Inference Under Information Constraints III: Local Privacy Constraints

Jayadev Acharya, *Member, IEEE*, Clément L. Canonne, Cody Freitag, Ziteng Sun, and Himanshu Tyagi, *Senior Member, IEEE*

***Abstract*—We study goodness-of-fit and independence testing of discrete distributions in a setting where samples are distributed across multiple users. The users wish to preserve the privacy of their data while enabling a central server to perform the tests. Under the notion of local differential privacy, we propose simple, sample-optimal, and communication-efficient protocols for these two questions in the noninteractive setting, where in addition users may or may not share a common random seed. In particular, we show that the availability of shared (public) randomness greatly reduces the sample complexity. Underlying our public-coin protocols are privacy-preserving mappings which, when applied to the samples, minimally contract the distance between their respective probability distributions.**

***Index Terms*—Distributed inference, privacy, goodness-of-fit, local differential privacy.**

I. Introduction

INFERRING statistical properties of data sources while maintaining their privacy is a core problem in privacy-preserving statistics. A widely established notion to achieve this is *local differential privacy (LDP)*, introduced in [30], [38]. The data samples are distributed across users ("players"), who do not trust the centralized data curator, which can be, e.g., corporate entities or government agencies. The data samples are privatized via a noise addition mechanism that is locally differentially private (see Eq. 1). This falls under the general setting of statistical inference under *local information constraints*, namely constraints on information that each player can reveal about its sample.

Manuscript received August 14, 2020; revised December 4, 2020; accepted January 14, 2021. Date of publication January 22, 2021; date of current version March 16, 2021. The work of Jayadev Acharya was supported by NSF-CCF-1846300 (CAREER), NSF-CCF-1815893, and a Google Faculty Research Award. The work of Clément L. Canonne was supported by the Goldstine Fellowship. The work of Cody Freitag was supported in part by NSF GRFP under Award DGE-1650441. The work of Ziteng Sun was supported in part by NSF-CCF-1846300 (CAREER). The work of Himanshu Tyagi was supported in part by a Research Grant from the Robert Bosch Center for Cyberphysical Systems (RBCCPS), Indian Institute of Science, Bangalore. A preliminary version of this work containing partial results appeared in the *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019 [1]. *(Corresponding author: Jayadev Acharya.)*

Jayadev Acharya and Ziteng Sun are with the Department of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14850 USA (e-mail: acharya@cornell.edu; zs335@cornell.edu).

Clément L. Canonne was with IBM Research, San Jose, CA 95120, USA. He is now with the School of Computer Science, University of Sydney, Sydney, NSW 2006, Australia (e-mail: ccanonne@cs.columbia.edu).

Cody Freitag is with the Department of Computer Science, Cornell Tech, New York, NY 10044 USA (e-mail: cfreitag@cs.cornell.edu).

Himanshu Tyagi is with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore 560012, India (e-mail: htyagi@iisc.ac.in).

Digital Object Identifier 10.1109/JSAIT.2021.3053569

Recently, a subset of the authors have initiated a systematic study of such problems under general constraints. In particular, [2] provides a framework for deriving lower bounds for such problems and [3] provides sample-optimal algorithms for communication constraints. This article, the third in this series, focuses on local privacy constraints. Specifically, we consider two of the most fundamental goodness-of-fit tasks, testing identity and independence of discrete distributions, and design sample-optimal LDP mechanisms for these tasks. We restrict to simultaneous message passing protocols and lay special emphasis on the availability of *public randomness* at the players (*i.e.*, a common random seed shared by all parties)[1] and seek to answer the following.

What is the sample complexity of testing identity and independence of discrete distrbutions under local differential privacy? Does the sample complexity depend on whether public randomness is available?

The role of public randomness in the design and analysis of distributed statistical inference has hitherto been largely overlooked. We fully resolve this question by providing tight bounds on the sample complexity of identity testing and independence testing of discrete distributions under local differential privacy, both with and without public randomness. Our results show that, for these two composite hypothesis testing tasks, schemes which allow for public randomness can achieve significantly smaller sample complexity than those who do not. Interestingly, this is in contrast with the seminal work of Tsitsiklis [45], which established that public randomness provides no advantage in the context of distributed *simple* hypothesis testing without local privacy constraints.

A. Results and Techniques

We study two inference problems over discrete distributions, identity testing and independence testing under ρ-LDP (at a high level, the privacy parameter $\rho > 0$ bounds the (worst-case) statistical leakage of any player's data, and smaller values imply stronger privacy guarantees; see Section II for formal definitions). Our results are summarized in Table I; we outline and discuss them below.

In the identity testing question, there is a known reference distribution $\mathbf{q}$ over $[k] := \{1, \ldots, k\}$, and the players' samples are i.i.d. from an unknown distribution $\mathbf{p}$. The goal is to test the hypotheses $\mathcal{H}_0 : \mathbf{p} = \mathbf{q}$ and $\mathcal{H}_1 : \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$

[1] We assume private randomness is always available at the players. Formal definitions can be found in Section II.

TABLE I
SUMMARY OF OUR RESULTS AND PREVIOUS WORK

	This work		**Previous work**	
	Private-Coin	Public-Coin	Private-Coin	Public-Coin
Identity Testing	$O\left(\frac{k^{3/2}}{\varepsilon^2\rho^2}\right)$	$O\left(\frac{k}{\varepsilon^2\rho^2}\right)$	$O\left(\frac{k^2}{\varepsilon^2\rho^2}\right)$, $\Omega\left(\frac{k^{3/2}}{\varepsilon^2\rho^2}\right)$	$\Omega\left(\frac{k}{\varepsilon^2\rho^2}\right)$
Independence Testing	$\Theta\left(\frac{k^3}{\varepsilon^2\rho^4}\right)$	$\Theta\left(\frac{k^2}{\varepsilon^2\rho^2}\right)$	$O\left(\frac{k^4}{\varepsilon^2\rho^2}\right)$	

using ρ-LDP mechanisms. We seek to characterize the *sample complexity* of this task, which is the minimum number of players to solve this problem with a (small) constant two-sided error. Without privacy constraints, when the true samples of $\mathbf{p}$ are available to the central data curator ("referee"), the optimal sample complexity of identity testing is known to be $\Theta(k^{1/2}/\varepsilon^2)$.

There are two parts of the problem. The first is to design *privacy-preserving mechanisms* that the players use to encode their data to be sent the server. The second is to design *post-processing algorithms* that the server uses to decide the output of the test given the privatized messages.

We first consider the task of designing optimal post-processing algorithms for existing ρ-LDP mechanisms. This is of interest in cases where the privatization mechanisms are in place, and changing them is impossible or too expensive—for instance, when an organization has already deployed a data aggregation pipeline, and seeks to add a statistical inference component to it without overhauling the entire system.

Arguably the simplest privatization scheme is k-randomized response (see [48]). Unfortunately, it was shown in [44] that the sample complexity of any test relying on this scheme is $\Theta(k^{5/2}/\varepsilon^2\rho^2)$, far from optimal. Our first result considers the now well established privatization scheme RAPPOR (Randomized Aggregatable Privacy-Preserving Ordinal Response [29], [37]) (see III-A1). In Theorem 1, we design an identity testing algorithm that, given samples from the RAPPOR mechanism, has sample complexity $O\left(\frac{k^{3/2}}{\varepsilon^2\rho^2}\right)$—a factor k improvement over randomized response.

The RAPPOR mechanism produces privatized messages with $\Omega(k)$ bits of entropy, and as a result those messages are k-bit long. Thus, RAPPOR requires a large communication bandwidth. We provide a new mechanism based on the recently proposed Hadamard Response (HR) that produces only one-bit messages, leading to an identity testing algorithm with the same sample complexity $O\left(\frac{k^{3/2}}{\varepsilon^2\rho^2}\right)$. This result is given in Theorem 2.

All the schemes above require no publicly agreed upon randomness, which we refer to as *private-coin* mechanisms, and are highly desirable when it is too inefficient or infeasible to setup a common random seed. However, in [2], it was established that any testing algorithm based on any private-coin ρ-LDP mechanism must use $\Omega(\frac{k^{3/2}}{\varepsilon^2\rho^2})$ players. Therefore, the algorithms we propose based on RAPPOR and HR are the best possible, and more significantly, are optimal among all LDP schemes that do not use public randomness.

This raises the question of building LDP mechanisms that do use public randomness, which we refer to as *public-coin* mechanisms, and post-processing algorithms that require fewer samples. We emphasize that the public randomness is used only for added utility and we require the same strong privacy guarantees. In this context, we design a new public-coin ρ-LDP mechanism and a corresponding algorithm whose sample complexity is $O\left(\frac{k}{\varepsilon^2\rho^2}\right)$, a factor $\sqrt{k}$ improvement over the best possible without using public randomness. Furthermore, this is asymptotically optimal from the result of [2], and the mechanism only uses one bit of communication from each player, making it as communication-efficient as possible. Our result relies on a randomized one-bit isometry, where the players use the common random seed to randomly project the original domain $[k]$ to a binary domain and perform testing over this new domain. This result is given in Theorem 4.

We then turn to the task of independence testing. Here, the underlying distribution $\mathbf{p}$ is over the product domain $[k]\times[k]$, and the goal is to test whether the marginals of $\mathbf{p}$ are independent (*i.e.*, if $\mathbf{p}$ is a product distribution) or at least ε away from all product distributions. We design schemes without and with public randomness which achieve sample complexity $O\left(\frac{k^3}{\varepsilon^2\rho^2}\right)$ and $O\left(\frac{k^2}{\varepsilon^2\rho^2}\right)$, respectively. These results are given in Theorem 5 and Theorem 8. Interestingly, in the case where public randomness is available, our protocol relies on a one-bit isometry similar to the one used in the identity testing case, but suitably generalized to handle the product structure of the domain. Finally, we prove the optimality of both these bounds, establishing matching lower bounds in Theorem 9. This is done by providing a formal reduction from independence testing over $[k]\times[k]$ to the identity testing problem over $[k^2]$. We believe this general reduction, which is not specific to the locally private setting, to be of independent interest.

The conceptual takeaway message of our results is that, for composite hypothesis testing problems, public randomness can prove very helpful, and its availability leads to significantly more sample-efficient protocols.

We finally remark that although this work is concerned with noninteractive protocols, more complicated *adaptive* LDP schemes are possible where the players sequentially choose their privatization schemes upon observing the messages of all previous players and the available public randomness. Recent works in this setting [5], [12], [17] show that, for identity testing, adaptivity does not allow for more efficient protocols than public randomness, and by our reduction for independence

testing this carries over to the independence testing problem as well.

B. Related Prior Work

Testing properties of distributions from their samples has a long history in statistics, which dates back more than a century. Recently, this problem has gathered renewed interest in the computer science community, with a particular focus on the study of discrete distributions in the finite-sample regime. In this section, we only focus on closely related papers and we refer an interested reader to surveys and books [13], [21], [34], [43] for a comprehensive treatment.

Following a long line of work, the optimal sample complexity for identity testing has been established as $\Theta(k^{1/2}/\varepsilon^2)$ [33], [42], [46] under constant error probability. Reference [25], [35] establish the optimal dependence on the error probability. Reference [19], [46] also study the "instance-optimal" variant of the problem, introduced in [46]. The optimal sample complexity for the independence testing problem where both observations are from the same set $[k]$ was studied in [16], [41], and shown to be $\Theta(k/\varepsilon^2)$ in [6], [26].

Distribution testing has also been studied under privacy constraints on the samples. Under the notion of (global) *differential privacy* (DP) [28], identity testing has been considered in [11], [20], with a complete characterization of the sample complexity derived in [8]. Reference [23] focuses on the class of product distributions in high dimensions, including product of Bernoulli's and Gaussians with known variances. Both these works show that, in certain parameter regimes, the sample complexity can match the sample complexity of the non-private counterpart of the problem, which is in sharp contrast to the more stringent case of *local* privacy (LDP) considered in this article. Finally, [12] and [14] consider uniformity testing (a specific case of identity testing) under the notions of *pan-privacy* and *shuffle privacy*, respectively, which provide privacy guarantees in-between DP and LDP.

Independence testing under *differentially privacy* has been studied in [31], [40], [47] and the first algorithm with finite sample guarantee was given in [10].

The works most closely related to ours are those that consider distribution testing under LDP constraints [1], [3], [4], [5], [12], [17], [32], [42]. Reference [44] considers both identity testing and independence testing with private-coin, noninteractive schemes. Our results improve upon theirs by a factor of k and k^2, respectively. Reference [3] establishes lower bounds for identity testing using both private-coin and public-coin noninteractive schemes, which match our bounds in both cases and imply the optimality of our results. Reference [4] considers noninteractive schemes where only a limited amount of public randomness is available, and obtains the optimal sample complexity which interpolates smoothly between the private-coin and public-coin cases. References [5], [12], [17] consider identity testing using sequentially interactive schemes, which combined with our results prove that interactivity cannot lead to an improvement in the sample complexity over public-coin noninteractive schemes. We note that the recent work of Joseph *et al.* [36] also considers the role of interactivity in LDP hypothesis testing; however, they focus on simple hypothesis testing (as well as a generalization to *convex* hypothesis classes). Their results do not apply to identity testing, and are incomparable to ours.

Another class of problems of statistical inference, density estimation, requires learning the unknown distribution up to a desired accuracy of ε in total variation distance. The optimal sample complexity of locally private learning discrete k-ary distributions is known to be $\Theta(k^2/(\varepsilon^2\rho^2))$; see [7], [9], [27], [29], [37], [49]. The private-coin identity testing schemes in this article are based on the same LDP randomization schemes proposed in these papers at the user side. Specifically, RAPPOR was independently proposed in [9], [27] and analyzed in [37]. Hadamard Response and its one-bit variant are proposed in [7], [9]. Also, [15] uses Hadamard transform together with sampling to reduce user communication to $O(1)$ bits in a public-coin scheme. Moreover, our private-coin independence testing protocol also involves a step that learns both marginal distributions, which relies on the scheme from [7].

C. Organization

The rest of the article is organized as follows. In Sections III-A1 and III-A2 we provide two private-coin LDP schemes for identity testing based on RAPPOR and Hadamard Response respectively, and analyze their sample complexity. In Section III-B we establish an upper bound on the sample complexity of public-coin protocols for identity testing. In Sections IV-A and IV-B we establish the upper bounds on private-coin and public-coin independence testing, respectively. Finally, in Section IV-C we provide a reduction between identity and independence testing and use it to prove the optimality of the proposed independence tests both for private- and public-coin protocols.

