X - Y\|_2^2 \geq \frac{3\epsilon^2 t^2}{4n} + 2t$ Now, suppose Δ is such that $\frac{\Delta}{\|X - Y\|_2^2} \in (1 - \alpha, 1 + \alpha)$. If $\|p - q\|_1 = 0$, then $\|p - q\|_2 = 0$ and hence $\Delta \leq (1 + \alpha)(\frac{\epsilon^2 t^2}{4n} + 2t) \leq t(\frac{\epsilon^2 t}{4n} + 2 + 2\alpha + \frac{\alpha\epsilon^2 t}{4n})$. On the other hand, if $\|p - q\|_1 > \epsilon$, then $\|p - q\|_2 > \epsilon/\sqrt{n}$ and hence $\Delta \geq (1 - \alpha)(\frac{3\epsilon^2 t^2}{4n} + 2t) \geq t(\frac{3\epsilon^2 t}{4n} + 2 - 2\alpha - \frac{3\alpha\epsilon^2 t}{4n})$.

We distinguish the two cases, by comparing Δ to $\tau = \frac{\epsilon^2 t^2}{2n} + 2t$: namely $p = q$ iff $\Delta \leq \tau$. Indeed we argue that $t(\frac{3\epsilon^2 t}{4n} + 2 - 2\alpha - \frac{3\alpha\epsilon^2 t}{4n}) - \tau \geq \tau - t(\frac{\epsilon^2 t}{4n} + 2 + 2\alpha + \frac{\alpha\epsilon^2 t}{4n})$. We have that $\frac{\epsilon^2 t}{4n} - 2\alpha - \frac{3\alpha\epsilon^2 t}{4n} \geq 0$, or $\alpha \leq \frac{\epsilon^2 t}{4n \cdot (2 + 3\epsilon^2 t/4n)}$. Since $t = O(Un\epsilon^{-2})$, the conclusion follows. ◀

3.2 Communication vs number of samples

We now provide a (non-secure) protocol for 2PCT with a trade-off between communication and number of samples.

▶ **Theorem 7** (Closeness, insecure)**.** *Fix $n > 1$ and $\epsilon \leq 2$. There exists some constant $C > 0$ such that for all $t \geq C \cdot \max\left(n^{2/3} \cdot \epsilon^{-4/3}, \sqrt{n} \cdot \epsilon^{-2}\right)$, the problem $2\mathrm{PCT}_{n,t,\epsilon}$ can be solved using $\tilde{O}\left(\frac{n^2}{t^2\epsilon^4} + 1\right)$ bits of communication.*

The protocol uses Lemma 6 as the main algorithmic tool and proceeds as follows. Bob generates multi-set S using samples from b and sends S to Alice. Then, Alice and Bob each simulate samples from a_S and b_S respectively, and together approximate the ℓ_2 difference of the resulting occurrence vectors using sketching methods [9].

Proof of Theorem 7. We note that, according to Lemma 4, $\mathbb{E}[\|b_S\|_2^2] \leq t^2\epsilon^4/n^2$ and hence $\|b_S\|_2^2 = O(t^2\epsilon^4/n^2)$ with at least 90% probability. Furthermore, since $t = \Omega(\sqrt{n}/\epsilon^2)$, we have that $|S| = O(n)$ with high probability. From now on, we condition on these two events.

Non-Secure 2pCT(a, b, t)
Alice's input: t samples from a
Bob's input: t samples from b
1. Fix $\alpha = \Omega(t \cdot \epsilon^2/n)$.
2. Bob generates multi-set S using $\mathrm{Poi}(\frac{n^2}{t^2\epsilon^4})$ samples from b.
3. Bob sends S to Alice.
4. Alice and Bob recast their samples as being from distributions a_S, b_S (see Def. 2), and set A_S, B_S to be the respective occurrence vectors.
5. Alice and Bob each estimate $\|a_S\|_2$ and $\|b_S\|_2$ up to factor 2; if the two estimates are not within factor 4, output "ϵ-FAR";
6. Alice and Bob approximate $\Delta = \|A_S - B_S\|_2^2$ up to $(1 + \alpha)$ factor, using, say, [9].
7. If Δ is less than $\tau = \frac{\epsilon^2 t^2}{2n} + 2t$ output "SAME", and, otherwise, output "ϵ-FAR".

