Test without Trust: Optimal Locally Private Distribution Testing

Jayadev Acharya
Cornell University

Clément L. Canonne
Stanford University

Cody Freitag
Cornell University

Himanshu Tyagi
Indian Institute of Science

Abstract

We study the problem of distribution testing when the samples can only be accessed using a locally differentially private mechanism and consider two representative testing questions of identity (goodness-of-fit) and independence testing for discrete distributions. First, we construct tests that use existing, general-purpose locally differentially private mechanisms such as the popular RAPPOR or the recently introduced Hadamard Response for collecting data and propose tests that are sample optimal, when we insist on using these mechanisms. Next, we allow bespoke mechanisms designed specifically for testing and introduce the *Randomized Aggregated Private Testing Optimal Response* (RAPTOR) mechanism which is remarkably simple and requires only one bit of communication per sample. We show that our proposed mechanism yields sample-optimal tests, and in particular, outperforms any test based on RAPPOR or Hadamard Response. A distinguishing feature of our optimal mechanism is that, in contrast to existing mechanisms, it uses public randomness.

1 Introduction

Locally differentially private (LDP) mechanisms have gained prominence as methods of choice for sharing sensitive data with untrusted curators. This strong notion of privacy, introduced in [20, 27, 15] as a "local" variant of differential privacy [18, 17], requires each user to report only a noisy version of its data such that the distribution of the reported data does not change multiplicatively beyond a prespecified factor when the underlying user data changes. With the proliferation of user data accumulated using such locally private mechanisms, there is an increasing demand for designing data analytics toolkits for operating on the collated user data. In this paper, we provide algorithms for enabling such a toolkit comprising statistical tests for the underlying user data distribution.

Specifically, we consider the following setting for hypothesis testing under privacy constraints: Samples $X_1, \ldots, X_n$ generated independently from an unknown distribution p on $[k] = \{1, \ldots, k\}$ are distributed across n users, with user i having access to X_i. Each user describes its sample to a central curator using a (possibly different) mechanism W, namely a channel which upon observing an input $x \in [k]$ sends $z \in \mathcal{Z}$ to the curator with probability $W(z \mid x)$.

However, users must maintain the privacy of their data and are allowed to describe it only using ε-LDP mechanisms W, *i.e.*, a W satisfying (*cf.* [27, 15])

$$\max_{z} \max_{x,x'} \log \frac{W(z \mid x)}{W(z \mid x')} \leq \varepsilon. \tag{1}$$

The parameter $\varepsilon > 0$ indicates the privacy level, with smaller values of ε indicating stronger privacy guarantees. In this work, we focus on the high-privacy regime and assume throughout that $\varepsilon \in (0, 1]$.

The central curator receives the outputs $Z_1, \ldots, Z_n$ of ε-LDP mechanisms $W^n = (W_1, \ldots, W_n)$, where W_i is the mechanism used by user i. At a high-level, we seek to address the following question: *How can the curator conduct statistical testing for p using observations $Z^n = (Z_1, \ldots, Z_n)$?*

In particular, we consider *uniformity testing*, which is the prototypical *identity testing* (goodness-of-fit) problem where the curator seeks to determine if $p = u$, the uniform distribution on $[k]$, or if $\mathrm{d}_{\mathrm{TV}}(p, u) \geq \gamma$ (where d_{TV} denotes the total variation distance). We seek algorithms that are efficient in the number of LDP user data samples required and can also be implemented practically. Our main focus is the uniformity testing problem, but we obtain results for *independence testing* as well using similar techniques.

Our results are organized into two categories based on

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

the choices of W^n: In the first category, each user applies the same mechanism W, *i.e.*, $W_i = W$ for $1 \le i \le n$, which is set to an existing LDP data release mechanisms. In particular, we set W to the popular RAPPOR mechanism of [19] and a recently introduced mechanism called Hadamard Response (HR) [5]. Because these mechanisms have utility beyond our specific use-case of distribution testing – RAPPOR, for instance, is already deployed in many applications – it is natural to build a more comprehensive data analytics toolkit using the data accumulated by these mechanisms. To this end, we provide uniformity testing algorithms with optimal sample complexity for both of these mechanisms. Further, we provide an algorithm for independence testing using HR and analyze its performance.

In the next category, we allow the more general class of *public-coin mechanisms* where the users can choose mechanisms $W^n = W_U^n$ as a function of public randomness U which is available to each of them and also to the curator. Note that since U is available to the curator, for each fixed realization $U = u$, to maintain privacy the mechanism $W_{i,u}$ applied to X_i must satisfy (1).

We note that ε-LDP mechanisms with constant (*i.e.,* degenerate) public randomness U can be viewed as *private-coin mechanisms* since private randomness is required to implement any ε-LDP W. RAPPOR and HR, too, are private-coin mechanisms, with the additional restriction that each user applies the same mechanism.

We present a new public-coin mechanism, *Randomized Aggregated Private Testing Optimal Response* (RAPTOR), that only requires users to send a single privatized bit indicating whether their data point is in a (publicly known) random subset of the domain. Using RAPTOR, we obtain simple algorithms for uniformity and independence testing that are sample-optimal even among tests based on public-coin mechanisms.

We provide below a detailed description of our results, followed by a brief discussion of the relevant literature to put them in perspective. We remark at the outset that the problems studied here have been considered earlier in [31, 22]. However, [22] did not consider finite sample performance analysis and the sample complexity upper bounds shown in [31] are far from optimal. Further, none of the prior works considered the role of public randomness – an often available resource used critically in our optimal test. Moreover, lower bounds for sample complexity were not available prior to our work. We fill this gap and establish tight lower bounds for both our settings: when we restrict to RAPPOR or HR for collecting data, and when no such restriction is imposed and even public-coin mechanisms are allowed.