II. The Setup: Local Privacy and Inference Protocols

A. Notation

Throughout the article, we denote by log the natural logarithm and $\log_2$ the base 2 logarithm. We use standard asymptotic notation $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$ for complexity orders.[2]

For a known and fixed discrete domain $\mathcal{X}$ let $\Delta_{\mathcal{X}}$ be the set of probability distributions over $\mathcal{X}$, *i.e.*,

$$\Delta_{\mathcal{X}} = \left\{\mathbf{p}\colon \mathcal{X} \to [0,1] : \ \|\mathbf{p}\|_1 = 1\right\},$$

where we identify a probability distribution to its probability mass function. We denote by $\mathbf{u}_{\mathcal{X}}$ the uniform distribution on $\mathcal{X}$ and omit the subscript when the domain is clear from context.

We are mostly interested in k-ary discrete distributions, and assume without loss of generality that $\mathcal{X} = [k] := \{1, 2, \ldots, k\}$. We use $\Delta_{[k]}$ and Δ_k interchangeably to denote the probability simplex consisting of all distributions over $[k]$.

[2]Namely, for two non-negative sequences $(a_n)_n$ and $(b_n)_n$, we write $a_n = O(b_n)$ (resp., $a_n = \Omega(b_n)$) if there exist $C > 0$ and $N \geq 0$ such that $a_n \leq Cb_n$ (resp., $a_n \geq Cb_n$) for all $n \geq N$. Further, we write $a_n = \Theta(b_n)$ when both $a_n = O(b_n)$ and $a_n = \Omega(b_n)$ hold.

The *total variation distance* between distributions $\mathbf{p}, \mathbf{q} \in \Delta_{\mathcal{X}}$ is

$$\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \coloneqq \sup_{S \subseteq \mathcal{X}} (\mathbf{p}(S) - \mathbf{q}(S)) = \frac{1}{2} \sum_{x \in \mathcal{X}} |\mathbf{p}(x) - \mathbf{q}(x)|,$$

namely, $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})$ is equal to half of the ℓ_1 distance of $\mathbf{p}$ and $\mathbf{q}$. For a distance parameter $\varepsilon \in (0, 1]$, we say that $\mathbf{p}, \mathbf{q} \in \Delta_{\mathcal{X}}$ are *ε-far* if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$. Finally, for two distributions $\mathbf{p}_1$ and $\mathbf{p}_2$ over $\mathcal{X}$, we denote by $\mathbf{p}_1 \otimes \mathbf{p}_2$ the product distribution over $\mathcal{X} \times \mathcal{X}$ defined by $(\mathbf{p}_1 \otimes \mathbf{p}_2)(x_1, x_2) = \mathbf{p}_1(x_1) \cdot \mathbf{p}_2(x_2)$ for all $x_1, x_2 \in \mathcal{X}$.

B. Local Differential Privacy and Protocols

For a data domain $\mathcal{X}$ and some set $\mathcal{Y}$ (which denotes the message set), a channel $W \colon \mathcal{X} \to \mathcal{Y}$ is *ρ-locally differentially private* (ρ-LDP) *mechanism* if [28], [29], [30]

$$\max_{y \in \mathcal{Y}} \max_{x, x' \in \mathcal{X}} \frac{W(y|x')}{W(y|x)} \leq e^{\rho}. \tag{1}$$

where, slightly overloading notation, we write $W(\cdot|x)$ for the output distribution (on $\mathcal{Y}$) for input $x \in \mathcal{X}$. Loosely speaking, no output message from a user can reveal too much about their sample. Let $\mathcal{W}_\rho$ be the set of all ρ-LDP channels with output $\{0, 1\}^*$ the set of all binary strings.

Our setup is depicted in Fig. 1. There are n independent samples $X^n \coloneqq X_1, \ldots, X_n$ from an unknown distribution $\mathbf{p}$ distributed across n players, with player i holding X_i. Player i passes X_i through a privatization channel $W_i \in \mathcal{W}_\rho$ and the output Y_i is their message. Note that once the channel W_i is fixed the output distribution of messages is only a function of X_i. We now describe the various communication protocols which restrict how the choice of W_is can be performed.

We restrict ourselves to *simultaneous message passing* (SMP) protocols of communication, *i.e.*, noninteractive LDP mechanisms, where the W_is are all selected simultaneously. Within SMP protocols, we distinguish between the case where a common random seed (public randomness) is available across players and can be used by them to select the W_is, and the case there is no public randomness available and they must choose the W_is independently. In both cases, however, the players are assumed to have access to private randomness, which is needed to implement any privatization mechanism. We describe these two cases in more detail below.

Definition 1 (Private-Coin SMP Protocols): Let $U_1, \ldots, U_n$ be independent random variables, which are also independent jointly of $(X_1, \ldots, X_n)$. U_i is the private randomness available to player i. A ρ-LDP *private-coin* SMP protocol π consists of the following two steps: (a) Player i selects their channel $W_i \in \mathcal{W}_\rho$ (possibly as a function of U_i), (b) and sends their message $Y_i \in \mathcal{Y}$, which is obtained by passing X_i through W_i, to the referee. The referee receives the messages $(Y_1, \ldots, Y_n) \coloneqq \pi(X^n)$. We assume that the protocol is decided ahead of time, so that the distribution of the U_i is known to the referee, but not their instantiation.

Since the random variables X_i and U_i are independent across players, and the message Y_i from player i is a randomized function of (X_i, U_i) the messages $(Y_1, \ldots, Y_n)$ are all independent across players.

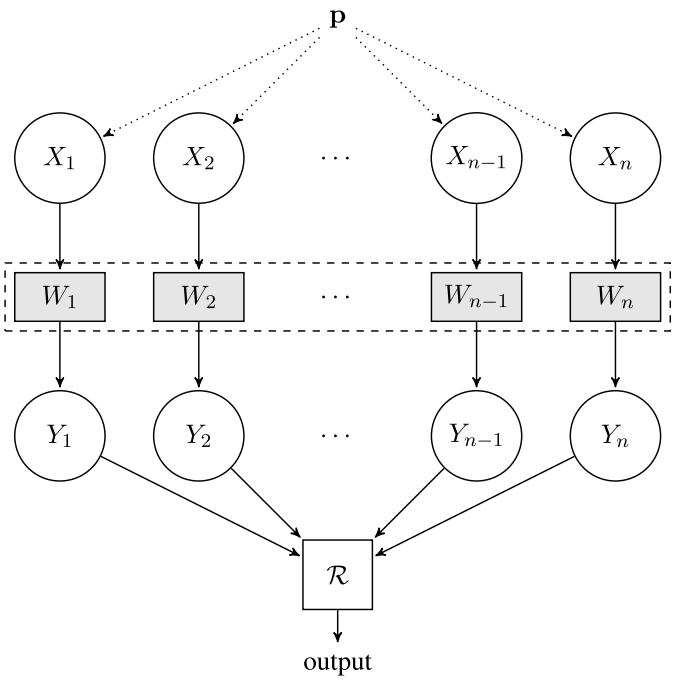

Fig. 1. The locally private distributed model, where each $Y_i \in \mathcal{Y}$. In the private-coin setting the channels $W_1, \ldots, W_n$ are independent, while in the public-coin setting they are jointly randomized.

Definition 2 (Public-Coin SMP Protocols): In addition to the private randomness $U_1, \ldots, U_n$ at the players as above, let V be a random variable jointly independent of the random variables X_i and U_i, which denotes the public randomness and is available to all players. A ρ-LDP *public-coin* SMP protocol π consists of the following two steps: (a) Player i selects their channel $W_i \in \mathcal{W}_\rho$ as a function of V (and possibly of U_i), and (b) sends their messages $Y_i \in \mathcal{Y}$, by passing X_i through W_i, to the referee. The referee receives the messages $(Y_1, \ldots, Y_n) \coloneqq \pi(X^n, V)$ and the public randomness V, but does not have access to the private randomness $(U_1, \ldots, U_n)$ of the players.

In contrast to private-coin protocols, in a public-coin SMP protocol, the message Y_i from player i is a function of V as well as (X_i, U_i), so the resulting messages Y_i are not independent. They are, however, independent conditioned on the shared randomness V.

We emphasize that private randomness is available even in the public-coin setting and, as previously mentioned, is *required* in order for the protocol to satisfy local privacy (see Eq. 1). This is because the channels must satisfy the LDP condition even when all the information available to the referee, including the public randomness V, is fully "leaked."

C. Distributed Inference Protocols

We now provide the formal description of the distributed inference tasks considered in this work, identity and independence testing.

a) *Identity testing:* Let $\mathbf{q} \in \Delta_k$ be a known reference distribution. In the (k, ε, δ)-identity testing problem, we seek to use n i.i.d. samples from an unknown $\mathbf{p} \in \Delta_k$ to test if $\mathbf{p}$ equals $\mathbf{q}$ or if it is ε-far from $\mathbf{q}$ in total variation distance. A private-coin (resp., public-coin) *ρ-LDP protocol for*

(k, ε, δ)-*identity testing* then consists of a private-coin (resp. public-coin) ρ-LDP protocol π along with a (randomized) mapping $\mathcal{T}\colon \mathcal{Y}^n \to \{0, 1\}$ such that

$$\Pr_{X^n \sim \mathbf{p}^n}\left[\mathcal{T}\left(\pi\left(X^n\right)\right) = 1\right] > 1 - \delta, \quad \text{if } \mathbf{p} = \mathbf{q},$$
$$\Pr_{X^n \sim \mathbf{p}^n}\left[\mathcal{T}\left(\pi\left(X^n\right)\right) = 0\right] > 1 - \delta, \quad \text{if } \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon.$$

Namely, after running the protocol π on the independent samples X^n held by the players, the referee applies the mapping $\mathcal{T}$ to the resulting messages $(Y_1, \ldots, Y_n) = \pi(X^n)$, which should "accept" with high constant probability if the samples come from the reference distribution $\mathbf{q}$ and "reject" with high constant probability if they come from a distribution significantly far from $\mathbf{q}$. The special case of identity testing for $\mathbf{u}_k$ is termed the (k, ε, δ)-*uniformity testing* problem.

The sample complexity of private-coin (resp. public-coin) ρ-LDP (k, ε, δ)-identity testing is the minimum n for which a ρ-LDP protocol for (k, ε, δ)-identity testing with n players exists for $\mathbf{q}$. While this quantity can depend on the reference distribution $\mathbf{q}$, it is customary to consider sample complexity over the worst-case $\mathbf{q}$.[3]

b) *Independence testing:* In the (k, ε, δ)-independence testing problem, we seek to use samples from an unknown $\mathbf{p} \in \Delta_{[k]\times[k]}$ (with unknown marginals $\mathbf{p}_1, \mathbf{p}_2 \in \Delta_k$) to test if $\mathbf{p}$ equals $\mathbf{p}_1 \otimes \mathbf{p}_2$ or if it is ε-far from *every* product distribution in total variation distance. A private-coin (resp., public-coin) ρ-*LDP protocol for* (k, ε, δ)-*independence testing* then consists of a private-coin (resp. public-coin) ρ-LDP protocol π along with a (randomized) mapping $\mathcal{T}\colon \mathcal{Y}^n \to \{0, 1\}$ such that

$$\Pr_{X^n \sim \mathbf{p}^n}\left[\mathcal{T}\left(\pi\left(X^n\right)\right) = 1\right] > 1 - \delta, \quad \text{if } \mathbf{p} = \mathbf{p}_1 \otimes \mathbf{p}_2,$$
$$\Pr_{X^n \sim \mathbf{p}^n}\left[\mathcal{T}(\pi(X^n)) = 0\right] > 1 - \delta, \quad \text{if } \inf_{\mathbf{q}_1, \mathbf{q}_2 \in \Delta_k} \mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}_1 \otimes \mathbf{q}_2) > \varepsilon.$$

The sample complexity of private-coin (resp. public-coin) ρ-LDP (k, ε, δ)-independence testing is the minimum n for which a ρ-LDP protocol for (k, ε, δ)-independence testing with n players exists for $\mathbf{q}$.

Remark 1: We note that the formulation above can be generalized to testing independence over $[k_1] \times [k_2]$ for arbitrary k_1, k_2, or even over general discrete product spaces $[k_1] \times \cdots \times [k_d]$. Some of our protocols may generalize to these more general settings, but for simplicity with focus on the simple and arguably fundamental case of independence over $[k] \times [k]$.

III. Locally Private Identity Testing

We begin by recalling lower bounds from [2] which show that for $\rho \in [0, 1)$, a private-coin protocol ρ-LDP identity testing protocol requires at least $\Omega(k^{3/2}/\rho^2\varepsilon^2)$ players and a public-coin protocol requires at least $\Omega(k/\rho^2\varepsilon^2)$ players. In this section, we propose both private- and public-coin protocols that attain these bounds, establishing a strict separation between the sample complexity of private- and public-coin protocols. In addition, we give protocols with optimal sample complexity for both settings that require only 1 bit of communication per player.

We note that our upper bounds for identity are phrased in terms of the domain size k, or, equivalently, as a worst-case among all possible reference distributions $\mathbf{q}$. However, they immediately imply more refined bounds parameterized by a functional of the reference $\mathbf{q}$ itself (*i.e.*, "instance-optimal" bounds, to follow [46]) *via* the reduction described in [3, Appendix D].

A. Private-Coin Protocols

We now present private-coin protocols based on RAPPOR and Hadamard Response that are both sample-optimal, with different communication requirements.

1) A Mechanism Based on RAPPOR: We begin by describing the *randomized aggregatable privacy-preserving ordinal response* (RAPPOR) mechanism, which is a ρ-LDP mechanism introduced in [29]. Its simplest implementation, k-RAPPOR, maps $\mathcal{X} = [k]$ to $\mathcal{Y} = \{0, 1\}^k$ in two steps. First, "one-hot encoding" is applied to the input $x \in [k]$ to obtain the vector $y' \in \{0, 1\}^k$ such that $y'_j = \mathbb{1}_{\{x=j\}}$ for all $j \in \mathcal{X}$. The privatized output $y \in \mathcal{Y}$ of k-RAPPOR is then a k-bit vector obtained by flipping each bit of y' independently with probability $\frac{1}{e^{\rho/2}+1}$.