If $\|a_S\|_2 \neq \Theta(\|b_S\|_2)$ then distributions are different and we output "ϵ-far" is step 5. Otherwise, we have that $\|a_S\|_2^2 = O(\|b_S\|_2^2) = O(t^2\epsilon^4/n^2)$. Hence we can use Lemma 6, where $U = O(t\epsilon^2/n)$ and $\alpha = \Omega(U)$, to claim the correctness of the protocol.

In terms of communication complexity, first, communicating S takes $|S| \log n = \tilde{O}(n^2/t^2\epsilon^4)$ bits with high probability. Second, estimating Δ up to approximation $1+\alpha$ takes $\tilde{O}(1/\alpha^2) = \tilde{O}(n^2/t^2\epsilon^4)$ bits, using standard ℓ_2 estimation algorithms [9, 47]. ◀

▶ Remark 8. Another application of this protocol is that it can be simulated by a *single party* to obtain a space bounded *streaming algorithm* with the same space/sample trade-offs. While we are not formalizing this argument in this paper, this can essentially be done by storing S and sketching $\|A_S - B_S\|_2^2$.

4 Closeness Testing: Communication Lower Bounds

We now prove that the protocol for 2pCT from Section 3 is near-tight, showing the following:

▶ **Theorem 9.** *Let a, b be some distributions over alphabet $[n]$, where Alice and Bob each receive $\mathrm{Poi}(t)$ samples from a, b respectively, for $t \leq n/\log^c n$ for some large enough $c > 1$. Then any (two-way) communication protocol Π that distinguishes between $a = b$ and $\|a - b\|_1 \geq 1/2$ requires $s = \tilde{\Omega}(n^2/t^2)$ communication.*

Intuitively, our proof formalizes the concept that in testing distributions for closeness, "collisions is all that matters", *even in the communication model*. This is similar to the intuition from the "canonical tester" from [59], which shows a similar principle when all the samples are accessible. Our result can be seen to extending it to saying that the canonical tester is still the best even if we have more-than-strictly-necessary number of samples that we could potentially compress in a communication protocol.

To prove the theorem, we rely on the following communication complexity lower bound, which is a variant of the Gap-Hamming-Distance (GHD) lower bound [23, 57]. Somewhat surprisingly, there does not seem to be a proof in the *one-way communication* model, which would be simpler than the two-way proof from the lemma below.

▶ **Lemma 10.** *Let $n \geq 1$ be even. There exists some $\beta = \beta(n) \in [\Theta(\sqrt{n}), \Theta(\sqrt{n \log n})]$, satisfying the following. Consider a two-way communication protocol $\mathcal{A}$ that, with probability at least 0.9, for $x, y \in \{0,1\}^n$ with $\|x\|_1 = \|y\|_1 = n/2$, can distinguish between the case when $\|x - y\|_1 = n/2$ versus $\|x - y\|_1 - n/2 \in [\beta, 2\beta]$. Then $\mathcal{A}$ must exchange at least $\Omega(\frac{n}{\log n \cdot \log\log n \cdot \log\log\log n})$ bits of communication.*

The proof of this lemma is presented in the full version of this paper.

Proof of Theorem 9. The idea is to reduce an instance of the GHD problem from Lemma 10 to an instance of closeness testing by carefully molding the input (x, y) into a couple of related occurrence vectors $(A, B) \in \mathbf{N}^n \times \mathbf{N}^n$ (recall that an occurrence vector precisely describes a set of samples).

Fix input vectors x, y, of length $m = \frac{n^2}{t^2 \log^3 n}$, to the above GHD problem. Let $\Delta = \beta(m) = \Omega(\sqrt{m})$, and $\delta = \frac{1}{2}(\|x - y\|_1 - m/2) \in \{0\} \cup [\Delta/2, \Delta]$. The case of $\delta = 0$ will correspond to "same" case (i.e. $a = b$), and $\delta \in [\Delta/2, \Delta]$ – to "far" case (i.e. $\|a - b\|_1 \in [1/2, 1]$).