1.1 Algorithms and results

Consider uniformity testing using a locally private mechanism. Given a public-coin ε-LDP mechanism W^n with public randomness $U \in \mathcal{U}$, a mapping $\tau\colon \mathcal{Z}^n \times \mathcal{U} \to \{0,1\}$ constitutes a *uniformity test using* W^n if the output Z^n of W^n satisfies $\Pr_u[\,\tau(Z^n,U)=1\,] \ge 2/3$ and $\Pr_p[\,\tau(Z^n,U)=0\,] \ge 2/3$ for every p such that $\mathrm{d_{TV}}(p,u) \ge \gamma$.

We begin with the results for uniformity testing using RAPPOR and HR. We briefly describe these mechanisms here. The mechanism RAPPOR is given by the channel W_R with output alphabet $\mathcal{Z} = \{0,1\}^k$ and such that for a given input x the output bits $Z^{(1)},\dots,Z^{(k)}$ are independent with $Z^{(x)}$ distributed as $\mathrm{Bern}(\alpha_R+\beta_R)$ and $Z^{(x')}$, $x' \ne x$, distributed as $\mathrm{Bern}(\beta_R)$, where

$$\alpha_R := \frac{e^{\varepsilon/2}-1}{e^{\varepsilon/2}+1} = \Theta(\varepsilon), \quad \beta_R := \frac{1}{e^{\varepsilon/2}+1}\,. \tag{2}$$

It can be verified (*cf.* [19]) that W_R is ε-LDP, and that $\mathbb{E}\big[Z^{(x)}\big] = \alpha_R p(x) + \beta_R$ for every $x \in [k]$.

HR, on the other hand, is a generalization of the classic Randomized Response (RR) [34] and can be described as follows. Let $(C_x)_{x\in[k]}$ be a collection of sets each of size $k/2$ such that every pair of sets overlap in exactly $k/4$ elements. Note that such a set system can be defined using the Hadamard code, and can be implemented efficiently using a Hadamard matrix.[1] HR is given by a channel W_H with output alphabet $\mathcal{Z} = [k]$ and such that

$$W_H(z \mid x) = \begin{cases} \frac{2}{k}\cdot\frac{e^\varepsilon}{e^\varepsilon+1} & \text{if } z \in C_x, \\ \frac{2}{k}\cdot\frac{1}{e^\varepsilon+1} & \text{if } z \in [k]\setminus C_x. \end{cases} \tag{3}$$

It was shown in [5] that W_H is an ε-LDP mechanism.

We say that a test τ is a *uniformity test using* RAPPOR (resp. HR) if the public randomness U is constant and τ constitutes a uniformity test using W^n with each $W_i = W_R$ (resp. W_H).

We first propose a uniformity test using RAPPOR, described in Algorithm 1 (a formal description is provided in Section 2.1). Moving now to uniformity test using HR, denote by q^* the output distribution of HR when the underlying samples are generated from the uniform distribution. Note that q^* can be computed explicitly. Invoking Parseval's theorem, we show that the ℓ_2 distance between the q^* and the output distribution of HR is roughly $\varepsilon/\sqrt{k}$ times the ℓ_2 distance

[1] We assume here for simplicity of exposition that k is a power of 2, and omit some technical details.

Algorithm 1 Uniformity testing using RAPPOR

1: Obtain $Z_1, \ldots, Z_n$ using RAPPOR.
2: For each x in $[k]$, compute the number N_x of k-bit vectors Z_i for which the x-th entry is 1.
3: Compute the test statistic T described in (5) which is, in essence, a bias-corrected version of the collision statistic $\sum_x (N_x^2 - N_x)$.
4: If T is less than roughly $n^2\gamma^2\varepsilon^2/k$, declare uniform; else declare not uniform.

between the uniform and the user data distributions. This motivates our second uniformity test, described in Algorithm 2. We analyze the sample complexity of

Algorithm 2 Uniformity testing using HR

1: Obtain $Z_1, \ldots, Z_n$ using HR.
2: Using an appropriate ℓ_2-test, test if the ℓ_2 distance between the distribution of Z_i's and q^* is less than roughly $\gamma\varepsilon/k$; in this case declare uniform. Else declare not uniform.

the tests above and show that it is order-wise optimal among all uniformity tests using RAPPOR or HR.

Result 1 (Sample complexity of uniformity testing using RAPPOR)**.** *The tests described in Algorithm 1 and Algorithm 2, respectively, constitute uniformity tests using* RAPPOR *and* HR *for $n = O(k^{3/2}/(\gamma^2\varepsilon^2))$ samples. Furthermore, any uniformity test using* RAPPOR *or* HR *must use $\Omega(k^{3/2}/(\gamma^2\varepsilon^2))$ samples.*

Thus, both tests proposed above provably cannot be improved beyond this barrier of $\Omega(k^{3/2}/(\gamma^2\varepsilon^2))$ samples, as long as the mechanisms are restricted to RAPPOR and HR. Interestingly, this was conjectured by Sheffet to be the optimal sample complexity of locally private uniformity testing [31], although no algorithm achieving this sample complexity was provided. Yet, our next result shows that we can make do with much fewer samples when public randomness is allowed.