Note that if X is drawn from $\mathbf{p} \in \Delta_k$, this leads to $Y \in \{0, 1\}^k$ such that the coordinates are (correlated) Bernoulli random variables, with Y_j distributed as $\mathrm{Bern}(\alpha \cdot \mathbf{p}(j) + \beta)$, $j \in [k]$, with α, β defined as

$$\alpha := \frac{e^{\rho/2} - 1}{e^{\rho/2} + 1} = \frac{\rho}{4} + o(\rho), \quad \beta := \frac{1}{e^{\rho/2} + 1} = \frac{1}{2} + o(\rho). \tag{2}$$

Given n independent samples from $\mathbf{p}$, let the output of RAPPOR applied to these samples be denoted by $Y_1, \ldots, Y_n \in \{0, 1\}^k$, where $Y_i = (Y_{i1}, \ldots, Y_{ik})$ for $i \in [n]$. The following fact is a simple consequence of the definition of RAPPOR.

Fact 1: Let $i, j \in [n]$, and $x, y \in [k]$.

$$\Pr\left[Y_{ix} = 1, Y_{jy} = 1\right] = \begin{cases} (\alpha\mathbf{p}(x) + \beta)(\alpha\mathbf{p}(y) + \beta), & \text{if } i \neq j \\ (\alpha\mathbf{p}(x) + \beta)(\alpha\mathbf{p}(y) + \beta) - \alpha^2\mathbf{p}(x)\mathbf{p}(y), & \text{if } i = j,\ x \neq y \\ \alpha\mathbf{p}(x) + \beta, & \text{if } i = j,\ x = y, \end{cases}$$

where α, β are defined as in (2). Note that vectors Y_i and Y_j are independent for distinct $i, j \in [n]$.

We now propose our testing mechanism based on RAPPOR, which, in essence, uses a privatized version of a χ^2-type statistic of [6], [24], [46]. We note that our choice of using such a χ^2-type statistic instead of a (perhaps more natural) "collision-based" unbiased estimator for $\|\mathbf{p}\|_2^2$ stems from the fact the latter has a high variance, leading to a suboptimal sample complexity. For $x \in [k]$, let the number of occurrences of x among the n (privatized) outputs of RAPPOR be

$$N_x := \sum_{j=1}^{n} \mathbb{1}_{\{Y_{jx}=1\}}, \tag{3}$$

[3]The sample complexity for a fixed $\mathbf{q}$, without privacy constraints, has been studied under the "instance-optimal" setting (see [18], [46]); and under local privacy constraints by [17]. See also Section III for a discussion of the relation between worst-case and instance-optimal settings.

Algorithm 1 Locally Private Identity Testing Using RAPPOR

Require: Privacy parameter $\rho > 0$, distance parameter $\varepsilon \in (0, 1)$, n players

1: Set
$$\alpha \leftarrow \frac{e^{\rho/2} - 1}{e^{\rho/2} + 1}, \quad \beta \leftarrow \frac{1}{e^{\rho/2} + 1}$$
as in (2).
2: Player i applies (ρ-LDP) RAPPOR to X_i, sends result $Y_i \in \{0, 1\}^k$ ▷ Time $O(k)$ per user
3: Server computes N_x for every $x \in [k]$, as defined in (3) ▷ Time $O(kn)$
4: Server computes T, as defined in (4) ▷ Time $O(k)$
5: **if** $T < n(n-1)\alpha^2\varepsilon^2/k$ **then**
6: **return** accept
7: **else**
8: **return** reject

which by the definition of RAPPOR follows a $\mathrm{Bin}(n, \alpha\mathbf{p}(x) + \beta)$ distribution. We consider the following test statistic T:

$$T \coloneqq \sum_{x\in[k]} \Big((N_x - (n-1)(\alpha\mathbf{q}(x) + \beta))^2 - N_x + (n-1)(\alpha\mathbf{q}(x) + \beta)^2 \Big). \tag{4}$$

This statistic is motivated from the fact that it constitutes an unbiased estimator of the squared ℓ_2 distance between $\mathbf{p}$ and $\mathbf{q}$. Using this property we threshold T to test whether $\mathbf{p} = \mathbf{q}$ or not. Keeping in mind that N_x is typically concentrated around its expected value of roughly $n/2$, our new statistic can be seen to take the form

$$T \approx \sum_{x\in[k]} \Big(N_x^2 - nN_x \Big) + \Theta\Big(kn^2\Big),$$

since $\beta \approx 1/2$. In particular, the subtracted linear term reduces the fluctuation of the quadratic part, bringing down the variance of the statistic.

This motivates our testing protocol, Algorithm 1, and leads to the main result of this section below.

Theorem 1: For every $k \geq 1$ and $\rho \in (0, 1]$, there exists a private-coin ρ-LDP protocol for (k, ε, δ)-identity testing over $[k]$ using RAPPOR and $n = O\Big(\frac{k^{3/2}}{\varepsilon^2\rho^2} \log \frac{1}{\delta}\Big)$ players.

Proof: Each player reports its data using RAPPOR, which is a ρ-LDP mechanism. Thus, we only need to analyze the error performance of the proposed test, which we do simply by using Chebyshev's inequality. Towards that, we evaluate the expected value and the variance of T.

The following evaluation of expected value of statistic T uses a simple calculation entailing moments of a Binomial random variable. ■

Lemma 1: For T defined in (4), we have

$$\mathbb{E}[T] = n(n-1)\alpha^2\|\mathbf{p} - \mathbf{q}\|_2^2,$$

where the expectation is taken over the private coins used by RAPPOR and the samples drawn from $\mathbf{p}$. In particular, (i) if $\mathbf{p} = \mathbf{q}$, then $\mathbb{E}[T] = 0$; and (ii) if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$, then $\mathbb{E}[T] > 4n(n-1)\frac{\alpha^2\varepsilon^2}{k}$.

Proof: Letting $\lambda_x \coloneqq \alpha\mathbf{q}(x) + \beta$, $\mu_x \coloneqq \alpha\mathbf{p}(x) + \beta$ for $x \in [k]$, and using the fact that N_x is Binomial with parameters n and μ_x, we have

$$\begin{aligned}
\mathbb{E}[T] &= \sum_{x\in[k]} \mathbb{E}\Big[(N_x - (n-1)\lambda_x)^2 - N_x + (n-1)\lambda_x^2\Big] \\
&= \sum_{x\in[k]} \Big(\mathbb{E}\Big[N_x^2 - N_x\Big] - 2(n-1)\lambda_x\mathbb{E}[N_x] \\
&\qquad + \Big((n-1)^2 + n - 1\Big)\lambda_x^2 \Big) \\
&= \sum_{x\in[k]} \Big(n(n-1)\mu_x^2 - 2n(n-1)\lambda_x\mu_x + n(n-1)\lambda_x^2 \Big) \\
&= \sum_{x\in[k]} n(n-1)(\lambda_x - \mu_x)^2 \\
&= n(n-1)\alpha^2 \sum_{x\in[k]} (\mathbf{p}(x) - \mathbf{q}(x))^2.
\end{aligned}$$

Claim (i) is immediate; claim (ii) follows upon noting that $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_1 \leq \frac{\sqrt{k}}{2}\|\mathbf{p} - \mathbf{q}\|_2$. ■

Turning to the variance, we are able to obtain the following:

Lemma 2: For T defined in (4), we have

$$\mathrm{Var}[T] \leq 2kn^2 + 5n^3\alpha^2\|\mathbf{p} - \mathbf{q}\|_2^2 \leq 2kn^2 + 4n\mathbb{E}[T].$$

The proof of this lemma is technical and relies on the analysis of the covariance of the random variables $(N_x)_{x\in[k]}$, in view of bounding quantities of the form $\mathrm{Cov}(f(N_x), f(N_y))$. We defer the details to Appendix A.

With these two lemmata, we are in a position to conclude the argument.

First, consider the case when $\mathbf{p} = \mathbf{q}$. In this case $\mathbb{E}[T] = 0$ and $\mathrm{Var}[T] \leq 2kn^2$ by Lemmas III.3 and III.4. Therefore, by Chebyshev's inequality we get

$$\Pr\left[T \geq n^2\frac{\alpha^2\varepsilon^2}{k}\right] \leq \frac{k^2\,\mathrm{Var}[T]}{n^4\alpha^4\varepsilon^4} \leq \frac{2k^3}{n^2\alpha^4\varepsilon^4},$$

which is at most $1/3$ for $n \geq \frac{3k^{3/2}}{\alpha^2\varepsilon^2}$.

Next, when $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \varepsilon$, we get

$$\begin{aligned}
\mathbb{E}[T] &> 4\frac{n(n-1)}{k}\alpha^2\varepsilon^2, \\
\mathrm{Var}[T] &\leq 2kn^2 + 4n\mathbb{E}[T].
\end{aligned}$$

Using Chebyshev's inequality yields

$$\begin{aligned}
\Pr\left[T < n^2\frac{\alpha^2\varepsilon^2}{k}\right] &\leq \Pr\left[T < \frac{1}{2}\mathbb{E}[T]\right] \leq \frac{4\,\mathrm{Var}[T]}{\mathbb{E}[T]^2} \\
&\leq \frac{k^3}{2(n-1)^2\alpha^4\varepsilon^4} + \frac{4k}{(n-1)\alpha^2\varepsilon^2},
\end{aligned}$$

which is at most $1/3$ for $n \geq \frac{9k^{3/2}}{\alpha^2\varepsilon^2} + 1$ and $k \geq 2$. Recalling that $\alpha = \Theta(\rho)$ concludes the proof of Theorem 1, for probability of error δ set to $1/3$.

Finally, we can reduce this probability of error to an arbitrary $\delta > 0$, at the cost of a $O(\log(1/\delta))$ factor in the number of players, using a standard "amplification" argument: repeat independently the protocol on $O(\log(1/\delta))$ disjoint sets of players and taking the majority output. ■

2) A Mechanism Based on Hadamard Response: While sample-optimal among private-coin protocols, Algorithm 1 requires each player to communicate k bits. We now present a private-coin protocol that is sample-optimal and requires only 1 bit of communication per player.

Since we seek to send only a 1-bit message per player, each player can simply indicate if its observation lies in a subset or not. To make this communication LDP, we flip this bit with appropriate probability. In fact, we divide n players into K subgroups and associate a subset $C_j \subset [k]$, $1 \le j \le K$, with the jth subgroup. Thus, the bits received at the referee can be viewed as n/K independent samples from a product-Bernoulli distribution on $\{0,1\}^K$.[4]

Suppose that the mean $\mu(\mathbf{p})$ of the resulting product-Bernoulli distribution satisfies $\|\mu(\mathbf{q}) - \mu(\mathbf{p})\|_2 > \alpha$ if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{q}) \ge \varepsilon$. Then, we can use a test for mean of product-Bernoulli distributions (see, for instance, [22, Sec. 2.1] or [23, Lemma 4.2]) to determine if the mean is $\mu(\mathbf{q})$ or α-far from $\mu(\mathbf{q})$ in ℓ_2 distance.

The key question that remains is how large can α be. The answer to this question was provided in [7], which introduced the Hadamard Response (HR) mechanism that uses the Hadamard matrix to select C_js that yield a large α.

Formally, the HR mechanism can be described as follows. Let $K := 2^{\lceil \log_2(k+1) \rceil}$, which is the smallest power of two larger than k, and let $H^{(K)}$ be the $K \times K$ Hadamard matrix. Note that $K \le 2k$. Let C_j be the location of 1s in the jth column, *i.e.*, $C_j = \left\{ i \in [K] : H^{(K)}_{ij} = 1 \right\}$. For any distribution $\mathbf{p}$ over $[k]$ and $C \subset [K]$, let $\mathbf{p}(C)$ be the probability that a sample from $\mathbf{p}$ falls in set C. Here we assign zero probability to elements outside $[k]$. The key property of the sets $(C_1, \ldots, C_K)$, which was observed in [7], is the following.

Lemma 3: For any two distributions $\mathbf{p}, \mathbf{q}$ over $[k]$,

$$\sum_{j=1}^{k} \big(\mathbf{p}(C_j) - \mathbf{q}(C_j)\big)^2 = \frac{K}{4}\|\mathbf{p} - \mathbf{q}\|_2^2.$$

Proof: Let $\mathbf{p}_K, \mathbf{q}_K$ be K-dimensional probability vectors obtained by appending zeros to the end of $\mathbf{p}$ and $\mathbf{q}$, respectively, and let $\mathbf{p}(C) := (\mathbf{p}(C_1), \mathbf{p}(C_2), \ldots, \mathbf{p}(C_K))$. By the definition of $\mathbf{p}(C_j)$s, we have

$$\mathbf{p}(C) = \frac{1}{2}\Big(H^{(K)}\mathbf{p}_K + \mathbf{1}_K\Big), \quad \mathbf{q}(C) = \frac{1}{2}\Big(H^{(K)}\mathbf{q}_K + \mathbf{1}_K\Big),$$

where $\mathbf{1}_K$ is an all-one vector of dimension K. Hence by the fact that $(H^{(K)})^T H^{(K)} = K\mathbb{I}$, we obtain

$$\begin{aligned}\sum_{j=1}^{k} \big(\mathbf{p}(C_j) - \mathbf{q}(C_j)\big)^2 &= \|\mathbf{p}(C) - \mathbf{q}(C)\|_2^2 \\ &= \frac{1}{4}(\mathbf{p}_K - \mathbf{q}_K)^T \Big(H^{(K)}\Big)^T H^{(K)}(\mathbf{p}_K - \mathbf{q}_K) \\ &= \frac{K}{4}\|\mathbf{p}_K - \mathbf{q}_K\|_2^2 = \frac{K}{4}\|\mathbf{p} - \mathbf{q}\|_2^2.\end{aligned}$$

■

[4] A K-dimensional product-Bernoulli distribution is a distribution over $\{0,1\}^K$, whose coordinates are independently distributed.