Fix $d = n/10$ and $l = C \cdot t \cdot \log n$ (where C is some constant that we shall fix later), which have the following meaning: each distribution a, b has half mass over $[d]$ items uniformly (called *dense items*), and the other half on $[l]$ items uniformly (called *large items*). When $a = b$, these are the same items, and when $a \neq b$, the large items are the same while the dense items have supports with a large difference. In particular, the dense items are supported on sets S_A, S_B respectively, with $|S_A| = |S_B| = d$, and $S_A \cap S_B = d \cdot \frac{\Delta - \delta}{\Delta}$; we hence also have that $|S_A \setminus S_B| = d \cdot \frac{\delta}{\Delta}$.

Now for $i \geq 0$, let $D(i) = \Pr[\mathrm{Poi}(t/2d) = i]$, i.e., probability a dense number is sampled i times (when sampling $\mathrm{Poi}(t)$ items from one of the distributions). For simplicity, we write $D(i, j) = D(i) \cdot D(j)$. Similarly we define $L(i) = \Pr[\mathrm{Poi}(t/2l) = i]$ and $L(i, j) = L(i) \cdot L(j)$. We also set $k = \Theta(\log n)$, which should be thought of as an upper bound on the count of any fixed item (whp). The algorithm constructs the occurrence vectors A, B iteratively coordinate by coordinate. Let $m_c = m/4 - \Delta$.

2pCT Lower Bound Reduction
Input: (x, y) size m input bits for the Exact GHD problem
Output: (A, B) occurrence vectors for $\mathrm{Poi}(t)$ samples for the 2pCT Problem.

1. For each $i, j \in \{1, \dots k\}$, and for each $c \in [m]$ (corresponding to a coordinate of x, y), we take $z_c = \mathrm{Poi}(\frac{d}{\Delta} \cdot D(i, j))$, and generate z_c pairs $(i \cdot x_c, j \cdot y_c)$ (i.e., we set the corresponding coordinate of A or B to i or j iff $x_c = 1$ or $y_c = 1$ respectively);
2. For each $i \in \{1, \dots k\}$, generate $\mathrm{Poi}(d \cdot D(i, 0))$ pairs $(i, 0)$, and similarly-distributed number of pairs $(0, i)$;
3. For each $i, j \in \{1, \dots k\}$, generate $\mathrm{Poi}(l \cdot L(i, j) - m_c \cdot \frac{d}{\Delta} D(i, j))$ pairs (i, j);
4. For each $i \in \{1, \dots k\}$, generate $\mathrm{Poi}(l \cdot L(i, 0) - \frac{m}{4} \cdot \frac{d}{\Delta} \sum_{j=1}^{k} D(i, j))$ pairs $(i, 0)$, and similarly-distributed number of pairs $(0, i)$.
5. Generate the required number of $(0, 0)$ pairs so that A, B have length precisely n;
6. Permute the coordinates of A, B using a common randomly picked permutation over $[n]$.

In each of steps (1)-(5) above, we say we "generate a pair (i, j)" which corresponds to setting the next coordinate of A and B to i and j respectively. We only use the input vectors (x, y) in step (1). We note that all random variables are chosen using shared randomness.

We first claim that the above reduction is well-defined, and in particular all arguments of the Poisson variables are positive.

▷ Claim 11. All the Poisson random variables from above have positive argument.

Proof. We only need to prove this for steps 3 and 4 as the other ones are obvious. Indeed, for $i, j \geq 1$, we get that $l \cdot L(i, j) = t \log n \cdot (\Omega(1/\log n))^{i+j} = t/\log n \cdot (\Omega(1/\log n))^{i+j-2}$, whereas, $m_c \cdot \frac{d}{\Delta} D(i, j) = O(\sqrt{m} \cdot n \cdot (t/2d)^{i+j}) \leq \frac{n^2}{t \log^{1.5} n} \cdot O(t^2/n^2) \cdot (O(t/n))^{i+j-2} \leq \frac{t}{\log^{1.5} n}(O(t/n))^{i+j-2}$. Thus $l \cdot L(i, j) - m_c \cdot \frac{d}{\Delta} D(i, j) \geq 0$ for all $i, j \geq 1$.