We propose a new public-coin mechanism RAPTOR, described in Algorithm 3. Note that RAPTOR can be

Algorithm 3 The RAPTOR mechanism

1: The curator and the users sample a uniformly random subset S of $[k]$ of cardinality $k/2$.
2: Each user computes the bit indicator $B_i = \mathbb{1}_{\{X_i \in S\}}$ and sends it using RR, *i.e.*, flips it with probability $1/(1+e^{\varepsilon})$ and sends the outcome to the curator.

cast in our notation for public-coin mechanisms by setting $U = S$ and channels W_i, $1 \le i \le n$, such that on input x_i the output is $Z_i = \mathbb{1}_{\{x_i \in S\}}$. We call a uniformity test using W^n a *uniformity test using* RAPTOR *with n samples.*

To build a uniformity test using RAPTOR, we observe that the bits B_i preserve the statistical distance between the two hypothesis classes, up to a shrinkage factor of $\Omega(1/\sqrt{k})$. Specifically, when the underlying distribution is γ-far from uniform, the bias of B_i is $1/2 + \Omega(\gamma/\sqrt{k})$ with constant probability (over the choice of S). Clearly, uniform distribution the bits B_i are unbiased. Thus, we can simply test for uniformity by learning the bias of the bits up to an accuracy of $\gamma/\sqrt{k}$, which can be done using $n = O(k/(\gamma^2\varepsilon^2))$ samples from RAPTOR. In fact, we further show that (up to constant factors) this number of samples cannot be improved upon.

Result 2 (Sample complexity of locally private uniformity testing)**.** *There exists a uniformity test using* RAPTOR *with $O(k/(\gamma^2\varepsilon^2))$ samples. Furthermore, any uniformity test using a public-coin mechanism requires $\Omega(k/(\gamma^2\varepsilon^2))$ samples.*

Although we have stated the previous three results for uniformity testing, our proofs extend easily to identity testing, *i.e.*, the problem of testing equality of the underlying distribution to a fixed known distribution q which is not necessarily uniform. In fact, if we allow simple preprocessing of user observations before applying locally private mechanisms, a reduction argument due to Goldreich [23] can be used to directly convert identity testing to uniformity testing. We defer the details to the extended version of the paper [1].

Our final set of results are for independence testing, where user data consists of two-dimensional vectors (X_i, Y_i) from $[k] \times [k]$. We only state this problem and the results informally here and leave the details to the full version. We seek to ascertain if these vectors were generated from a product distribution $p_1 \otimes p_2$ or a distribution that is γ-far in total variation distance from every independent distribution. For this problem, a natural counterpart of RAPTOR which simply applies RAPTOR to each of the two coordinate using independently generated sets yields a sample-optimal test – indeed, we simply need to test whether the pair of indicator bits at the output of this mechanism are independent. This can be done using $O(k^2/(\gamma^2\varepsilon^2))$, leading to the next result.

Result 3 (Sample complexity of locally private independence testing)**.** *There exists an independence test using* RAPTOR *with $O(k^2/(\gamma^2\varepsilon^2))$ samples. Furthermore, any independence test using a public-coin mechanism requires $\Omega(k^2/(\gamma^2\varepsilon^2))$ samples.*

For completeness, we also consider independence testing using existing mechanisms and provide an independence test using HR which requires $O(k^3/(\gamma^2\varepsilon^4))$ samples. The proposed test builds on a technique in-

troduced in [4] and relies on learning in χ^2 divergence. Although this result maybe suboptimal in the dependence on the privacy parameter ε, it improves on sample complexity of both [31] and the testing-by-learning baseline approach by a factor of roughly k. We summarize all our results in Table 1 and compare them with the best known prior bounds from [31].

Testing	This work		Previous [31]
	Private-Coin	Public-Coin	Private-Coin
Uniformity	$O\left(\frac{k^{3/2}}{\gamma^2\varepsilon^2}\right)$	$\Theta\left(\frac{k}{\gamma^2\varepsilon^2}\right)$	$O\left(\frac{k^2}{\gamma^2\varepsilon^2}\right)$
Independence	$O\left(\frac{k^3}{\gamma^2\varepsilon^4}\right)$	$\Theta\left(\frac{k^2}{\gamma^2\varepsilon^2}\right)$	$O\left(\frac{k^4}{\gamma^2\varepsilon^2}\right)$

Table 1: Summary of our results and comparison with prior work. Private-coin result for uniformity testing is achieved for both RAPPOR and HR and is optimal for any test based on these mechanisms. The private-coin result for independence testing uses HR.

Conceptually, our main contribution is a quantitative characterization of how information constraints imposed by local privacy requirements affect the difficulty of a testing problem. Our lower bounds rely on the approach proposed in [2] (see [3] for more general results) to analyze the contractions in chi-square distance due to such constraints. On the other hand, our algorithms yield quantitatively optimal mitigation for the information lost due to these contractions. For tests based on RAPPOR and HR, this not only requires a careful construction of test statistic based on sanitized samples, but also a non-trivial analysis that sheds light on how these mechanisms modify data statistics. The key message of our work is that public randomness can be used gainfully to optimally alleviate the information loss by privacy constraints.

1.2 Proof techniques

We start by describing the analysis of our tests based on existing ε-LDP mechanisms. Recall that a standard (non-private) uniformity test entails estimating the ℓ_2 norm of the underlying distribution by counting the number of collisions in the observed samples. When applying the same idea on the data collected via RAPPOR, we can naively try to estimate the number of collisions by adding the number of pairs of output vectors with 1s in the x-th coordinate, for each x. However, the resulting statistic has a prohibitively high variance stemming from the noise added by RAPPOR. We fix this shortcoming by considering a bias-corrected version of this statistic that closely resembles the classic χ^2 statistic. However, analyzing the variance of this new statistic turns out to be rather technical and involves handling the covariance of quadratic functions of correlated binomial random variables. Our main technical effort in this part goes into analyzing this covariance, which may find further applications.