Algorithm 2 Locally Private Identity Testing Using Hadamard Response

Require: Privacy parameter $\rho > 0$, distance parameter $\varepsilon \in (0,1)$, n players

1: Define $C_j = \left\{ i \in [K] : H^{(K)}_{ij} = 1 \right\}$, $j \in [K]$.
2: n players are divided into K disjoint subgroups of equal size (using an explicit partition fixed ahead of time). Players in the jth subgroup, $j \in [K]$, are assigned to the set C_j, and they use (5) to generate their output bits (independent copies of B_j).
3: Taking one player from each block and viewing the resulting collection of messages as a length-K binary vector, the referee gets n/K independent copies of $(B_1, B_2, \ldots, B_K)$ generated a product-Bernoulli distribution on $\{0,1\}^K$ with mean vector $\mu(\mathbf{p})$.
4: The referee uses these n/K samples to test whether the mean vector $\mu(\mathbf{p})$ is (i) a prespecified vector $\mu = \mu(\mathbf{q}) \in \mathbb{R}^K$ or (ii) at ℓ_2 distance at least $\alpha = \varepsilon/2 \in (0,1]$ from $\mu(\mathbf{q})$. It can use the test from, for instance, [22, Sec. 2.1], which requires $O\Big(\sqrt{K}(\log 1/\delta)/\alpha^2\Big)$ samples to do this. It accepts $\mathbf{q}$ if the mean is $\mu(\mathbf{q})$, and rejects otherwise.

In HR, a player observing $X \in [k]$ assigned a subset C_j sends a random bit B_j with distribution given by

$$\Pr\big[B_j = 1 | X\big] = \begin{cases} \frac{e^\rho}{e^\rho+1}, & \text{if } X \in C_j, \\ \frac{1}{e^\rho+1}, & \text{otherwise.} \end{cases} \tag{5}$$

Let $\mu(\mathbf{p})$ denote the mean of the product-Bernoulli distribution induced on bits $(B_1, \ldots, B_K)$ (corresponding to any K players assigned sets $(C_1, \ldots, C_K)$) when the observations of players have distribution $\mathbf{p}$, *i.e.*,

$$\mu(\mathbf{p})_j := \mathbb{E}_{\mathbf{p}}\big[\Pr\big[B_j = 1|X\big]\big].$$

Following the same computations as in [7], we have that for all $j \in [K]$

$$\begin{aligned}\mu(\mathbf{p})_j &= \sum_{x \in C_j} \mathbf{p}(x)\frac{e^\rho}{e^\rho+1} + \sum_{x \notin C_j} \mathbf{p}(x)\frac{1}{e^\rho+1} = \frac{e^\rho - 1}{e^\rho+1}\mathbf{p}\big(C_j\big) \\ &\quad + \frac{1}{e^\rho+1}.\end{aligned}$$

Then, by Lemma 3,

$$\begin{aligned}\|\mu(\mathbf{p}) - \mu(\mathbf{q})\|_2 &= \frac{\sqrt{K}(e^\rho - 1)}{2(e^\rho+1)}\|\mathbf{p} - \mathbf{q}\|_2 \\ &\ge \frac{(e^\rho - 1)}{2(e^\rho+1)}\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{q}),\end{aligned} \tag{6}$$

where we used the observation that $K \ge k$.

Motivated by this observation, we obtain Algorithm 2 for LDP identity testing.[5] The result below summarizes the performance of Algorithm 2.

Theorem 2: For every $k \ge 1$ and $\rho \in (0,1]$, there exists a private-coin ρ-LDP protocol for (k,ε,δ)-identity testing using one bit of communication per player and $n = O\Big(\frac{k^{3/2}}{\varepsilon^2\rho^2}\log\frac{1}{\delta}\Big)$ players.

Proof: We have already outline the proof in the discussion above. It is easy to check that the mechanism in (5)

[5] Without loss of generality, we assume K divides n (as otherwise we can ignore the last $(n - K\lfloor \frac{n}{K} \rfloor)$ players without changing the number of samples by a factor of 2).

is ρ-LDP. Further, by (6), the test in [22, Sec. 2.1] gives the correct outcome with probability of error less than δ if $n/K \gtrsim \sqrt{K}\log(1/\delta)/\alpha^2$ with $\alpha = \varepsilon/2$, *i.e.*,

$$n = O\left(k^{3/2}\frac{(e^\rho+1)^2}{(e^\rho-1)^2\varepsilon^2}\log\frac{1}{\delta}\right) = O\left(\frac{k^{3/2}}{\varepsilon^2\rho^2}\log\frac{1}{\delta}\right),$$

suffices as claimed. ∎

Remark 2: From the proof of Theorem 2, it is clear that the protocol provides a stronger, ℓ_2, guarantee: it allows one to distinguish with probability $1-\delta$ between $\|\mathbf{p}-\mathbf{q}\|_2^2 \le \frac{\varepsilon^2}{k}$ and $\|\mathbf{p}-\mathbf{q}\|_2^2 \ge \frac{4\varepsilon^2}{k}$ with $n = O\left(\frac{k^{3/2}}{\varepsilon^2\rho^2}\log\frac{1}{\delta}\right)$ players (by Cauchy–Schwarz, this implies the total variation testing guarantee). Moreover, the protocol does not require the players to have knowledge of the reference distribution $\mathbf{q}$; it is sufficient that the referee knows it. Both these points are useful, later, for our independence testing results.

B. Public-Coin Protocols

The HR based identity testing protocol generates samples from a product-Bernoulli distribution by assigning different subsets to different subgroups of players. Specifically, we found subsets such that, for any two distributions $\mathbf{p}$ and $\mathbf{q}$, the ℓ_2 distance between the means of the induced approximately k-dimensional product distributions is roughly equal to the ℓ_2 distance between $\mathbf{p}$ and $\mathbf{q}$.

Interestingly, we can interpret Lemma 3 to get that for I distributed uniformly over $[K]$, $\mathbb{E}\left[(\mathbf{p}(C_I)-\mathbf{q}(C_I))^2\right] \ge \|\mathbf{p}-\mathbf{q}\|_2^2/4$. This suggests the possibility of finding a random subset S such that $(\mathbf{p}(S)-\mathbf{q}(S))^2 \gtrsim \varepsilon^2/k$ if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{q}) \ge \varepsilon$. Such a set is very handy: We can simply implement a version of Algorithm 2 with $K=1$ using this set and get a test that works with roughly $k/(\rho^2\varepsilon^2)$ samples. This saving in sample-complexity arises from the fact that we were able to retain the same "per dimension" ℓ_2 distance as that using HR, while using much smaller (only one) dimensional observations. But the players need to use public coins to share this set S. We formalize this protocol in this section.

The first component of our protocol is the following lemma from [3], specialized to a target domain of size 2.

Theorem 3 [3, Th. VI.2]: Fix any k-ary distributions $\mathbf{p},\mathbf{q}$. If $S \subseteq [k]$ is a set chosen uniformly at random, we have the following. (i) if $\mathbf{p}=\mathbf{q}$, then $\mathbf{p}(S)=\mathbf{q}(S)$ with probability one; and (ii) if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p},\mathbf{q}) > \varepsilon$, then

$$\Pr_S\left[(\mathbf{p}(S)-\mathbf{q}(S))^2 > \frac{\varepsilon^2}{2k}\right] \ge c.$$

where $c = 1/228$.

Thus, indeed, we can find our desired random set S.

Next, we present an LDP protocol for testing the bias of coins, our LDP identity testing problem for $k=2$. The protocol below can be viewed as a special case of our protocol in Section III-A2; we include this simpler result here for completeness. We have the following.

Lemma 4 (Locally Private Bias Estimation, Warmup): For every $\rho \in (0,1]$, there exists a private-coin ρ-LDP protocol for $(2,\varepsilon,\delta)$-identity testing using one bit of communication per player and $n = O\left(\frac{1}{\varepsilon^2\rho^2}\log\frac{1}{\delta}\right)$ players. Moreover, the players do not need to know the reference distribution.

Proof: Assume without loss of generality that the reference distribution is $\mathbf{q} = \mathrm{Bern}(q)$. The algorithm uses a simple Randomized Response (RR) scheme [48], where each sample is flipped with probability $1/(e^\rho+1)$. When the input is $\mathrm{Bern}(p)$, the output distribution is $\mathrm{Bern}((1+p(e^\rho-1)/(e^\rho+1)))$. Therefore, if $p-q>\varepsilon$, then the bias of the output distribution of applying RR to $\mathrm{Bern}(p)$, and $\mathrm{Bern}(q)$ differ by $(p-q)(e^\rho-1))/(e^\rho+1)$, which is $\Omega(\varepsilon\rho)$ for $\rho = O(1)$. To distinguish these two Bernoulli distributions with a constant probability, $O(1/(\rho^2\varepsilon^2))$ samples suffice, and the success probability can be boosted to $1-\delta$ by repeating $O(\log(1/\delta))$ times. ∎

Motivated by these observations, we propose Algorithm 3 for public-coin LDP identity testing.

Algorithm 3 Locally Private Identity Testing

Require: Privacy parameter $\rho > 0$, distance parameter $\varepsilon \in (0,1)$, n players

Set

$$c \leftarrow \frac{1}{288},\ \delta_0 \leftarrow \frac{c}{2(1+c)},\ \varepsilon' \leftarrow \frac{\varepsilon}{\sqrt{2k}},\ T = \Theta(1), \qquad m \leftarrow \frac{n}{T}.$$

Partition the players in T subgroups $G_1,\dots,G_T$ of m players
for t from 1 to T **do** ▷ In parallel
    Players in G_t generate uniformly at random a common subset $S_t \subseteq [k]$.
    for all $i \in G_t$ **do**
        Player i converts their sample X_i to $X_i' \coloneqq \mathbb{1}_{\{X_i \in S_t\}}$.
    Players in G_t (and the referee) run the protocol from Lemma 4 on the samples $(X_i')_{i\in G_t}$ to test identity of $\mathbf{p}(S_t)$ to $\mathbf{q}(S_t)$, with distance parameter ε' and failure probability δ_0
▷ At the referee
Let τ denote the fraction of the T protocols that returned accept
if $\tau > 1-(\delta_0+\frac{c}{4})$ **then**
    return accept
else
    return reject

Motivated by these observations, we propose Algorithm 3 for public-coin LDP identity testing.

We close this section with a characterization of performance of our proposed algorithm.

Theorem 4: For every $k \ge 1$ and $\rho \in (0,1]$, there exists a public-coin ρ-LDP protocol for (k,ε,δ)-identity testing using one bit of communication per player and $n = O\left(\frac{k}{\varepsilon^2\rho^2}\log\frac{1}{\delta}\right)$ players.

Proof: The proof of correctness follows the foregoing outline, which we describe in more detail. Let $c \coloneqq 1/288$ be the constant from Theorem 3, let $\delta_0 \coloneqq \frac{c}{2(1+c)} = 1/458$, and set $\varepsilon' \coloneqq \frac{\varepsilon}{\sqrt{2k}}$.

Consider the t-th test from Algorithm 3 (where $1 \le t \le T$), and let b_t be the indicator that the protocol run by players in G_t returned accept. If $\mathbf{p}=\mathbf{q}$, then by the above we have $\Pr[b_t=1] \ge 1-\delta_0$ (where the probability is over the choice of the random subset S_t, and the randomness of protocol from Lemma 4). However, if $\mathbf{p}$ is ε-far from $\mathbf{q}$, by Theorem 3 it the case that $\Pr[b_t=1] \le (1-c)+c\delta_0 = 1-(\delta_0+\frac{c}{2})$. Therefore, for a sufficiently large constant in the choice of $T = \Theta(1/c^2) = \Theta(1)$, a Chernoff bound argument ensures that we

Algorithm 4 Locally Private Independence Testing (Private-Coin)

Require: Privacy parameter $\rho > 0$, distance parameter $\varepsilon \in (0, 1)$, $n = O\left(\frac{k^3}{\varepsilon^2\rho^2}\right)$ players

1: Partition the players in two groups, L ("learning") and T ("testing"), each of size $\frac{n}{2}$.
2: Players in group L run a ρ-LDP *learning* protocol to estimate $\mathbf{p}_1 \otimes \mathbf{p}_2$ in ℓ_2 distance, obtaining $\widehat{\mathbf{p}}_1 \otimes \widehat{\mathbf{p}}_2$ such that $\|\widehat{\mathbf{p}}_1 \otimes \widehat{\mathbf{p}}_2 - \mathbf{p}_1 \otimes \mathbf{p}_2\|_2^2 \le \frac{\varepsilon^2}{2k^2}$ (using the protocol of Lemma 5).
3: Players in group T run a ρ-LDP *identity testing* protocol on $\mathbf{p}$, to distinguish between $\|\mathbf{p} - \widehat{\mathbf{p}}_1 \otimes \widehat{\mathbf{p}}_2\|_2^2 \le \frac{\varepsilon^2}{2k^2}$ and $\|\mathbf{p} - \widehat{\mathbf{p}}_1 \otimes \widehat{\mathbf{p}}_2\|_2^2 \ge \frac{2\varepsilon^2}{k^2}$ (using the protocol of Theorem 2).

can distinguish between these two cases with probability at least 2/3. ∎

IV. Locally Private Independence Testing

In this section, we establish the sample complexity of testing independence of discrete distrbutions. We present private-coin and public-coin protocols for LDP independence testing that require $\Omega\left(\frac{k^3}{\varepsilon^2\rho^2}\right)$ and $\Omega\left(\frac{k^2}{\varepsilon^2\rho^2}\right)$ players, respectively. In fact, we show matching lower bounds for these sample complexities in the final subsection, establishing their optimality among private-coin and public-coin protocols, respectively. The lower bound is a consequence of a general reduction between independence and uniformity testing, which may be of independent interest.

A. Private-Coin Protocols

To design a private-coin LDP independence testing protocol using $O\left(\frac{k^3}{\rho^2\varepsilon^2}\right)$ players, the first observation we make is that we can find a product distribution $\widehat{\mathbf{p}}$ that is ε/k-close in ℓ_2 distance from the product distribution $\mathbf{p}_1 \times \mathbf{p}_2$ using $O\left(\frac{k^3}{\rho^2\varepsilon^2}\right)$ players. When the generating distribution $\mathbf{p}$ is not a product distribution, from the separation between our hypothesis, we know that $\mathbf{p}$ must have ℓ_2 distance exceeding ε/k from the product distribution $\widehat{\mathbf{p}}$ we find which is close to $\mathbf{p}_1 \times \mathbf{p}_2$. After this point, treating $\widehat{\mathbf{p}}$ as the reference, we can use an private-coin LDP identity testing protocol to test if the samples are generated from a distribution that is close to the (product) reference distribution (in ℓ_2 distance) or far from it.

Formally, we describe the algorithm in Algorithm 4, and present its performance in Theorem 5.

Theorem 5: For every $k \ge 1$ and $\rho \in (0, 1]$, there exists a private-coin ρ-LDP protocol for (k, ε, δ)-independence testing using one bit of communication per player and $n = O\left(\frac{k^3}{\varepsilon^2\rho^2}\log\frac{1}{\delta}\right)$ players, where $\varepsilon \in (0, 1]$ is the distance parameter.