Similarly, for step 5, for $i \geq 1$, we have, $l \cdot L(i,0) = \Omega(t \cdot (O(1/\log n))^{i-1})$, whereas, $m/4 \cdot \frac{d}{\Delta} \sum_{j\geq 1} D(i,j) \leq O(\sqrt{m} \cdot n \cdot \sum_{j \geq 1} (O(t/n))^{i+j}) \leq O(\frac{n^2}{t \log^{1.5} n} \cdot (O(t/n))^{i+1}) \leq O(\frac{t}{\log^{1.5} n} \cdot (O(t/n))^{i-1})$. We again have $l \cdot L(i,0) - m/4 \cdot \frac{d}{\Delta} \sum_{j \geq 1} D(i,j) \geq 0$ as required. ◁

We now prove the core of the reduction: that the distribution of (A,B) (denoted $\hat{\mathcal{D}}$) is close to the distribution $\mathcal{D}$ of occurrence vectors of $\mathrm{Poi}(t)$ i.i.d. samples from (a,b), such that $a = b$ if $\|x-y\|_1 = m/2$, and similarly, $\|a-b\|_1 \geq 1/2$ when $\|x-y\|_1 \geq m/2 + \beta$. We will prove that, for distribution of (co-)occurrences of large items is nearly same in the two instances; and similarly for the dense items. We partition the coordinates of (x,y) in the following four groups, each corresponding to either occurrences of dense or large items:

- large: $m_c = m/4 - \Delta$ coordinates for each of $(1,1)$ and $(0,0)$ coordinate pairs (i.e., coordinates $i \in [m]$ where $(x_i, y_i) = (1,1)$ or $(x_i, y_i) = (0,0)$);
- large: $m/4$ coordinates for each of $(1,0)$ and $(0,1)$ pairs;
- dense: $\Delta - \delta$ coordinates for each of $(1,1)$ and $(0,0)$ pairs;
- dense: δ coordinates for each of $(1,0)$ and $(0,1)$ pairs.

Note that this accounts for all coordinates for a pair x,y such that $\|x-y\|_1 = m/2 + 2\delta$.

Next, we compare the distributions of occurrences for large and dense items in the generated vectors (A,B) as opposed to occurrences of items coming from distributions a,b defined above. In particular, we consider the distribution of counts $c_{i,j}$, where $i+j > 0$, where $c_{i,j}$ is the number of large items which where sampled i times on Alice's side and j times on the Bob's side; we will refer to them as (i,j) occurrence pairs.

We denote by $\hat{\mathcal{D}}^L, \hat{\mathcal{D}}^D$ the distribution of, respectively, large and dense $\{c_{i,j}\}_{i+j>0}$ occurrence pairs in (A,B). Similarly, we denote $\mathcal{D}^L, \mathcal{D}^D$ the distribution of, respectively, large and dense $\{c_{i,j}\}_{i+j>0}$ occurrence pairs randomly drawn from (a,b). Note that $\mathcal{D}^L, \mathcal{D}^D$ are multinomial distributions, formally defined as follows:

▶ **Definition 12.** *Fix $n, k \geq 1$, vector $\vec{p} \in \mathbb{R}_+^k$, where $\sum_{i=1}^k p_i \leq 1$. The k-dimensional random variable $(M_1, \ldots M_k) = \mathrm{Mult}_{\text{-}0}(n; \vec{p})$ is obtained by drawing a Multinomial r.v. with parameters n and probability vector $(1 - \sum_{i=1}^k p_i, \vec{p})$, and dropping the first coordinate.*

In particular, $\mathcal{D}^L = \mathrm{Mult}_{\text{-}0}(l; \vec{p}_L)$ where $\vec{p}_L = (L(i,j))_{i,j\geq 0; i+j>0}$; $\mathcal{D}^D$ will be clarified later. We now deduce $\hat{\mathcal{D}}^L$ and $\hat{\mathcal{D}}^D$. Below we use the fact that the sum of Poisson random variables is also Poisson.