For testing uniformity using HR, we follow a different approach. In this case, we exploit the structure of Hadamard transform and take recourse to Parseval's theorem to show that the ℓ_2 distance to uniformity of the original distribution p is equal, up to an $\varepsilon/\sqrt{k}$ factor, to the ℓ_2 distance of the Fourier transform $H(p)$ to some (explicit) fixed distribution q. Further, it can be shown that $\|q\|_2 = O(1/\sqrt{k})$. With this structural result in hand, we can test identity of $H(p)$ to q in the Fourier domain, by invoking the (non-private) ℓ_2 tester of Chan et al. [12] with the corresponding distance parameter $\gamma\varepsilon/\sqrt{k}$. Exploiting the fact that q has a small ℓ_2 norm leads to the stated sample complexity.

As mentioned above, our main results – the optimal public-coin mechanisms for identity and independence testing – are remarkably simple. The key heuristic underlying both stems from the following claim: *If p is γ-far from uniform, then with constant probability a uniformly random subset $S \subseteq [k]$ of size $k/2$ will satisfy $p(S) = 1/2 \pm \Omega(\gamma/\sqrt{k})$.* On the other hand, if p is uniform then $p(S) = 1/2$ always holds. Thus, one can reduce the original testing problem (over alphabet size k) to the much simpler question of estimating the bias of a coin. This latter task is easy to perform optimally in a locally private manner – for instance it can be completed only using the classic randomized response – and requires each player to send only *one* bit to the server. Hence, the main technical difficulty is to prove this quite intuitive claim. We do this by showing anticoncentration bounds for a suitable random variable by bounding its fourth moment and invoking the Paley–Zygmund inequality. As a byproduct, we obtain a more general version, Theorem 7, which we believe to be of independent interest.

Our information-theoretic lower bounds are all based on a general approach introduced recently by Acharya, Canonne, and Tyagi [2] (see [3] for more general lower bounds) that allows us to handle the change in distances between distributions when information constraints are imposed on samples. We utilize the by-now-standard "Paninski construction" [29], a collection $\mathcal{C}$ of $2^{k/2}$ distributions obtained by adding a small pointwise perturbation to the k-ary uniform distribution. In order to obtain a lower bound for the sample complexity of locally private uniformity testing, following [2], we consider n noisy channels $(W_j \colon [k] \to \{0,1\}^*)_{j\in[n]}$ and the distribution $\mathcal{W}(p)$ of the tuple of n messages when the underlying distribution of the samples is p. The key step then is to bound the χ^2 divergence between (i) $\mathcal{W}(u)$, the distribution of the messages under the uniform distribution; and

(ii) $\mathbb{E}_{p\in\mathcal{C}}[\mathcal{W}(p)]$, the *average* distribution of the messages when p is chosen uniformly at random among the "perturbed distributions."

Using the results of [2] (*cf.* [3]), this in turn is tantamount to obtaining an upper bound for the Frobenius norm of specific $[k/2]\times[k/2]$ matrices $\mathbf{H}_1,\ldots,\mathbf{H}_n$ that capture the information constraints imposed by W_j's. Deriving these bounds for Frobenius norms constitutes the main technical part of the lower bounds and relies on a careful analysis of the underlying mechanism and of the LDP constraints it must satisfy. Due to lack of space, we omit a more detailed discussion of lower bound proofs.

1.3 Related prior work

Testing properties of a distribution by observing samples from it is a central problem in statistics and has been studied for over a century. It has seen renewed interest in the computer science community under the broad title of *distribution testing*, with a particular focus on sample-optimal algorithms for discrete distributions. This literature itself spans the last two decades; we refer an interested reader to surveys and books [30, 11, 24, 8] for a comprehensive review. Due to lack of space, we only touch upon few results in this area that are related directly to our paper.

The sample complexity for uniformity testing was shown to be $\Theta(k^{1/2}/\gamma^2)$ in [29], following a long line of work. Several tests achieving this optimal sample complexity are now available and even the optimal dependence on error probability is known (*cf.* [25, 13]). We remark that it is easy to extend our uniformity testing results to the more general problem of identity testing using a reduction argument from Goldreich [23]. In fact, in a manner similar to [2], this reduction can be used in conjunction with results from [9] to extend our results to the instance-optimal setting of [32]. The optimal sample complexity for the independence testing problem where both observations are from the same set $[k]$ was shown to be $\Theta(k/\gamma^2)$ in [4, 14].

Moving now to distribution testing settings with privacy constraints, the setting of *central differentially private* (DP) testing has been extensively studied. Here the algorithm itself is run by a trusted curator who has access to all the user data, but needs to ensure that the output of the test maintains differential privacy; see [21, 28, 33] for a sampling of results on identity and independence testing. Identity testing in the finite sample setting has been considered in [10, 7], with a complete characterization of sample complexity derived in [6]. Interestingly, in several parameter ranges of interest the sample complexity here matches the sample complexity for the non-private case discussed earlier, showing that "privacy often comes at no additional cost" in this setting. As we show in this work, this is in stark contrast to what can be achieved in the more stringent locally private setting.

Coming to the literature most closely related to our work, locally private hypothesis testing was considered by Sheffet in [31] where, too, both identity and independence testing were considered. However, as remarked earlier, this work did not consider the role of public randomness, and even among private-coin mechanisms the algorithms proposed in [31] have suboptimal sample complexity. Indeed, note that the problem of learning the unknown k-ary distribution up to an accuracy of γ in total variation distance in the locally private setting has received a lot of attention, and its optimal sample complexity is known to be $\Theta(k^2/(\gamma^2\varepsilon^2))$; see [16, 19, 35, 26, 5]. Clearly, the testing problems we consider can be solved by privately learning the distributions (to accuracy γ). This readily implies a sample complexity upper bound of $O(k^2/(\gamma^2\varepsilon^2))$ for locally private identity testing, and of $O(k^4/(\gamma^2\varepsilon^2))$ for independence testing. In this respect the performance guarantees obtained in [31] are not entirely satisfactory, since the same (and in some cases even better) performance can be achieved by this "testing-by-learning" approach.