Proof: We first note that Algorithm 4 can be implemented in the SMP setting. Recall that the protocol of Theorem 2 does not require the players to know the reference distribution, and therefore the protocol can be performed in the SMP setting, where players all send their messages simultaneously to the referee. Indeed, in our case, this reference distribution is the product distribution $\widehat{\mathbf{p}}_1 \otimes \widehat{\mathbf{p}}_2$ computed from the messages of the players in L, so the fact that the players' messages (from the group T) do not require knowledge of the reference distribution is crucial to obtain an SMP protocol.

Lemma 5: Given samples from a distribution $\mathbf{p}$ over $[k]\times[k]$ with marginals $\mathbf{p}_1$ and $\mathbf{p}_2$, there exists a private-coin ρ-LDP protocol with $O\left(\frac{k^3}{\rho^2\varepsilon^2}\right)$ players that outputs distributions $\widehat{\mathbf{p}}_1$, $\widehat{\mathbf{p}}_2$ such that $\|\widehat{\mathbf{p}}_1 \otimes \widehat{\mathbf{p}}_2 - \mathbf{p}_1 \otimes \mathbf{p}_2\|_2^2 \le \frac{\varepsilon^2}{2k^2}$ with probability at least 5/6. Moreover, each player sends one bit.

Proof: From the known results on LDP distribution estimation [7], [9], [37], with $O\left(\frac{k}{\rho^2(\varepsilon/k)^2}\right) = O\left(\frac{k^3}{\rho^2\varepsilon^2}\right)$ players one can under ρ-LDP output distributions $\widehat{\mathbf{p}}_1$, $\widehat{\mathbf{p}}_2$ such that

$$\|\widehat{\mathbf{p}}_1 - \mathbf{p}_1\|_2^2 \le \frac{\varepsilon^2}{8k^2}, \quad \|\widehat{\mathbf{p}}_2 - \mathbf{p}_2\|_2^2 \le \frac{\varepsilon^2}{8k^2}$$

with probability at least 5/6. Whenever this guarantee holds, it implies that

$$\begin{aligned}\|\widehat{\mathbf{p}}_1 \otimes \widehat{\mathbf{p}}_2 - \mathbf{p}_1 \otimes \mathbf{p}_2\|_2^2 &\le 2 \cdot \|\widehat{\mathbf{p}}_1 \otimes \mathbf{p}_2 - \mathbf{p}_1 \otimes \mathbf{p}_2\|_2^2 \\ &\quad + 2 \cdot \|\widehat{\mathbf{p}}_1 \otimes \widehat{\mathbf{p}}_2 - \widehat{\mathbf{p}}_1 \otimes \mathbf{p}_2\|_2^2 \\ &\le 2\left(\|\widehat{\mathbf{p}}_1 - \mathbf{p}_1\|_2^2 + \|\widehat{\mathbf{p}}_2 - \mathbf{p}_2\|_2^2\right) \le \frac{\varepsilon^2}{2k^2}\end{aligned}$$

proving the lemma. The bound on the per-player communication follows from the protocol of [7]. ∎

Using the protocol from Lemma 5, we get the following with probability 5/6. If $\mathbf{p}$ is a product distribution with marginals $\mathbf{p}_1$ and $\mathbf{p}_2$, then

$$\|\widehat{\mathbf{p}}_1 \otimes \widehat{\mathbf{p}}_2 - \mathbf{p}\|_2^2 = \|\widehat{\mathbf{p}}_1 \otimes \widehat{\mathbf{p}}_2 - \mathbf{p}_1 \otimes \mathbf{p}_2\|_2^2 \le \frac{\varepsilon^2}{2k^2}.$$

If however $\mathbf{p}$ is ε-far from being a product distribution, then, by the Cauchy–Schwarz inequality,

$$\|\mathbf{p} - \widehat{\mathbf{p}}_1 \otimes \widehat{\mathbf{p}}_2\|_2^2 \ge \frac{4}{k^2}\|\mathbf{p} - \widehat{\mathbf{p}}_1 \otimes \widehat{\mathbf{p}}_2\|_1^2 > 4\frac{\varepsilon^2}{k^2}.$$

We can use the protocol from Theorem 2 (specifically, recalling Remark 2) to distinguish the two cases with $O\left(\frac{(k^2)^{3/2}}{\rho^2\varepsilon^2}\right) = O\left(\frac{k^3}{\rho^2\varepsilon^2}\right)$ players, and probability of success 5/6. By a union bound over the two protocols used, the overall tester is successful with probability at least 2/3. Amplifying the probability of success to $1 - \delta$ by running the protocol in parallel on $O(\log(1/\delta))$ disjoint sets of players and taking the majority output yields the result. ∎

B. Public-Coin Protocols

We now present our public-coin protocol for LDP independence testing. Our approach is similar to the one we followed for our public-coin LDP identity testing protocol: namely, we first use public coins to "embed" the problem in a smaller domain of size $k = 2$, and then apply an LDP independence test for $k = 2$. For this strategy to work, we first need a result guaranteeing that randomly hashing the domain $[k]\times[k]$ to $\{0, 1\} \times \{0, 1\}$ while respecting the product structure preserves distances. This is what we provide next, establishing an analogue of Theorem 3 tailored to the product space setting.

Theorem 6: Fix any distribution $\mathbf{p}$ over $[k] \times [k]$ with marginals $\mathbf{p}_1, \mathbf{p}_2$. If $S_1, S_2 \subseteq [k]$ are two sets chosen independently and uniformly at random, we have the following. (i) if $\mathbf{p} = \mathbf{p}_1 \otimes \mathbf{p}_2$, then $\mathbf{p}(S_1 \times S_2) = \mathbf{p}_1(S_1)\mathbf{p}_2(S_2)$ with probability one; and (ii) if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{p}_1 \otimes \mathbf{p}_2) > \varepsilon$, then

$$\Pr_{S_1,S_2}\left[(\mathbf{p}(S_1 \times S_2) - \mathbf{p}_1(S_1)\mathbf{p}_2(S_2))^2 > \frac{\varepsilon^2}{8k}\right] \geq c.$$

for some absolute constant $c > 0$. (Moreover, one can take $c = 1/4096$.)

We emphasize that Theorem 6 is not a direct consequence of Theorem 3, due to the product structure of the random subset $S_1 \times S_2$ (while the previous theorem would apply to a random subset $S \subseteq [k] \times [k]$). And indeed, proving Theorem 6 requires the following hashing lemma, proven in a fashion similar to [3, Th. A.6]:

Theorem 7 (Joint Probability Perturbation Hashing): Consider a matrix $\delta \in \mathbb{R}^{k \times k}$ such that, for every $i_0, j_0 \in [k]$, $\sum_{j \in [k]} \delta_{i_0,j} = \sum_{i \in [k]} \delta_{i,j_0} = 0$. Let random variables $X = (X_1, \ldots, X_k)$ and $Y = (Y_1, \ldots, Y_k)$ be independent and uniformly distributed over k-length binary sequences. Define $Z = \sum_{(i,j) \in [k] \times [k]} \delta_{ij} X_i Y_j$. Then, for every $\alpha \in (0, 1/16)$, there exists a constant $c_\alpha > 0$ such that

$$\Pr\left[Z^2 \geq \alpha \|\delta\|_F^2\right] \geq c_\alpha.$$

The proof of this theorem is deferred to Appendix B. We now show how this implies Theorem 6.

Proof of Theorem 6: Let $\mathbf{p}$ be as in the statement. Item (i) is from the definition. We just focus on proving item (ii). Define $\delta \in \mathbb{R}^{k \times k}$ by $\delta_{ij} = \mathbf{p}(i,j) - \mathbf{p}_1(i)\mathbf{p}_2(j)$ for $i, j \in [k]$. Since $\mathbf{p}$ has marginals $\mathbf{p}_1, \mathbf{p}_2$, δ satisfies the assumptions of Theorem 7, we can apply the theorem, observing that if X (resp. Y) is the indicator vector of the set S_1 (resp. S_2) then

$$\begin{aligned} Z &= \sum_{(i,j) \in [k] \times [k]} (\mathbf{p}(i,j) - \mathbf{p}_1(i)\mathbf{p}_2(j)) X_i Y_j \\ &= \mathbf{p}(S_1 \times S_2) - \mathbf{p}_1(S_1)\mathbf{p}_2(S_2), \end{aligned}$$

and that $\|\delta\|_F^2 = \|\mathbf{p} - \mathbf{p}_1 \otimes \mathbf{p}_2\|_2^2 \geq \frac{4\varepsilon^2}{k^2}$ (the inequality being Cauchy–Schwarz). Taking $\alpha = 1/32$ yields the result. ∎

It only remains to describe an LDP independence testing protocol for $k = 2$. Note that while we can set $k = 2$ in the protocol of Theorem 5, it leads to a complicated protocol. We instead provide a simple test for $k = 2$.

Lemma 6 (Locally Private Bias Estimation): Let $\rho \in (0, 1]$. There exists a private-coin ρ-LDP protocol for $(2, 2, \varepsilon, \delta)$-independence testing using one bit of communication per player and $n = O\left(\frac{1}{\varepsilon^2 \rho^2} \log \frac{1}{\delta}\right)$ players.

Proof: Consider a distribution $\mathbf{p}$ over $\{0, 1\} \times \{0, 1\}$ with marginals $\mathbf{p}_1$ and $\mathbf{p}_2$. We use the fact that

$$|\mathbf{p}(0,0) - \mathbf{p}_1(0)\mathbf{p}_2(0)| = |\mathbf{p}(x,y) - \mathbf{p}_1(x)\mathbf{p}_2(y)|, \quad x, y \in \{0, 1\},$$

which holds since

$$\begin{aligned} |\mathbf{p}(0,1) - \mathbf{p}_1(0)\mathbf{p}_2(1)| &= |(\mathbf{p}_1(0) - \mathbf{p}(0,0)) - \mathbf{p}_1(0)(1 - \mathbf{p}_2(0))| \\ &= |\mathbf{p}_1(0)\mathbf{p}_2(0) - \mathbf{p}(0,0)|. \end{aligned}$$

Thus, if $\mathbf{p}$ is ε-far in total variation distance from every product distribution, it must hold that $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{p}_1 \otimes \mathbf{p}_2) \geq \varepsilon$, which in view of the equation above yields $|\mathbf{p}(0,0) - \mathbf{p}_1(0)\mathbf{p}_2(0)| \geq \varepsilon/2$. Using this observation, we can test for independence using $O(1/(\rho^2\varepsilon^2))$ samples as follows.

The n players are partitioned in 3 sets A, B, C of size $n/3$. Since, for any symbol (x, y), $\mathbf{p}(x, y)$ (resp. $\mathbf{p}_1(x)$, $\mathbf{p}_2(y)$) can be estimated up to accuracy ε by converting the observation (X, Y) to the binary observation $\mathbb{1}_{\{(X,Y)=(x,y)\}}$ (resp. $\mathbb{1}_{\{X=x\}}$, $\mathbb{1}_{\{Y=y\}}$) and proceeding as in Lemma 4, we can estimate $\mathbf{p}(0,0)$, $\mathbf{p}_1(0)$, and $\mathbf{p}_2(0)$ up to an additive accuracy $\varepsilon/16$ by assigning $|A| = |B| = |C| = O(1/(\rho^2\varepsilon^2)\log(1/\delta))$ players for each of them, so that the three estimates are simultaneously accurate with probability at least $1 - \delta$. Denote these estimates by $\tilde{\mathbf{p}}(0,0)$, $\tilde{\mathbf{p}}_1(0)$, and $\tilde{\mathbf{p}}_2(0)$, respectively. When $\mathbf{p}(0,0) = \mathbf{p}_1(0)\mathbf{p}_2(0)$,

$$\begin{aligned} \left|\tilde{\mathbf{p}}(0,0) - \tilde{\mathbf{p}}_1(0)\tilde{\mathbf{p}}_2(0)\right| &\leq \left|\tilde{\mathbf{p}}(0,0) - \mathbf{p}(0,0)\right| + \left|\tilde{\mathbf{p}}_1(0) - \mathbf{p}_1(0)\right| \\ &\quad + \left|\tilde{\mathbf{p}}_2(0) - \mathbf{p}_2(0)\right| \leq \frac{3}{16}\varepsilon. \end{aligned}$$

On the other hand, when $|\mathbf{p}(0,0) - \mathbf{p}_1(0)\mathbf{p}_2(0)| \geq \varepsilon/2$, we have

$$\begin{aligned} &\left|\tilde{\mathbf{p}}(0,0) - \tilde{\mathbf{p}}_1(0)\tilde{\mathbf{p}}_2(0)\right| \\ &\quad \geq |\mathbf{p}(0,0) - \mathbf{p}_1(0)\mathbf{p}_2(0)| - \left|\tilde{\mathbf{p}}(0,0) - \mathbf{p}(0,0)\right| \\ &\qquad - \left|\tilde{\mathbf{p}}_1(0) - \mathbf{p}_1(0)\right| - \left|\tilde{\mathbf{p}}_2(0) - \mathbf{p}_2(0)\right| \geq \frac{5}{16}\varepsilon. \end{aligned}$$

Thus, it is sufficient for the referee to form the estimates $\tilde{\mathbf{p}}(0,0)$, $\tilde{\mathbf{p}}_1(0)$, and $\tilde{\mathbf{p}}_2(0)$ and compare $\left|\tilde{\mathbf{p}}(0,0) - \tilde{\mathbf{p}}_1(0)\tilde{\mathbf{p}}_2(0)\right|$ to the threshold $\varepsilon/4$. ∎

We summarize the overall algorithm and its performance below.

Theorem 8: For every $k \geq 1$ and $\rho \in (0, 1]$, there exists a public-coin ρ-LDP protocol for (k, ε, δ)-independence testing using one bit of communication per player and $n = O\left(\frac{k^2}{\varepsilon^2\rho^2} \log \frac{1}{\delta}\right)$ players.

Proof: The proof proceeds as follows: Using the public randomness, the players select two uniformly random subsets $S_1, S_2 \subseteq [k]$, and from their samples allow the referee to estimate the quantities $\mathbf{p}(S_1 \times S_2)$, $\mathbf{p}_1(S_1)$, and $\mathbf{p}_2(S_2)$. By Theorem 6, this in turn is enough to detect (with constant probability over the choice of S_1, S_2) if $\mathbf{p}$ is far from $\mathbf{p}_1 \otimes \mathbf{p}_2$; it then suffices to repeat this in parallel on disjoint groups of players in order to amplify the probability of success.