▷ Claim 13. $\hat{\mathcal{D}}^L$ is distributed as $\mathrm{Poi}(l \cdot \vec{p}_L)$. Also $\hat{\mathcal{D}}^D$ is distributed as $\mathrm{Poi}(\frac{\Delta-\delta}{\Delta} \cdot d \cdot \vec{p}_D) + \mathrm{Poi}(\frac{\delta}{\Delta} \cdot d \cdot \vec{p}_{D^0})$ where $\vec{p}_D = (D(i,j))_{i,j\geq 0; i+j>0}$ and $\vec{p}_{D^0} = (D(i) \cdot \mathbb{1}[j=0] + D(j) \cdot \mathbb{1}[i=0])_{i,j\geq 0; i+j>0}$.

Proof. For $i,j \in \{1, \ldots k\}$, $\hat{\mathcal{D}}^L_{i,j}$ is distributed as $\mathrm{Poi}(l \cdot L(i,j))$, which is composed of $\mathrm{Poi}(m_c \cdot \frac{d}{\Delta} D(i,j))$ (from the first step: there are m_c coordinate pairs $(1,1)$), plus $\mathrm{Poi}(l \cdot L(i,j) - m_c \cdot \frac{d}{\Delta} \cdot D(i,j))$ (from the third step).

Similarly, $\hat{\mathcal{D}}^L_{i,0}$ (and by symmetric argument also $\hat{\mathcal{D}}^L_{0,i}$) is distributed as $\mathrm{Poi}(l \cdot L(i,0))$, composed of $\mathrm{Poi}(m/4 \cdot \frac{d}{\Delta} \sum_{j=1}^k D(i,j))$ (from the first step: there are $m/4$ coordinate pairs $(1,0)$), plus $\mathrm{Poi}(l \cdot L(i,0) - m/4 \cdot \frac{d}{\Delta} \sum_{j=1}^k D(i,j))$ (from step 4).

For $i,j \geq 1$, $\hat{\mathcal{D}}^D_{i,j}$ are distributed as $\mathrm{Poi}((\Delta - \delta) \cdot \frac{d}{\Delta} D(i,j))$ since there are $\Delta - \delta$ coordinate pairs $(1,1)$. For $\hat{\mathcal{D}}^D_{i,0}$ (and similarly $\hat{\mathcal{D}}^L_{0,i}$), the distribution is $\mathrm{Poi}(\delta \cdot \frac{d}{\Delta} \sum_{j=1}^k D(i,j))$ (from the first step: there are δ coordinate pairs $(1,0)$), plus $\mathrm{Poi}(d \cdot D(i,0))$ (from the second step). This amounts to $\mathrm{Poi}(d \cdot \frac{\delta}{\Delta} \sum_{j=1}^k D(i,j) + d \cdot D(i,0)) = \mathrm{Poi}(d \frac{\Delta-\delta}{\Delta} \cdot D(i,0) + d \frac{\delta}{\Delta} \cdot (D(i,0) + \sum_{j=1}^k D(i,j))) = \mathrm{Poi}(d \cdot \frac{\Delta-\delta}{\Delta} \cdot D(i,0) + d \cdot \frac{\delta}{\Delta} \cdot D(i))$. ◁

We prove $\left\|\hat{\mathcal{D}}-\mathcal{D}\right\|_{TV} \le 0.01+o(1)$ by showing (i) $\left\|\hat{\mathcal{D}}^L-\mathcal{D}^L\right\|_{TV} \le 0.01+o(1)$ and (ii) $\left\|\hat{\mathcal{D}}^D-\mathcal{D}^D\right\|_{TV} = o(1)$. We compare $\hat{\mathcal{D}}^L$ and $\hat{\mathcal{D}}^D$ versus $\mathcal{D}^L$ and $\mathcal{D}^D$ using the following estimate on the TV distance between Multinomial and Poisson random variables. Note that the identity of items is not important, as the items are randomly permuted inside the domain, for both A, B as well as in distributions a, b.