Subsequent work. Recent results by a subset of the authors [3], studying inference under general local information constraints (of which local differential privacy is an example), supersede the private-coin lower bounds obtained in the current paper. Specifically, [3] establishes an $\Omega(k^{3/2}/(\gamma^2\varepsilon^2))$ sample complexity lower bound for all uniformity tests using private-coin LDP mechanisms; thus showing that the sample complexity of both Algorithms 4 and 5 is indeed order-wise optimal among all such tests. Hence, all locally private uniformity tests provided in the current paper (both using public- and private-coin mechanisms) achieve optimal sample complexity for the respective class of mechanisms.

Organization. In the interest of space, in this extended abstract we only provide the statements and main lemmata of our results. Omitted details, as well as the sections on lower bounds and independence testing, are deferred to the full version [1].

2 Locally Private Uniformity Testing using Existing Mechanisms

In this section, we provide two locally private mechanisms for uniformity testing. As discussed earlier, this in turn provides similar mechanisms for identity testing as well. These two tests, based respectively on

the private-coin mechanisms RAPPOR and HR, will be seen to both have sample complexity $O(k^{3/2}/\gamma^2\varepsilon^2)$. However, the first has the advantage of being based on a widespread mechanism, while the second is more efficient in terms of both time and communication.

2.1 A uniformity test using Rappor

Given n independent samples from p, let the output of RAPPOR applied to these samples be denoted by $\mathbf{b}_1, \dots, \mathbf{b}_n \in \{0,1\}^k$, where $\mathbf{b}_i = (\mathbf{b}_{i1}, \dots, \mathbf{b}_{ik})$ for $i \in [n]$. The following fact is a simple consequence of the definition of RAPPOR.

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

$$\Pr[\,\mathbf{b}_{ix} = 1, \mathbf{b}_{jy} = 1\,] = \begin{cases} (\alpha_R p(x) + \beta_R)(\alpha_R p(y) + \beta_R), & \text{if } i \neq j, \\ \alpha_R p(x) + \beta_R, & \text{if } i = j,\ x = y, \\ (\alpha_R p(x) + \beta_R)(\alpha_R p(y) + \beta_R) - \alpha_R^2 p(x)p(y), & \text{o.w.}, \end{cases}$$

where α_R, β_R are defined as in (2).

First idea: Counting Collisions. A natural idea would be to try and estimate $\|p\|_2^2$ by counting the collisions from the output of RAPPOR. Since this only adds post-processing to RAPPOR, which is LDP, the overall procedure does not violate the ε-LDP constraint. For $\sigma^x_{i,j}$ defined as $\mathbb{1}_{\{\mathbf{b}_{ix}=1, \mathbf{b}_{jx}=1\}}$, $x \in [k]$, $i \neq j$, the statistic $S := \sum_{1 \le i < j \le n} \sum_{x \in [k]} \sigma^x_{i,j}$ counting collisions over all samples and differentially private symbols can be seen to have expectation

$$\mathbb{E}[S] = \binom{n}{2}\left(\alpha_R^2 \|p\|_2^2 + 2\alpha_R\beta_R + k\beta_R^2\right) \asymp \varepsilon^2 n^2 \|p\|_2^2 + k.$$

Up to the constant normalizing factor, this suggests an unbiased estimator for $\|p\|_2^2$, and thereby also for $\|p-u\|_2^2 = \|p\|_2^2 - 1/k$. However, the issue lies with the variance of this estimator. Indeed, it can be shown that $\mathrm{Var}(S) \approx n^3 k$ (for constant ε). Thus, if we use this statistic to distinguish between $\|p\|_2^2 = 1/k$ and $\|p\|_2^2 > (1 + \Omega(\gamma^2))/k$ for uniformity testing, we need $\sqrt{n^3 k} \ll n^2\varepsilon^2 \cdot (\gamma^2/k)$ *i.e.*, $n \gg k^3/(\gamma^4\varepsilon^4)$. This sample requirement turns out to be off by a quadratic order, and even worse than the trivial upper bound obtained by learning p.

An Optimal Mechanism. We now propose our testing mechanism based on RAPPOR, which, in essence, uses a privatized version of a χ^2-type statistic of [12, 4, 32]. For $x \in [k]$, let the number of occurrences of x among the n (privatized) outputs of RAPPOR be

$$N_x := \sum_{j=1}^{n} \mathbb{1}_{\{\mathbf{b}_{jx}=1\}} \tag{4}$$

which by the definition of RAPPOR follows a $\mathrm{Bin}(n, \alpha_R p(x) + \beta_R)$ distribution. Set

$$T := \sum_{x \in [k]} \left(\left(N_x - (n-1)\left(\frac{\alpha_R}{k} + \beta_R\right)\right)^2 - N_x \right) + k(n-1)\left(\frac{\alpha_R}{k} + \beta_R\right)^2. \tag{5}$$

This T is a statistic, applied to the output of RAPPOR, which (as we shall see) is up to normalization an unbiased estimator for the squared ℓ_2 distance of p to uniform. The main difference with the naive approach we discussed previously lies in the extra linear term. Indeed, the collision-based statistic of the previous section has the form $S \propto \sum_{x \in [k]} \left(N_x^2 - N_x\right)$, and in comparison, 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

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

since $\beta_R \approx 1/2$. The fluctuations of the quadratic term are reduced significantly by the subtracted linear term, bringing down the variance of the statistic. This leads to our algorithm based on RAPPOR, Algorithm 4, and yields the main result of this section:

Theorem 2. *The test described in Algorithm 4 constitutes a uniformity test using* RAPPOR *for $n = O\left(k^{3/2}/(\gamma^2\varepsilon^2)\right)$ samples.*