To wit, the proof of correctness follows the foregoing outline, which we describe in more detail. Let $c := 1/4096$ be the constant from Theorem 6, let $\delta_0 := \frac{c}{2(1+c)}$, and set $\varepsilon' := \frac{\varepsilon}{\sqrt{8k}}$.

Consider the t-th test from Algorithm 5 (where $1 \leq t \leq T$), and let b_t be the indicator that the protocol run by players in B_t returned accept. If $\mathbf{p} = \mathbf{p}_1 \otimes \mathbf{p}_2$, then by the above we have $\Pr[b_t = 1] \geq 1 - \delta_0$ (where the probability is over the choice of the random subsets $S_{t,1}$ and $S_{t,2}$, and the randomness of protocol from Lemma 6). However, if $\mathbf{p}$ is ε-far from $\mathbf{p}_1 \otimes \mathbf{p}_2$, by Theorem 6 it the case that $\Pr[b_t = 1] \leq (1 - c) + c\delta_0 = 1 - (\delta_0 + \frac{c}{2})$. Therefore, for a sufficiently large constant in the choice of $T = \Theta(1/c^2) = \Theta(1)$, a Chernoff bound argument ensures that we can distinguish between these two cases with probability at least $2/3$. ∎

Algorithm 5 Locally Private Independence Testing (Public-Coin)

Require: Privacy parameter $\rho > 0$, distance parameter $\varepsilon \in (0, 1)$, n players

Set

$$c \leftarrow \frac{1}{4096},\ \delta_0 \leftarrow \frac{c}{2(1+c)},\ \varepsilon' \leftarrow \frac{\varepsilon}{\sqrt{8k}},\ T = \Theta(1),\ m \leftarrow \frac{n}{3T}.$$

Partition the players in $3T$ groups $B_{1,1}, B_{1,2}, B_{1,3}, B_{2,1}, B_{2,2}, B_{2,3}, \ldots, B_{T,1}, B_{T,2}, B_{T,3}$ of m players

for t from 1 to T **do** ▷ In parallel

  Players in $B_{t,1} \cup B_{t,2} \cup B_{t,3}$ generate uniformly at random two common subsets $S_{1,t}, S_{2,t} \subseteq [k]$.

  for all $i \in B_{t,1}$ **do**

    Player i converts their sample (X_i, Y_i) to $X'_i := \mathbb{1}_{\{(X_i,Y_i)\in S_{t,1}\times S_{t,2}\}}$.

  for all $i \in B_{t,2}$ **do**

    Player i converts their sample (X_i, Y_i) to $X'_i := \mathbb{1}_{\{X_i\in S_{t,1}\}}$.

  for all $i \in B_{t,3}$ **do**

    Player i converts their sample (X_i, Y_i) to $X'_i := \mathbb{1}_{\{Y_i\in S_{t,2}\}}$.

  Players in $B_{t,1}\cup B_{t,2}\cup B_{t,3}$ (and the referee) run the protocol from Lemma 4 on the samples $(X'_i)_{i\in B_{t,1}\cup B_{t,2}\cup B_{t,3}}$ to test identity of $\mathbf{p}(S_{t,1} \times S_{t,2})$ to $\mathbf{p}_1(S_{t,1})\mathbf{p}_1(S_{t,2})$, with distance parameter ε' and failure probability δ_0

▷ At the referee

Let τ denote the fraction of the T protocols that returned accept

if $\tau > 1 - (\delta_0 + \frac{c}{4})$ **then**

  return accept

else

  return reject

C. Lower Bounds

The following theorem proves the tightness of our upper bounds for independence testing.

Theorem 9: For every $k \geq 1$ and $\rho \in (0, 1]$, every private-coin (resp., public-coin) ρ-LDP protocol for $(k, \varepsilon, 1/12)$-independence testing must have $\Omega\left(\frac{k^3}{\varepsilon^2\rho^2}\right)$ players (resp., $\Omega\left(\frac{k^2}{\varepsilon^2\rho^2}\right)$ players).

Proof: We show the following reduction, which implies our bounds for independence testing. If there exists a private-coin (resp., public-coin) ρ-LDP protocol for $(k, \varepsilon, 1/12)$-independence testing with n players, then there also exists a private-coin (resp., public-coin) ρ-LDP protocol for distinguishing the "Paninski construction" over $[k^2]$ with n players. Recall that for even integer k and a distance parameter $\gamma \in [0, 1/2]$, the Paninski construction is a family of $2^{k^2/2}$ distributions $\{\mathbf{p}_z\}_{z\in\{-1,+1\}^{k^2/2}}$ over $[k^2]$, where for $z \in \{-1, +1\}^{k^2/2}$ we have

$$\mathbf{p}_z(x) = \begin{cases} \frac{1-2\gamma z_i}{k^2}, & x = 2i-1 \\ \frac{1+2\gamma z_i}{k^2}, & x = 2i \end{cases}, \quad x \in [k^2]. \tag{7}$$

Note that every $\mathbf{p}_z$ is then at total variation distance exactly γ from $\mathbf{u}_{k^2}$. From the lower bounds on uniformity testing already established in [2] (listed in Table I), we then obtain the lower bounds on independence testing.

We first state the following useful fact (see, e.g., [16]) which states that if a distribution is close to a product distribution, then it must be close to the product of its own marginals.

Fact 2: Let $\mathbf{p}, \mathbf{q} \in \Delta_{\Omega\times\Omega}$ with $\mathbf{q}$, a product distribution. If $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) \leq \varepsilon$ then $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{p}_1 \otimes \mathbf{p}_2) \leq 3\varepsilon$.

Let $k = 2\ell$. For $z \in \{-1, +1\}^{2\ell^2}$, let $(\mathbf{p}_z)_z$ be the collection of distributions over $[4\ell^2] = [k^2]$ given in (7), each at a distance $\gamma := 3\varepsilon$ from the uniform distribution. We construct a mapping $\Phi\colon \Delta_{[k^2]} \to \Delta_{[2k]\times[2k]}$ such that:

1) Both marginals of $\Phi(\mathbf{p}_z)$ are $\mathbf{u}_{[2k]}$ for all z;
2) $\mathrm{d}_{\mathrm{TV}}\big(\Phi(\mathbf{p}_z), \mathbf{u}_{[2k]\times[2k]}\big) = \mathrm{d}_{\mathrm{TV}}\big(\mathbf{p}_z, \mathbf{u}_{k^2}\big)$, and $\Phi(\mathbf{u}_{k^2}) = \mathbf{u}_{[2k]\times[2k]}$;
3) There exists a mapping from $[k^2] \to [2k] \times [2k]$ that converts a sample from $\mathbf{p}_z$ into a sample from $\Phi(\mathbf{p}_z)$, and a sample from $\mathbf{u}_{[k^2]}$ into a sample from $\mathbf{u}_{[2k]\times[2k]}$.

By Fact 2, for any product distribution $\mathbf{q}$ over $[2k] \times [2k]$,

$$\begin{aligned}\mathrm{d}_{\mathrm{TV}}(\Phi(\mathbf{p}_z), \mathbf{q}) &\geq \mathrm{d}_{\mathrm{TV}}\big(\Phi(\mathbf{p}_z), \mathbf{u}_{[2k]\times[2k]}\big)/3 \\ &= \mathrm{d}_{\mathrm{TV}}\big(\mathbf{p}_z, \mathbf{u}_{k^2}\big) = \varepsilon,\end{aligned}$$

and the distribution $\Phi(\mathbf{p}_z)$ is at least ε-far from any product distribution. Now, by Item 3, if we obtain n samples from $\Phi(\mathbf{p}_z)$ for a uniformly chosen z, we can convert them to n samples from $\mathbf{p}_z$. Therefore, any algorithm for testing independence can be used to test uniformity for the Paninski class of distributions over $[k^2]$, for which the lower bounds were established in [2]. This proves Corollary 9, assuming the mapping Φ.

We now describe the function Φ satisfying the three conditions. To each $i \in [2\ell^2]$, we associate a collection $C_i = \{a_{i,j}, b_{i,j}\}_{1\leq j\leq 4} \subset [2k] \times [2k]$ of 8 elements and arrange them in a "block" B_i as

$$B_i := \begin{bmatrix} a_{i,1} & b_{i,1} & a_{i,2} & b_{i,2} \\ b_{i,3} & a_{i,3} & b_{i,4} & a_{i,4} \end{bmatrix},$$

The we can see the set of $(2k)^2 = 8 \cdot 2\ell^2$ elements $C := \bigcup_{i=1}^{2\ell^2} C_i$ as a $2k$-by-$2k$ matrix B, comprised of the $2\ell^2$ blocks as follows:

$$B := \begin{bmatrix} B_1 & B_2 & \ldots & B_\ell \\ B_{\ell+1} & B_{\ell+2} & \ldots & B_{2\ell} \\ \vdots & \vdots & \ddots & \vdots \\ B_{(2\ell-1)+1} & B_{(2\ell-1)+2} & \ldots & B_{2\ell^2} \end{bmatrix}.$$

This matrix B enables us to see the target domain $[2k] \times [2k]$ as this 2ℓ-by-ℓ grid of 2-by-4 blocks of elements. Explicitly, this correspondence is given by the indices

$$\begin{aligned} a_{i,1} &= (2r_i + 1, 4c_i + 1),\ b_{i,1} = (2r_i + 1, 4c_i + 2), \\ a_{i,2} &= (2r_i + 1, 4c_i + 3),\ b_{i,2} = (2r_i + 1, 4c_i + 4), \\ a_{i,3} &= (2r_i + 1, 4c_i + 2),\ b_{i,3} = (2r_i + 2, 4c_i + 1), \\ a_{i,4} &= (2r_i + 1, 4c_i + 4),\ b_{i,4} = (2r_i + 1, 4c_i + 3), \end{aligned}$$

where $r_i = \lfloor i/\ell \rfloor$ and $c_i = i \bmod \ell$, for $1 \leq i \leq 2\ell^2$. This enables us to define our mapping $\Phi\colon \Delta_{[k^2]} \to \Delta_{[2k]\times[2k]}$: given a distribution $\mathbf{p}$ over $[k^2] = [4\ell^2]$, let, for all $i \in [2\ell^2]$,

$$\begin{aligned} \Phi(\mathbf{p})\big(a_{i,1}\big) &= \Phi(\mathbf{p})\big(a_{i,2}\big) = \Phi(\mathbf{p})\big(a_{i,3}\big) = \Phi(\mathbf{p})\big(a_{i,4}\big) \\ &= \frac{1}{4}\mathbf{p}(2i-1), \\ \Phi(\mathbf{p})\big(b_{i,1}\big) &= \Phi(\mathbf{p})\big(b_{i,2}\big) = \Phi(\mathbf{p})\big(b_{i,3}\big) = \Phi(\mathbf{p})\big(b_{i,4}\big) \\ &= \frac{1}{4}\mathbf{p}(2i). \end{aligned}$$

Heuristically, for each $1 \le i \le 2\ell^2 = k^2/2$, recalling the layout of the block B_i, the mapping Φ "distributes" the probability masses $\mathbf{p}(2i-1)$ and $\mathbf{p}(2i)$ on 8 elements of C_i as follows:

$$\frac{1}{4}\begin{bmatrix} \mathbf{p}(2i-1) & \mathbf{p}(2i) & \mathbf{p}(2i-1) & \mathbf{p}(2i) \\ \mathbf{p}(2i) & \mathbf{p}(2i-1) & \mathbf{p}(2i) & \mathbf{p}(2i-1) \end{bmatrix}.$$

This implies Item 2, since ℓ_1 distance (and thus total variation) is preserved by this transformation. It also establishes Item 3, as upon seeing a sample x from $\mathbf{p}$, one can generate a sample from $\Phi(\mathbf{p})$ by returning uniformly at random one of the four corresponding elements from the block $B_{\lceil x/2 \rceil}$. Thus, it only remains to show Item 1. This in turn comes from the fact that for every $i \in [2\ell^2]$, by construction, the probabilities under $\Phi(\mathbf{p})$ of each row (resp., column) of block B_i sum to $\frac{1}{2}(\mathbf{p}(2i-1) + \mathbf{p}(2i))$ (resp., $\frac{1}{4}(\mathbf{p}(2i-1) + \mathbf{p}(2i))$), which are $1/k^2$ and $1/2k^2$ respectively for $\mathbf{p}_z$ and $\mathbf{u}_{[k^2]}$, independent of i. ∎

Appendix A
Proof of Lemma 2

In this section, we provide the proof of the variance bound for the RAPPOR-based statistic of Section III-A1.