▶ **Theorem 14** ([11]). *Let $n, k \ge 1$, as well as a vector $\vec{p} \in \mathbb{R}^k_+$, where $p = \sum_{i=1}^k p_i \le 1$. Consider the random variable $(M_1, \dots M_k)$ drawn from the Multinomial $\mathrm{Mult}_{\text{-}0}(n; \vec{p})$. Also consider the Poisson random variable $P = (P_1, \dots, P_k)$ where $P_i \sim \mathrm{Poi}(np_i)$. Then the variables $M = (M_1, \dots M_k)$ and $(P_1, \dots P_k)$ are at a statistical distance of $O(p \log n)$.*

By Theorem 14, the TV-distance between $\mathcal{D}^L = \mathrm{Mult}_{\text{-}0}(l; \vec{p}_L)$ and $\hat{\mathcal{D}}^L = \mathrm{Poi}(l \cdot \vec{p}_L)$ is bounded by: $O(\log n) \cdot \sum_{i,j \ge 0; i+j>0} L(i,j) \le O(\log n) \cdot \sum_{i,j\ge 0; i+j>0} (t/2l)^{i+j} \le O(1)/C \le 0.01$ (for sufficiently large constant C).

For (ii), we note $\mathcal{D}^D$ can be thought of as two distributions, corresponding to: (1) items in $S_A \cap S_B$, (2) items in $S_A \triangle S_B$. The occurrence counts for (1) are distributed as a Multinomial M^D with parameters $|S_A \cap S_B| = d \cdot \frac{\Delta-\delta}{\Delta}$ and probability vector $\vec{p}_D = (D(i,j))_{i,j\ge 0; i+j>0}$. By Theorem 14, the TV-distance between M^D and the distribution $\mathrm{Poi}(d \cdot \frac{\Delta-\delta}{\Delta} \cdot \vec{p}_D)$ is bounded by: $O(\log n) \cdot \sum_{i,j\ge 0; i+j>0} D(i,j) \le O(\log n) \cdot \sum_{i,j \ge 0; i+j>0} (t/2d)^{i+j} \le O(1/\log n)$.

For (2), we can only have $(i, 0)$ and $(0, j)$ pairs, and the occurrence counts are distributed as a Multinomial $M^{D^0} = \mathrm{Mult}_{\text{-}0}(|S_A \triangle S_B|/2; \vec{p}_{D^0})$. By Theorem 14, the TV-distance between M^{D^0} and $\mathrm{Poi}(\frac{\delta}{\Delta} \cdot d \cdot \vec{p}_{D^0})$ is at most $O(\log n) \cdot \sum_{i \ge 1} D(i) \le O(1/\log n)$.

Thus we conclude that the distributions $\hat{\mathcal{D}}$ and $\mathcal{D}$ are at a small TV distance. ◀

References

1. Jayadev Acharya, Clément L. Canonne, Cody Freitag, and Himanshu Tyagi. Test without Trust: Optimal Locally Private Distribution Testing. *CoRR*, abs/1808.02174, 2018.
2. Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Distributed Simulation and Distributed Inference. *CoRR*, abs/1804.06952, 2018. `arXiv:1804.06952`.
3. Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Shengjun Pan. Competitive closeness testing. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 47–68, 2011.
4. Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, and Ananda Suresh. Competitive classification and closeness testing. In *Conference on Learning Theory*, pages 22–1, 2012.
5. Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Sublinear algorithms for outlier detection and generalized closeness testing. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 3200–3204. IEEE, 2014.
6. Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially Private Testing of Identity and Closeness of Discrete Distributions. *CoRR*, abs/1707.05128, 2017. `arXiv:1707.05128`.
7. Rudolf Ahlswede and Imre Csiszár. Hypothesis testing with communication constraints. *IEEE transactions on information theory*, 32(4), 1986.
8. Maryam Aliakbarpour, Ilias Diakonikolas, and Ronitt Rubinfeld. Differentially Private Identity and Closeness Testing of Discrete Distributions. *CoRR*, abs/1707.05497, 2017. `arXiv:1707.05497`.
9. Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comp. Sys. Sci.*, 58:137–147, 1999. Previously appeared in STOC'96.
10. S Amari et al. Statistical inference under multiterminal data compression. *IEEE Transactions on Information Theory*, 44(6):2300–2324, 1998.

11 Andrew D Barbour. Stein's method and Poisson process convergence. *Journal of Applied Probability*, 25(A):175–184, 1988.

12 Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D Smith, and Patrick White. Testing closeness of discrete distributions. *Journal of the ACM (JACM)*, 60(1):4, 2013.

13 Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing Closeness of Discrete Distributions. *J. ACM*, 60(1):4:1–4:25, February 2013. doi:10.1145/2432622.2432626.