Algorithm 4 LDP Uniformity Testing using RAPPOR

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

1: Set $\alpha_R \leftarrow \frac{e^{\varepsilon/2}-1}{e^{\varepsilon/2}+1}$, $\beta_R \leftarrow \frac{1}{e^{\varepsilon/2}+1}$ as in (2).
2: Apply (ε-LDP) RAPPOR to the samples to obtain $(\mathbf{b}_i)_{1 \le i \le n}$ ▷ Time $O(k)$ per user
3: Compute $(N_x)_{x \in [k]}$, as per (4) ▷ Time $O(kn)$
4: Compute T, as defined in (5) ▷ Time $O(k)$
5: **if** $T < n(n-1)\alpha_R^2\gamma^2/k$ **return** uniform
6: **else return** not uniform

2.2 A uniformity test using Hadamard Response

Although the RAPPOR-based mechanism of Section 2.1 achieves a significantly improved sample complexity over the naive learning-and-testing approach, it suffers several shortcomings. The most apparent is its time complexity: inherently, the one-hot encoding procedure used in RAPPOR leads to a time complexity of $\Theta(kn)$, with an extra linear dependence on the alphabet size k, which is far from the "gold standard" of $O(n)$ time complexity. A more time-efficient procedure can be obtained using HR. In fact, we describe

a uniformity test using HR that has the same sample complexity as the one using RAPPOR described above, but is much more time-efficient.

Theorem 3. *The test described in Algorithm 5 constitutes a uniformity test using* HR *for* $n = O\big(k^{3/2}/(\gamma^2\varepsilon^2)\big)$ *samples. Moreover, the test runs in time near-linear in the number of samples.*[2]

Algorithm 5 LDP Uniformity Testing using HR

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

1: Set $\alpha_H \leftarrow \frac{e^\varepsilon - 1}{e^\varepsilon + 1}$, $K \leftarrow 2^{\lceil \log(k+1) \rceil}$
2: Apply HR (with parameters ε, K) to the samples to get n samples in $[K]$ ▷ Time $O(\log k)$ per user
3: Invoke the testing algorithm TEST-ℓ_2 of Theorem 5 on these n samples, with parameters $b \leftarrow \frac{1+\alpha_H}{\sqrt{K}}$, $\gamma' \leftarrow \frac{2\alpha_H\gamma}{kK}$ and q^* being the explicit distribution from Theorem 4 ▷ Time $O(n \log k + n \log n)$
4: **if** TEST-ℓ_2 accepts **return** uniform
5: **else return** not uniform

To describe the intuition behind this algorithm, suppose we feed inputs from an input distribution $p \in \Delta([k])$ to the HR mechanism, whose output then follows some induced distribution $q \in \Delta([K])$, where $\Delta(\mathcal{X})$ denotes the set of distributions on the set $\mathcal{X}$. It is natural to expect that whenever p is uniform (over $[k]$), then q is uniform (over $[K]$), too; and that conversely if p is not uniform, then q is neither, and that the distance to uniformity is preserved. This is not exactly what we will obtain. However, we can get something close to it in the next result, which suffices for our purpose.[3]

Theorem 4. *Let $\varepsilon \in (0,1]$, $K = O(k)$ be a power of 2, and denote by q the output distribution over $[K]$. Then, we have*

$$\|q - q^*\|_2^2 = \frac{\alpha_H^2}{K} \cdot \|p - u\|_2^2 \asymp \frac{\varepsilon^2}{k}\|p - u\|_2^2, \qquad (6)$$

where $\alpha_H := \frac{e^\varepsilon - 1}{e^\varepsilon + 1}$, and $q^* \in \Delta([K])$ is an explicit distribution, efficiently computable and independent of p, with $\|q^*\|_2 \le (1 + \alpha_H)/\sqrt{K}$. Moreover, q^* can be sampled in time $O(\log K)$.

Thus, when $p = u$, we get $q = q^*$. Otherwise when $\mathrm{d}_{\mathrm{TV}}(p,u) > \gamma$, then

$$\|q - q^*\|_2^2 > \frac{4\alpha_H^2\gamma^2}{kK} = \Theta\left(\frac{\varepsilon^2}{k^2}\gamma^2\right). \qquad (7)$$

[2] We say that a complexity is *near-linear* in a parameter t if it is of the form $O(t \operatorname{poly}(\log t))$.

[3] To see that our desired statement cannot hold as stated above, note that for $p = u$, the definition of HR (*cf.* (3)) implies $q(z_1) = \frac{1+\alpha_H}{K}$, since $|D_{z_1}| = k$ as the first column of H_K is the all-one vector. Thus the squared ℓ_2 distance of q to uniform is at least α_H^2/K.

The observation above suggests that if we can estimate the ℓ_2 distance between q and q^*, we can get our desired uniformity test. We facilitate this by invoking the result below, which follows from the ℓ_2-distance estimation algorithm of [12, Proposition 3.1], combined with an observation from [14, Lemma 2.3]:

Theorem 5 (Adapted from [12, Proposition 3.1])**.** *For two unknown distributions $p, q \in \Delta([k])$, there exists an algorithm* TEST-ℓ_2 *that distinguishes with probability at least 2/3 between the cases $\|p - q\|_2 \le \gamma/2$ and $\|p - q\|_2 > \gamma$ by observing $O(\min(\|p\|_2, \|q\|_2)/\gamma^2)$ samples from each. Moreover, this algorithm runs in time near-linear in the number of samples.*

We apply the algorithm of Theorem 5 to our case by generating desired number of samples from q^*, which can simply be obtained by passing samples from the uniform distribution via HR, and using them along with the samples observed from q at the output of HR. We need to distinguish between the cases $q = q^*$ and $\|q - q^*\|_2 > \gamma'/\sqrt{K}$, which by the previous result can be done using $O(\|q^*\|_2 K/\gamma'^2)$ samples where $\gamma' := 2\alpha_H\gamma/\sqrt{k}$. Substituting $K = O(k)$ and $\|q^*\|_2 = O(1/\sqrt{K})$, the number of samples we need is

$$O\left(\frac{1}{\sqrt{K}} \cdot K \cdot \left(\frac{\sqrt{k}}{\gamma\varepsilon}\right)^2\right) = O\left(\frac{k^{3/2}}{\gamma^2\varepsilon^2}\right),$$

which is our claimed sample complexity.