Lemma 7 (Lemma 2, Restated): For T defined as in (4), we have

$$\mathrm{Var}[T] \le 2kn^2 + 5n^3\alpha^2\|\mathbf{p}-\mathbf{q}\|_2^2 \le 2kn^2 + 4n\mathbb{E}[T].$$

Proof: We let $\lambda_x := \alpha\mathbf{q}(x)+\beta$ and $\mu_x := \frac{1}{n}\mathbb{E}[N_x] = \alpha\mathbf{p}(x)+\beta$ for $x \in [k]$. Dropping the constant terms from T, we define T' such that $\mathrm{Var}[T'] = \mathrm{Var}[T]$ as

$$T' := \sum_{x\in[k]}\Big(N_x^2 - (2(n-1)\lambda_x+1)N_x\Big) = \sum_{x\in[k]} g(N_x,\lambda_x),$$

where $g\colon [0,\infty)\times[0,1] \to \mathbb{R}$ is given by $g(t,\lambda) = t^2 - (2(n-1)\lambda+1)t$. The key difficulty in the analysis arises from the fact that the multiplicities of the N_x terms that arise from RAPPOR are correlated random variables. Because g is not monotone in its first input, the cross covariance terms may be positive even though the N_x terms are negatively associated. As a result, we fully expand out the variance and analyze the terms separately. Recall that

$$\begin{aligned}\mathrm{Var}[T'] &= \sum_{x\in[k]}\mathrm{Var}\big[g(N_x,\lambda_x)\big] \\ &\quad + 2\sum_{x<y}\mathrm{Cov}\big(g(N_x,\lambda_x), g\big(N_y,\lambda_y\big)\big). \end{aligned} \tag{8}$$

We first analyze the sum of variances. A direct computation gives that, for every $x \in [k]$,

$$\begin{aligned}\mathrm{Var}\big[g(N_x,\lambda_x)\big] &= 2n(n-1)\mu_x(1-\mu_x) \\ &\quad \times \Big(\mu_x(1-\mu_x) + 2(n-1)(\lambda_x-\mu_x)^2\Big) \\ &\le \frac{1}{8}n^2 + \alpha^2 n(n-1)^2(\mathbf{p}(x)-\mathbf{q}(x))^2,\end{aligned}$$

where the inequality holds since $\mu_x \in [0,1]$ so $\mu_x(1-\mu_x) \le 1/4$. It follows that

$$\sum_{x\in[k]}\mathrm{Var}\big[g(N_x,\lambda_x)\big] \le \frac{1}{8}n^2k + \alpha^2 n(n-1)^2\|\mathbf{p}-\mathbf{q}\|_2^2. \tag{9}$$

We now turn to the sum of the covariance terms. Fix any $x < y$ in $[k]$. By expanding the corresponding covariance term, we get

$$\begin{aligned}&\mathrm{Cov}\big(g(N_x,\lambda_x), g\big(N_y,\lambda_y\big)\big) \\ &= \mathbb{E}\big[g(N_x,\lambda_x)g\big(N_y,\lambda_y\big)\big] - \mathbb{E}\big[g(N_x,\lambda_x)\big]\mathbb{E}\big[g\big(N_y,\lambda_y\big)\big] \\ &= \mathbb{E}\Big[N_x^2N_y^2\Big] - \big(2(n-1)\lambda_y+1\big)\mathbb{E}\Big[N_x^2N_y\Big] \\ &\quad - (2(n-1)\lambda_x+1)\mathbb{E}\Big[N_xN_y^2\Big] \\ &\quad + (2(n-1)\lambda_x+1)(2(n-1)\lambda_x+1)\mathbb{E}\big[N_xN_y\big] \\ &\quad - n^2(n-1)^2\mu_x\mu_y(\mu_x-2\lambda_x)\big(\mu_y-2\lambda_y\big)\end{aligned} \tag{10}$$

since $\mathbb{E}\big[N_x^2 - (2(n-1)\lambda_x+1)N_x\big] = n(n-1)\mu_x(\mu_x-2\lambda_x)$. We then proceed by evaluating the expressions for $\mathbb{E}\big[N_xN_y\big]$, $\mathbb{E}\big[N_x^2N_y\big]$, $\mathbb{E}\Big[N_xN_y^2\Big]$, and $\mathbb{E}\Big[N_x^2N_y^2\Big]$ separately.

First, by Fact 1, we have that

$$\begin{aligned}\mathbb{E}\big[N_xN_y\big] &= \sum_{1\le i,j\le n}\Pr\big[Y_{ix}=1, Y_{jy}=1\big] \\ &= \sum_{i=1}^{n}\Big(\mu_x\mu_y - \alpha^2\mathbf{p}(x)\mathbf{p}(y)\Big) + 2\sum_{i<j}\mu_x\mu_y \\ &= n^2\mu_x\mu_y - n\alpha^2\mathbf{p}(x)\mathbf{p}(y) \\ &= n^2\mu_x\mu_y - n(\mu_x-\beta)\big(\mu_y-\beta\big).\end{aligned} \tag{11}$$

Second, for $\mathbb{E}\big[N_x^2N_y\big]$, we get

$$\begin{aligned}\mathbb{E}\Big[N_x^2N_y\Big] &= \sum_{1\le i,j,\ell\le n}\Pr\big[Y_{ix}=1, Y_{jx}=1, Y_{\ell y}=1\big] \\ &= n\Pr\big[Y_{ix}=1, Y_{iy}=1\big] + 6\binom{n}{3}\mu_x^2\mu_y \\ &\quad + 2\binom{n}{2}\Big(\mu_x\mu_y + 2\mu_x\Big(\mu_x\mu_y - \alpha^2\mathbf{p}(x)\mathbf{p}(y)\Big)\Big) \\ &= n\mu_x\mu_y - n\alpha^2\mathbf{p}(x)\mathbf{p}(y) + 6\binom{n}{3}\mu_x^2\mu_y \\ &\quad + n(n-1)\mu_x\mu_y + 4\binom{n}{2}\mu_x^2\mu_y \\ &\quad - 4\binom{n}{2}\alpha^2\mu_x\mathbf{p}(x)\mathbf{p}(y),\end{aligned}$$

which, gathering the terms, yields

$$\begin{aligned}\mathbb{E}\Big[N_x^2N_y\Big] &= n^2\mu_x\mu_y - (2(n-1)\mu_x+1)n(\mu_x-\beta)\big(\mu_y-\beta\big) \\ &\quad + n^2(n-1)\mu_x^2\mu_y.\end{aligned} \tag{12}$$

The term $\mathbb{E}\Big[N_xN_y^2\Big]$ term follows similarly.

Finally, for $\mathbb{E}\Big[N_x^2N_y^2\Big]$, note that

$$\begin{aligned}\mathbb{E}\Big[N_x^2N_y^2\Big] &= \sum_{1\le i,j,i',j'\le n}\Pr\big[Y_{ix}=1, Y_{jx}=1, Y_{i'y}=1, Y_{j'y}=1\big] \\ &= n\Big(\mu_x\mu_y - \alpha^2\mathbf{p}(x)\mathbf{p}(y)\Big) + \binom{n}{2}\end{aligned}$$

$$\times \Big(2\mu_x\mu_y + 4\mu_x\Big(\mu_x\mu_y - \alpha^2\mathbf{p}(x)\mathbf{p}(y)\Big) + 4\mu_y\Big(\mu_x\mu_y - \alpha^2\mathbf{p}(x)\mathbf{p}(y)\Big) + 4\Big(\mu_x\mu_y - \alpha^2\mathbf{p}(x)\mathbf{p}(y)\Big)^2\Big) + \binom{n}{3}\Big(6\mu_x^2\mu_y + 6\mu_x\mu_y^2 + 24\mu_x\mu_y \times \Big(\mu_x\mu_y - \alpha^2\mathbf{p}(x)\mathbf{p}(y)\Big)\Big) + 24\binom{n}{4}\mu_x^2\mu_y^2$$

where the second equality follows from counting the different possibilities for the values taken by i, i', j, j'; we divide into cases based on the number of different values taken and apply Fact 1 for each subcase. Note that the total number of terms is $n + 14\binom{n}{2} + 36\binom{n}{3} + 24\binom{n}{4} = n^4$. This can be simplified to

$$\begin{aligned}\mathbb{E}\Big[N_x^2N_y^2\Big] &= n^2(n-1)^2\mu_x^2\mu_y^2 + n^2(n-1)\mu_x\mu_y\big(\mu_x+\mu_y\big) \\ &\quad + n^2\mu_x\mu_y - 4n(n-1)^2(\mu_x-\beta)\big(\mu_y-\beta\big)\mu_x\mu_y \\ &\quad - 2n(n-1)(\mu_x-\beta)\big(\mu_y-\beta\big)\big(\mu_x+\mu_y\big) \\ &\quad + 2n(n-1)(\mu_x-\beta)^2\big(\mu_y-\beta\big)^2 \\ &\quad - n(\mu_x-\beta)\big(\mu_y-\beta\big), \end{aligned} \tag{13}$$

Plugging the bounds from (11) to (13) into (10) and simplifying, we get

$$\begin{aligned}&\mathrm{Cov}\big(g(N_x,\lambda_x), g\big(N_y,\lambda_y\big)\big) \\ &\quad\le 2n(n-1)(\mu_x-\beta)\big(\mu_y-\beta\big) \\ &\qquad\times \big((\mu_x-\beta)\big(\mu_y-\beta\big) - 2(n-1)(\mu_x-\lambda_x)\big(\mu_y-\lambda_y\big)\big) \\ &\quad= 2\alpha^4 n(n-1)\mathbf{p}(x)\mathbf{p}(y) \\ &\qquad\times (\mathbf{p}(x)\mathbf{p}(y) - 2(n-1)(\mathbf{p}(x)-\mathbf{q}(x))(\mathbf{p}(y)-\mathbf{q}(y)))\end{aligned}$$

Summing over all distinct x, y, we have $\sum_{1\le x\ne y\le k}\mathbf{p}(x)^2\mathbf{p}(y)^2 = \|\mathbf{p}\|_2^2 - \|\mathbf{p}\|_4^4 \le \|\mathbf{p}\|_2^2$ and

$$\begin{aligned}&-\sum_{1\le x\ne y\le k}\mathbf{p}(x)\mathbf{p}(y)(\mathbf{p}(x)-\mathbf{q}(x))(\mathbf{p}(y)-\mathbf{q}(y)) \\ &\quad= \sum_{x\in[k]}\mathbf{p}(x)^2(\mathbf{p}(x)-\mathbf{q}(x))^2 - \left(\sum_{x\in[k]}\mathbf{p}(x)(\mathbf{p}(x)-\mathbf{q}(x))\right)^2,\end{aligned}$$

which is at most $\sum_{x\in[k]}\mathbf{p}(x)^2(\mathbf{p}(x)-\mathbf{q}(x))^2 \le \|\mathbf{p}-\mathbf{q}\|_2^2$. Thus,

$$2\sum_{x<y}\mathrm{Cov}\big(g(N_x,\lambda_x), g\big(N_y,\lambda_y\big)\big) \le 2\alpha^4n^2\|\mathbf{p}\|_2^2 + 4\alpha^4n^3\|\mathbf{p}-\mathbf{q}\|_2^2, \tag{14}$$

completing our bound for the cross-variance terms. Combining (9) and (14) into (8) lets us conclude that

$$\begin{aligned}\mathrm{Var}\,[T] &\le n^2\left(\frac{1}{8}k + 2\alpha^4\|\mathbf{p}\|_2^2\right) + \alpha^2n^3\|\mathbf{p}-\mathbf{q}\|_2^2\Big(1+4\alpha^2\Big) \\ &\le 2kn^2 + 5\alpha^2n^3\|\mathbf{p}-\mathbf{q}\|_2^2,\end{aligned}$$

which holds as long as $k \ge 2$, proving the lemma. ■

APPENDIX B
PROOF OF THEOREM 7

Theorem 10 (Joint Probability Perturbation Concentration, Restated): Consider a matrix $\delta \in \mathbb{R}^{k\times k}$ such that, for every $i_0, j_0 \in [k]$, $\sum_{j\in[k]}\delta_{i_0,j} = \sum_{i\in[k]}\delta_{i,j_0} = 0$. Let random variables $X = (X_1,\ldots,X_k)$ and $Y = (Y_1,\ldots,Y_k)$ be independent and uniformly distributed over length-k binary sequences. Define $Z = \sum_{(i,j)\in[k]\times[k]}\delta_{ij}X_iY_j$. Then, for every $\alpha \in (0, 1/16)$, there exists a constant $c_\alpha > 0$ such that

$$\Pr\Big[Z^2 \ge \alpha\|\delta\|_F^2\Big] \ge c_\alpha.$$

Moreover, one can take $c_\alpha = \frac{(1-16\alpha)^2}{1024}$.

Proof: The proof is similar in flavor to that of [3, Th. A.6] (for the case $L = 2$), as we proceed by bounding $\mathbb{E}[Z]$, $\mathbb{E}\big[Z^2\big]$, and $\mathbb{E}\big[Z^4\big]$, before applying the Paley–Zygmund inequality to Z^2. While we could follow the approach of [3, Th. A.6] and handle general 4-symmetric random variables by carefully keeping track of the various quantities in the expansion of $\mathbb{E}\big[Z^2\big]$ and $\mathbb{E}\big[Z^4\big]$, for conciseness we choose here to provide a simpler (albeit less general) proof relying on our specific choice of random variables.

As a first step, let $\theta_i := 2X_i - 1$ and $\theta'_j := 2Y_j - 1$ for $i, j \in [k]$, so that the θ_i and θ'_j are independent Rademacher random variables. Since the sum of entries of δ along any fixed row or column is zero by assumption, we note that

$$Z = \frac{1}{4}\sum_{i,j\in[k]}\delta_{ij}\theta_i\theta'_j. \tag{15}$$

Since θ_i and θ'_j are independent and $\mathbb{E}[\theta_i] = 0$, it follows that $\mathbb{E}[Z] = 0$. For Z^2, we again use independence of θ and θ' to obtain

$$\begin{aligned}\mathbb{E}\Big[Z^2\Big] &= \sum_{(i_1,j_1,i_2,j_2)\in[k]^4}\delta_{i_1j_1}\delta_{i_2j_2}\mathbb{E}\Big[\theta_{i_1}\theta_{i_2}\theta'_{j_1}\theta'_{j_2}\Big] \\ &= \sum_{(i_1,j_1,i_2,j_2)\in[k]^4}\delta_{i_1j_1}\delta_{i_2j_2}\mathbb{E}\big[\theta_{i_1}\theta_{i_2}\big]\mathbb{E}\Big[\theta'_{j_1}\theta'_{j_2}\Big].\end{aligned}$$

Moreover, since the coordinates are independent, we also have $\mathbb{E}\big[\theta_{i_1}\theta_{i_2}\big] = \mathbb{E}\Big[\theta'_{i_1}\theta'_{i_2}\Big] = \mathbb{1}_{\{i_1=i_2\}}$. Therefore,

$$\mathbb{E}\Big[Z^2\Big] = \frac{1}{16}\sum_{(i,j)\in[k]^2}\delta_{ij}^2 = \frac{1}{16}\|\delta\|_F^2.$$

It remains to bound the fourth moment of Z. Using the representation of Z as in (15), we bound the moment-generating function of Z as[6]

$$\log\mathbb{E}_{\theta\theta'}\Big[e^{\lambda Z}\Big] = \log\mathbb{E}_{\theta\theta'}\Big[e^{\frac{\lambda}{4}\theta^T\delta\theta'}\Big] \le \frac{\lambda^2}{32}\cdot\frac{\|\delta\|_F^2}{1-\frac{\lambda^2}{4}\rho\big(\delta^T\delta\big)}, \quad \forall\, 0<\lambda<\frac{2}{\sqrt{\rho\big(\delta^T\delta\big)}},$$

where $\rho(\delta^T\delta)$ is the spectral radius of $\delta^T\delta$. Now, by a standard Markov-based argument, we have that $\mathbb{E}\big[Z^4\big] \le \frac{4!}{\lambda^4}\mathbb{E}_{\theta\theta'}\big[e^{\lambda Z}\big]$

[6]See, e.g., [2, Claim IV.17], and note that the proof goes through even without the positive semi-definiteness assumption.

for all $\lambda > 0$. Therefore, combining the two and using the fact that $\sqrt{\rho(\delta^T\delta)} \le \|\delta\|_F$ we can write

$$\mathbb{E}\Big[Z^4\Big] \le \frac{24}{\lambda^4} e^{\frac{\lambda^2}{32}\cdot\frac{\|\delta\|_F^2}{1-\lambda^2\|\delta\|_F^2/4}}, \qquad \forall\, 0 < \lambda < \frac{2}{\|\delta\|_F}.$$

Setting $\lambda = \frac{1}{C\|\delta\|_F}$ for any constant $C > 0$ yields $\mathbb{E}\big[Z^4\big] \le 24 \cdot C^4 e^{\frac{1}{32(C^2-1/4)}} \|\delta\|_F^4$. Optimizing for $C > 1/2$, we can take $C = \frac{1+\sqrt{65}}{16}$ and get

$$\mathbb{E}\Big[Z^4\Big] \le 4\|\delta\|_F^4. \tag{16}$$

The remainder of the proof follows that of Theorem 3 using the Paley–Zygmund inequality: for every $t \in [0, 1]$

$$\Pr\Big[Z^2 > \frac{t}{16}\|\delta\|_F^2\Big] \ge (1-t)^2 \frac{\mathbb{E}\big[Z^2\big]^2}{\mathbb{E}\big[Z^4\big]} \ge \frac{(1-t)^2}{256 \cdot 4},$$

establishing the theorem (by choosing $t := 16\alpha$ and $c_\alpha := \frac{(1-16\alpha)^2}{1024}$). ∎

Remark 3: Although the proof of Theorem 7 uses full independence of the vectors X and Y (due to the use of the moment-generating function), it is easy to see that the statement still holds when X (resp. Y) is only 4-wise independent. This is because the Paley–Zygmund-based argument only relies on bounds on moments up to order four, and those moments are the same for 4-wise and fully independent vectors.