14 Amos Beimel, Paz Carmi, Kobbi Nissim, and Enav Weinreb. Private Approximation of Search Problems. *SIAM J. Comput.*, 38(5):1728–1760, 2008.

15 Amos Beimel, Renen Hallak, and Kobbi Nissim. Private Approximation of Clustering and Vertex Cover. *Computational Complexity*, 18(3):435–494, 2009.

16 Bhaswar Bhattacharya and Gregory Valiant. Testing closeness with unequal sized samples. In *Advances in Neural Information Processing Systems*, pages 2611–2619, 2015.

17 Elette Boyle, Niv Gilboa, and Yuval Ishai. Breaking the Circuit Size Barrier for Secure Computation Under DDH. In *CRYPTO (1)*, volume 9814 of *Lecture Notes in Computer Science*, pages 509–539. Springer, 2016.

18 Elette Boyle, Niv Gilboa, and Yuval Ishai. Group-Based Secure Computation: Optimizing Rounds, Communication, and Computation. In *EUROCRYPT (2)*, volume 10211 of *Lecture Notes in Computer Science*, pages 163–193, 2017.

19 Mark Braverman, Ankit Garg, Tengyu Ma, Huy L Nguyen, and David P Woodruff. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1011–1020. ACM, 2016.

20 Sergey Bravyi, Aram W Harrow, and Avinatan Hassidim. Quantum algorithms for testing properties of distributions. *IEEE Transactions on Information Theory*, 57(6):3971–3981, 2011.

21 Bryan Cai, Constantinos Daskalakis, and Gautam Kamath. Priv'IT: Private and Sample Efficient Identity Testing. *arXiv preprint*, 2017. arXiv:1703.10127.

22 Clément L Canonne. A survey on distribution testing: Your data is big. but is it blue? In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 22(63), pages 1–9, 2015.

23 Amit Chakrabarti and Oded Regev. An optimal lower bound on the communication complexity of gap-hamming-distance. *SIAM Journal on Computing*, 41(5):1299–1317, 2012.

24 Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1193–1203. Society for Industrial and Applied Mathematics, 2014.

25 Thomas M Cover. Hypothesis testing with finite statistics. *The Annals of Mathematical Statistics*, 40(3):828–835, 1969.

26 Michael Crouch, Andrew McGregor, Gregory Valiant, and David P Woodruff. Stochastic streams: Sample complexity vs. space complexity. In *LIPIcs-Leibniz International Proceedings in Informatics*, volume 57. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

27 Yuval Dagan and Ohad Shamir. Detecting Correlations with Little Memory and Communication. *CoRR*, abs/1803.01420, 2018. arXiv:1803.01420.

28 I. Diakonikolas, E. Grigorescu, J. Li, A. Natarajan, K. Onak, and L. Schmidt. Communication-Efficient Distributed Learning of Discrete Distributions. In *Advances in Neural Information Processing Systems*, 2017. To appear.

29 Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Collision-based testers are optimal for uniformity and closeness. *arXiv preprint*, 2016. arXiv:1611.03579.

30 Ilias Diakonikolas, Moritz Hardt, and Ludwig Schmidt. Differentially private learning of structured discrete distributions. In *Advances in Neural Information Processing Systems*, pages 2566–2574, 2015.

31 Ilias Diakonikolas and Daniel M Kane. A new approach for testing properties of discrete distributions. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 685–694. IEEE, 2016.

32 Joan Feigenbaum, Yuval Ishai, Tal Malkin, Kobbi Nissim, Martin J. Strauss, and Rebecca N. Wright. Secure Multiparty Computation of Approximations. *ACM Trans. Algorithms*, 2(3):435–472, July 2006. doi:10.1145/1159892.1159900.

33 Sumegha Garg, Ran Raz, and Avishay Tal. Extractor-based time-space lower bounds for learning. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 990–1002. ACM, 2018.

34 Craig Gentry. Fully homomorphic encryption using ideal lattices. In *STOC*, pages 169–178. ACM, 2009.

35 Oded Goldreich. Introduction to Property Testing (working draft), 2017. URL: www.wisdom.weizmann.ac.il/~oded/PDF/pt-v3.pdf.