The time complexity follows from the efficiency of Hadamard encoding (see [5, Section 4.1]), which allows each player to generate their private sample in time $O(\log K) = O(\log k)$, and to send only $O(\log k)$ bits.[4] After this, running the TEST-ℓ_2 algorithm takes time $O(n \log K + n \log n)$, the first term being the time required to generate n samples from q^*. Thus, to conclude the proof of Theorem 3, it only remains to establish Theorem 4, which we do in the full version.

3 Optimal Locally Private Uniformity Testing

In the foregoing treatment, we saw that existing (private-coin) mechanisms such as RAPPOR and HR can perform uniformity testing using $O(k^{3/2}/(\gamma^2\varepsilon^2))$ samples at best. In this section, we describe our public-coin mechanism, RAPTOR,[5] and use it to design an algorithm for testing uniformity that requires only $O(k/(\gamma^2\varepsilon^2))$ samples and constant communication per sample. Our algorithm builds upon the following folklore fact:

[4] This is significantly better than the $O(k)$ time and communication per player of Algorithm 4.

[5] Which stands for *Randomized Aggregated Private Testing Optimal Response*.

Fact 6. *For $\varepsilon \in (0,1]$, an estimate of the bias of a coin with an additive accuracy of γ can be obtained using $O\big(1/(\gamma^2\varepsilon^2)\big)$ samples via ε-LDP* RR.

Specifically, we use public randomness to reduce the uniformity testing problem for an arbitrary k to that for $k = 2$, albeit with γ replaced with $\gamma/\sqrt{k}$; and then apply the algorithm above.

To enable the aforementioned reduction, we need to show that the probabilities of a randomly generated set differ appropriately under the uniform distribution and a distribution that is γ far from uniform in total variation distance. To accomplish this, we prove a more general result which might be of independent interest. We say that random variables $X_1, X_2, \ldots, X_k$ are 4-*symmetric* if $\mathbb{E}[X_{i_1}X_{i_2}X_{i_3}X_{i_4}]$ depends only on the number of times each element appears in the multiset $\{i_1, i_2, i_3, i_4\}$. The following result constitutes a concentration bound for $Z = \sum_{i\in[k]} \delta_i X_i$ for a probability perturbation δ.

Theorem 7 (Probability perturbation concentration)**.** *Consider a vector δ such that $\sum_{i\in[k]} \delta_i = 0$. Let random variables $X_1, \ldots, X_k$ be 4-symmetric and $Z = \sum_{i\in[k]} \delta_i X_i$. Then, for every $\alpha \in (0, 1/4)$,*

$$\Pr\left[\left(\mathbb{E}\big[X_1^2\big] - \mathbb{E}[X_1X_2]\right) - \sqrt{\frac{38\alpha}{1-2\alpha}\mathbb{E}[X_1^4]} \le \frac{Z^2}{\|\delta\|_2^2} \le \frac{1}{1-2\alpha}\left(\mathbb{E}\big[X_1^2\big] - \mathbb{E}[X_1X_2]\right)\right] \ge \alpha.$$

The proof requires a careful evaluation of the second and the fourth moments of Z and is deferred to the full version [1]. As a corollary, we obtain the result below, which is at the core of our reduction argument.

Corollary 8. *Consider a distribution $p \in \Delta([k])$ such that $\mathrm{d}_{\mathrm{TV}}(p, u) > \gamma$. For a random subset S of $[k]$ distributed uniformly over all subsets of $[k]$ of cardinality $k/2$, it holds that $\Pr\left[\,\left|p(S) - \frac{1}{2}\right| > \frac{\gamma}{\sqrt{5k}}\,\right] > \frac{1}{477}$.*

Armed with this result, we can divide our LDP testing problem into two parts: A public-coin ε-LDP mechanism releases 1-bit per sample to the curator, and the curator applies a test to the received bits to accomplish uniformity testing. This specific mechanism suggested by the previous corollary is our Raptor (see Algorithm 3 for a description). While in this paper we have only considered its use for testing uniformity and independence, since it provides locally private 1-bit outputs that in essence preserve the ℓ_2 distance of the underlying distribution from any fixed one, we can foresee many other use-cases for Raptor and pose it as a standalone mechanism of independent interest.

Recall that in Raptor the curator and the users pick a random subset S of size $k/2$ from their shared randomness, and each user sends the indicator function that its input lies in this set S using ε-LDP RR. This is precisely the 1-bit information from samples required to enable the estimator of Fact 6. Note that when the underlying distribution p is uniform, the probability $p(S)$ of user bit being 1 is exactly $1/2$. Also, by Corollary 8 when p is γ-far from uniform we have $p(S) = 1/2 \pm \Omega(\gamma/\sqrt{k})$ with a constant probability (over the choice of S); by repeating the protocol a constant number of times,[6] we can ensure that with high constant probability at least one of the choices of S will indeed have this property. Therefore, we obtain an instance of the uniformity testing problem for $k = 2$, namely the problem of privately distinguishing a Bern(1/2) from Bern$\left(1/2 \pm \frac{c_1\gamma}{\sqrt{k}}\right)$. Thus, when we apply Raptor to the samples, the curator gets the 1-bit updates required by Fact 6 to which it can apply the estimator prescribed in Fact 6 to solve the underlying uniformity testing instance for $k = 2$ using

$$O\Big(\frac{k}{\gamma^2}\frac{(e^\varepsilon+1)^2}{(e^\varepsilon-1)^2}\Big)$$

samples. Since we used ε-LDP RR to send each bit, Raptor, too, is ε-LDP and thereby so is our overall uniformity test. This leads to the following theorem.