REFERENCES

[1] J. Acharya, C. Canonne, C. Freitag, and H. Tyagi, "Test without trust: Optimal locally private distribution testing," in *Proc. Mach. Learn. Res.*, Apr. 2019, pp. 2067–2076. [Online]. Available: http://proceedings.mlr.press/v89/acharya19b.html

[2] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints I: Lower bounds from chi-square contraction," 2018. [Online]. Available: arXiv:abs/1812.11476

[3] J. Acharya, C. L. Canonne, and H. Tyagi, "Inference under information constraints II: Communication constraints and shared randomness," 2019. [Online]. Available: arXiv:abs/1804.06952

[4] J. Acharya, C. L. Canonne, Y. Han, Z. Sun, and H. Tyagi, "Domain compression and its application to randomness-optimal distributed goodness-of-fit," in *Proc. 33rd Conf. Learn. Theory*, vol. 125, Jul. 2020, pp. 3–40. [Online]. Available: http://proceedings.mlr.press/v125/acharya20a.html

[5] J. Acharya, C. L. Canonne, Y. Liu, Z. Sun, and H. Tyagi, "Interactive inference under information constraints," 2020. [Online]. Available: arXiv:2007.10976

[6] J. Acharya, C. Daskalakis, and G. C. Kamath, "Optimal testing for properties of distributions," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett, and R. Garnett, Eds., Curran Assoc., Inc., 2015, pp. 3577–3598.

[7] J. Acharya and Z. Sun, "Communication complexity in locally private distribution estimation and heavy hitters," in *Proc. 36th Int. Conf. Mach. Learn.*, vol. 97, Jun. 2019, pp. 51–60. [Online]. Available: http://proceedings.mlr.press/v97/acharya19c.html

[8] J. Acharya, Z. Sun, and H. Zhang, "Differentially private testing of identity and closeness of discrete distributions," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Curran Assoc., Inc., 2018, pp. 6878–6891. [Online]. Available: http://papers.nips.cc/paper/7920-differentially-private-testing-of-identity-and-closeness-of-discrete-distributions.pdf

[9] J. Acharya, Z. Sun, and H. Zhang, "Hadamard response: Estimating distributions privately, efficiently, and with little communication," in *Proc. Mach. Learn. Res.*, Apr. 2019, pp. 1120–1129. [Online]. Available: http://proceedings.mlr.press/v89/acharya19a.html

[10] M. Aliakbarpour, I. Diakonikolas, D. Kane, and R. Rubinfeld, "Private testing of distributions via sample permutations," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Assoc., Inc., 2019, pp. 10878–10889. [Online]. Available: http://papers.nips.cc/paper/9270-private-testing-of-distributions-via-sample-permutations.pdf

[11] M. Aliakbarpour, I. Diakonikolas, and R. Rubinfeld, "Differentially private identity and equivalence testing of discrete distributions," in *Proc. 35th Int. Conf. Mach. Learn.*, Jul. 2018, pp. 169–178. [Online]. Available: http://proceedings.mlr.press/v80/aliakbarpour18a.html

[12] K. Amin, M. Joseph, and J. Mao, "Pan-private uniformity testing," in *Proc. 33rd Conf. Learn. Theory*, vol. 125, Jul. 2020, pp. 183–218. [Online]. Available: http://proceedings.mlr.press/v125/amin20a.html

[13] S. Balakrishnan and L. Wasserman, "Hypothesis testing for high-dimensional multinomials: A selective review," *Ann. Appl. Stat.*, vol. 12, no. 2, pp. 727–749, 2018. [Online]. Available: https://doi.org/10.1214/18-AOAS1155SF

[14] V. Balcer, A. Cheu, M. Joseph, and J. Mao, "Connecting robust shuffle privacy and pan-privacy," 2020. [Online]. Available: https://arxiv.org/abs/2004.09481

[15] R. Bassily, K. Nissim, U. Stemmer, and A. Guha Thakurta, "Practical locally private heavy hitters," in *Advances in Neural Information Processing Systems*, vol. 30, I. Guyon *et al.*, Eds. Red Hook, NY, USA: Curran Assoc., Inc., 2017, pp. 2288–2296. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3d779cae2d46cf6a8a99a35ba4167977-Paper.pdf

[16] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White, "Testing random variables for independence and identity," in *Proc. 42nd Annu. Symp. Found. Comput. Sci. (FOCS)*, Newport Beach, CA, USA, 2001, pp. 442–451.

[17] T. B. Berrett and C. Butucea, "Locally private non-asymptotic testing of discrete distributions is faster using interactive mechanisms," 2020. [Online]. Available: arXiv:2005.12601

[18] E. Blais, C. L. Canonne, and T. Gur, "Distribution testing lower bounds via reductions from communication complexity," in *Proc. 32nd Comput. Complexity Conf.*, vol. 79, 2017, pp. 1–40.

[19] E. Blais, C. L. Canonne, and T. Gur, "Distribution testing lower bounds via reductions from communication complexity," *ACM Trans. Comput. Theory*, vol. 11, no. 2, p. 6, 2019. [Online]. Available: https://doi.org/10.1145/3305270

[20] B. Cai, C. Daskalakis, and G. Kamath, "Priv'it: Private and sample efficient identity testing," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 635–644.

[21] C. L. Canonne, "Big data on the rise?" in *Proc. Int. Colloq. Automata Lang. Program. (ICALP)*, 2015, pp. 294–305. [Online]. Available: http://dx.doi.org/10.1007/978-3-662-47672-7_24

[22] C. L. Canonne, I. Diakonikolas, D. M. Kane, and A. Stewart, "Testing Bayesian networks," in *Proc. Conf. Learn. Theory*, vol. 65, Jul. 2017, pp. 370–448.

[23] C. L. Canonne, G. Kamath, A. McMillan, J. Ullman, and L. Zakynthinou, "Private identity testing for high-dimensional distributions," 2019. [Online]. Available: https://arxiv.org/abs/1905.11947

[24] S.-O. Chan, I. Diakonikolas, G. Valiant, and P. Valiant, "Optimal algorithms for testing closeness of discrete distributions," in *Proc. 25th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA)*, 2014, pp. 1193–1203.

[25] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price, "Sample-optimal identity testing with high probability," in *Proc. 45th Int. Colloq. Automata Lang. Program. (ICALP)*, vol. 107, 2018, pp. 1–14.

[26] I. Diakonikolas and D. M. Kane, "A new approach for testing properties of discrete distributions," in *Proc. 57th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, New Brunswick, NJ, USA, 2016, pp. 685–694.

[27] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proc. 54th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Berkeley, CA, USA, 2013, pp. 429–438.

[28] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography* (Lecture Notes in Computer Science), vol. 3876. Berlin, Germany: Springer, 2006, pp. 265–284.

[29] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized aggregatable privacy-preserving ordinal response," in *Proc. ACM Conf. Comput. Commun. Security*, 2014, pp. 1054–1067.

[30] A. V. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proc. ACM SIGMOD-SIGACT-SIGART Symp. Principles Database Syst. (PODS)*, 2003, pp. 211–222.

[31] M. Gaboardi, H. Lim, R. M. Rogers, and S. P. Vadhan, "Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 2111–2120.
[32] M. Gaboardi and R. Rogers, "Local private hypothesis testing: Chi-square tests," in *Proc. 35th Int. Conf. Mach. Learn.*, Jul. 2018, pp. 1626–1635. [Online]. Available: http://proceedings.mlr.press/v80/gaboardi18a.html
[33] O. Goldreich, "The uniform distribution is complete with respect to testing identity to a fixed distribution," *Electron. Colloq. Comput. Complexity*, vol. 23, p. 15, Mar. 2016. [Online]. Available: http://eccc.hpi-web.de/report/2016/015
[34] O. Goldreich, *Introduction to Property Testing*. Cambridge, U.K.: Cambridge Univ. Press, 2017. [Online]. Available: http://www.wisdom.weizmann.ac.il/∼oded/pt-intro.html
[35] D. Huang and S. Meyn, "Generalized error exponents for small sample universal hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8157–8181, Dec. 2013.
[36] M. Joseph, J. Mao, S. Neel, and A. Roth, "The role of interactivity in local differential privacy," in *Proc. IEEE 60th Annu. Symp. Found. Comput. Sci. (FOCS)*, 2019, pp. 94–105.
[37] P. Kairouz, K. Bonawitz, and D. Ramage, "Discrete distribution estimation under local privacy," in *Proc. 33rd Int. Conf. Mach. Learn. (ICML)*, vol. 48, 2016, pp. 2436–2444.
[38] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" in *Proc. 49th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Philadelphia, PA, USA, Oct. 2008, pp. 531–540.
[39] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith, "What can we learn privately?" *SIAM J. Comput.*, vol. 40, no. 3, pp. 793–826, 2011.
[40] D. Kifer and R. M. Rogers, "A new class of private chi-square hypothesis tests," in *Proc. 20th Int. Conf. Artif. Intell. Stat.*, 2017, pp. 991–1000.
[41] R. Levi, D. Ron, and R. Rubinfeld, "Testing properties of collections of distributions," *Theory Comput.*, vol. 9, pp. 295–347, Mar. 2013. [Online]. Available: http://dx.doi.org/10.4086/toc.2013.v009a008
[42] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4750–4755, Oct. 2008.
[43] R. Rubinfeld, "Taming big probability distributions," *XRDS Crossroads ACM Mag. Students*, vol. 19, no. 1, p. 24, Sep. 2012. [Online]. Available: http://dx.doi.org/10.1145/2331042.2331052
[44] O. Sheffet, "Locally private hypothesis testing," in *Proc. 35th Int. Conf. Mach. Learn.*, Jul. 2018, pp. 4612–4621.
[45] J. N. Tsitsiklis, "Decentralized detection," in *Advances in Statistical Signal Processing*, vol. 2, H. V. Poor and J. B. Thomas, Eds., JAI Press, 1993, pp. 297–344.
[46] G. Valiant and P. Valiant, "An automatic inequality prover and instance optimal identity testing," *SIAM J. Comput.*, vol. 46, no. 1, pp. 429–455, 2017.
[47] Y. Wang, J. Lee, and D. Kifer, "Revisiting differentially private hypothesis tests for categorical data," 2015. [Online]. Available: arXiv:1511.03376
[48] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *J. Amer. Stat. Assoc.*, vol. 60, no. 309, pp. 63–69, 1965.
[49] M. Ye and A. Barg, "Optimal schemes for discrete distribution estimation under locally differential privacy," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5662–5676, Aug. 2018. [Online]. Available: https://doi.org/10.1109/TIT.2018.2809790

Jayadev Acharya (Member, IEEE) received the Bachelor of Technology degree in electronics and electrical communication engineering from the Indian Institute of Technology Kharagpur, Kharagpur, India, in 2007, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California San Diego, San Diego, CA, USA, in 2009 and 2014, respectively. He is an Assistant Professor with the School of Electrical and Computer Engineering, Cornell University. He was a Postdoctoral Associate in electrical engineering and computer science with MIT from 2014 to 2016.

Clément L. Canonne is a Lecturer with the School of Computer Science, University of Sydney, Sydney, NSW, Australia. Prior to this, he was a Motwani Postdoctoral Fellow with Stanford University and a Goldstine Postdoctoral Fellow with IBM Research, after graduating from Columbia University in 2017, where he was advised by R. Servedio. His research focuses on the fields of property testing and sublinear algorithms, and more broadly on computational aspects of learning and statistical inference.

Cody Freitag is currently pursuing the Ph.D. degree in computer science with Cornell Tech, advised by R. Pass. His research interests include theoretical cryptography and privacy and their applications to blockchain technologies and learning theory.

Ziteng Sun received the B.S. degree from Tsinghua University. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Cornell University. His research interest lies in studying the tradeoffs between different resources in modern data science, including samples, privacy, communication, memory, and computation.

Himanshu Tyagi (Senior Member, IEEE) received the B.Tech. degree in electrical engineering and the M.Tech. degree in communication and information technology from the Indian Institute of Technology Delhi, New Delhi, India, in 2007, and the Ph.D. degree from the University of Maryland, College Park, MD, USA, in 2013. From 2013 to 2014, he was a Postdoctoral Researcher with the Information Theory and Applications Center, University of California San Diego, San Diego, CA, USA. Since January 2015, he has been an a Faculty Member with the Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore. His research interests broadly lie in information theory and its application in cryptography, statistics and computer science. Also, he is interested in communication and automation for city-scale systems.