36 Oded Goldreich and Dana Ron. On Testing Expansion in Bounded-Degree Graphs. *ECCC 7(20)*, 2000.

37 Michael Gutman. Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Transactions on Information Theory*, 35(2):401–408, 1989.

38 U. Hadar, J. Liu, Y. Polyanskiy, and O. Shayevitz. Communication Complexity of Estimating Correlations. In *Proceedings of the Symposium on Theory of Computing (STOC)*, 2019.

39 Shai Halevi, Robert Krauthgamer, Eyal Kushilevitz, and Kobbi Nissim. Private approximation of NP-hard functions. In *STOC*, pages 550–559. ACM, 2001.

40 Te Han. Hypothesis testing with multiterminal data compression. *IEEE transactions on information theory*, 33(6):759–772, 1987.

41 Martin E Hellman and Thomas M Cover. Learning with finite memory. *The Annals of Mathematical Statistics*, pages 765–782, 1970.

42 Piotr Indyk. Stable Distributions, Pseudorandom Generators, Embeddings and Data Stream Computation. *J. ACM*, 53(3):307–323, 2006. Previously appeared in FOCS'00.

43 Piotr Indyk and David Woodruff. Polylogarithmic Private Approximations and Efficient Matching. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 245–264, Berlin, Heidelberg, 2006. Springer-Verlag. doi:10.1007/11681878_13.

44 Yuval Ishai, Tal Malkin, Martin J. Strauss, and Rebecca N. Wright. Private multiparty sampling and approximation of vector combinations. *Theor. Comput. Sci.*, 410(18):1730–1745, 2009.

45 Joe Kilian, André Madeira, Martin J. Strauss, and Xuan Zheng. Fast Private Norm Estimation and Heavy Hitters. In *TCC*, volume 4948 of *Lecture Notes in Computer Science*, pages 176–193. Springer, 2008.

46 Gillat Kol, Ran Raz, and Avishay Tal. Time-space hardness of learning sparse parities. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1067–1080. ACM, 2017.

47 Eyal Kushilevitz, Rafail Ostrovsky, and Yuval Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM J. Comput.*, 30(2):457–474, 2000. Preliminary version appeared in STOC'98.

48 Dana Moshkovitz and Michal Moshkovitz. Mixing implies lower bounds for space bounded learning. In *Conference on Learning Theory*, pages 1516–1566, 2017.

49 Moni Naor and Kobbi Nissim. Communication Complexity and Secure Function Evaluation. *CoRR*, cs.CR/0109011, 2001. arXiv:cs.CR/0109011.

50 Ran Raz. Fast learning requires good memory: A time-space lower bound for parity learning. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 266–275. IEEE, 2016.

51 Ran Raz. A time-space lower bound for a large class of learning problems. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 732–742. IEEE, 2017.

52 Ronitt Rubinfeld. Sublinear time algorithms. In *International Congress of Mathematicians*, volume 3, pages 1095–1110, 2006.

53 Ronitt Rubinfeld. Taming big probability distributions. *XRDS: Crossroads, The ACM Magazine for Students*, 19(1):24–28, 2012.

54 KR Sahasranand and Himanshu Tyagi. Extra Samples can Reduce the Communication for Independence Testing. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 2316–2320. IEEE, 2018.

55 Ofer Shayevitz. On Rényi measures and hypothesis testing. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 894–898. IEEE, 2011.

56 Or Sheffet. Locally Private Hypothesis Testing. In *ICML*, 2018.

57 Alexander A. Sherstov. The Communication Complexity of Gap Hamming Distance. *Theory of Computing*, 8(1):197–208, 2012. doi:10.4086/toc.2012.v008a008.

58 Jayakrishnan Unnikrishnan. On optimal two sample homogeneity tests for finite alphabets. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, pages 2027–2031. Ieee, 2012.

59 Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011. Previously in STOC'08.

60 Elad Verbin and Wei Yu. The streaming complexity of cycle counting, sorting by reversals, and other problems. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 11–25. SIAM, 2011.

61 Andrew Chi-Chih Yao. Protocols for Secure Computations (Extended Abstract). In *Proceedings of the 23rd IEEE Symposium on Foundations of Computer Science*, FOCS '82, pages 160–164, 1982.