Theorem 9. *The test described in Algorithm 6 constitutes a uniformity test using* Raptor *with $n = O(k/(\gamma^2\varepsilon^2))$ samples.*

Algorithm 6 LDP Uniformity Testing using Raptor

Require: Privacy $\varepsilon > 0$, distance $\gamma \in (0,1)$, $n = mT$ samples

1: Set $c \leftarrow \frac{1}{477}$, $\delta \leftarrow \frac{c}{2(1+c)}$, $\gamma' \leftarrow \frac{\gamma}{\sqrt{5k}}$, $T \leftarrow \Theta(1)$
2: **for** t from 1 to T **do** ▷ In parallel
3: Generate u.a.r. $S_t \subseteq [k]$ of cardinality $k/2$
4: Apply Raptor using S_t to each sample in the mini-batch of m samples
5: Use the estimator of Fact 6 to test with probability of failure δ if $p(S_t) = 1/2$ (unbiased) or $|p(S_t) - 1/2| > \gamma'$ (biased)
6: **end for**
7: Let η denote the fraction of the T outcomes that returned unbiased
8: **if** $\eta > 1 - (\delta + \frac{c}{4})$ **return** uniform
9: **else return** not uniform

[6] To preserve the symmetry of our mechanism, we note that this can be done "in parallel" at each user. That is, each user considers the same $T = \Theta(1)$ many random subsets, and sends their corresponding T privatized (with parameter $\varepsilon' = \varepsilon/T$) indicator bits to the curator.

References

[1] Jayadev Acharya, Clément L. Canonne, Cody Freitag, and Himanshu Tyagi. Test without trust: Optimal locally private distribution testing. *CoRR*, abs/1808.02174, 2018.

[2] Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Distributed simulation and distributed inference. *CoRR*, abs/1804.06952, 2018.

[3] Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Inference under information constraints I: lower bounds from chi-square contraction. *CoRR*, abs/1812.11476, 2018.

[4] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28*, NeurIPS '15, pages 3577–3598. Curran Associates, Inc., 2015.

[5] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Communication efficient, sample optimal, linear time locally private discrete distribution estimation. *arXiv preprint arXiv:1802.04705*, 2018.

[6] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private testing of identity and closeness of discrete distributions. In *Advances in Neural Information Processing Systems 31*, NeurIPS '18, pages 6878–6891. 2018.

[7] Maryam Aliakbarpour, Ilias Diakonikolas, and Ronitt Rubinfeld. Differentially private identity and closeness testing of discrete distributions. *arXiv preprint arXiv:1707.05497*, 2017.

[8] Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics*, 12(2):727–749, 2018.

[9] Eric Blais, Clément L. Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. *ACM Trans. Comput. Theory*, 11(2):6:1–6:37, February 2019.

[10] Bryan Cai, Constantinos Daskalakis, and Gautam Kamath. Priv'it: Private and sample efficient identity testing. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 635–644. JMLR, Inc., 2017.

[11] Clément L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22(63), 2015.

[12] Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 1193–1203, Philadelphia, PA, USA, 2014. SIAM.

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

[14] Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 685–694, Washington, DC, USA, 2016. IEEE Computer Society.

[15] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *FOCS*, pages 429–438. IEEE Computer Society, 2013.

[16] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 2017.

[17] Cynthia Dwork. Differential privacy. In *ICALP (2)*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.

[18] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer.

[19] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM Conference on Computer and Communications Security*, CCS '14, pages 1054–1067, New York, NY, USA, 2014. ACM.

[20] Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, pages 211–222, New York, NY, USA, 2003. ACM.

[21] Marco Gaboardi, Hyun-Woo Lim, Ryan M. Rogers, and Salil P. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *Proceedings of the*

33rd International Conference on Machine Learning, ICML '16, pages 1395–1403. JMLR, Inc., 2016.

[22] Marco Gaboardi and Ryan Rogers. Local private hypothesis testing: Chi-square tests. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1612–1621. PMLR, 10–15 Jul 2018.

[23] Oded Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. *Electronic Colloquium on Computational Complexity (ECCC)*, 23(15), 2016.

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

[25] Dayu Huang and Sean Meyn. Generalized error exponents for small sample universal hypothesis testing. *IEEE Transactions on Information Theory*, 59(12):8157–8181, 2013.

[26] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2436–2444. JMLR.org, 2016.

[27] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826, 2011.

[28] Daniel Kifer and Ryan M. Rogers. A new class of private chi-square tests. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, AISTATS '17, pages 991–1000. JMLR, Inc., 2017.

[29] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[30] Ronitt Rubinfeld. Taming big probability distributions. *XRDS: Crossroads, The ACM Magazine for Students*, 19(1):24, sep 2012.

[31] Or Sheffet. Locally private hypothesis testing. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4612–4621, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[32] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

[33] Yue Wang, Jaewoo Lee, and Daniel Kifer. Revisiting differentially private hypothesis tests for categorical data. *arXiv preprint arXiv:1511.03376*, 2015.

[34] Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[35] Min Ye and Alexander Barg. Optimal schemes for discrete distribution estimation under local differential privacy. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 759–763, June 2017.