Priv'IT: *Priv*ate and Sample Efficient *I*dentity *T*esting

Bryan Cai [*1] Constantinos Daskalakis [*1] Gautam Kamath [*1]

Abstract

We develop differentially private hypothesis testing methods for the small sample regime. Given a sample $\mathcal{D}$ from a categorical distribution p over some domain Σ, an explicitly described distribution q over Σ, some privacy parameter ε, accuracy parameter α, and requirements β_{I} and β_{II} for the type I and type II errors of our test, the goal is to distinguish between $p = q$ and $d_{\mathrm{TV}}(p, q) \geq \alpha$. We provide theoretical bounds for the sample size $|\mathcal{D}|$ so that our method both satisfies $(\varepsilon, 0)$-differential privacy, and guarantees β_{I} and β_{II} type I and type II errors. We show that differential privacy may come for free in some regimes of parameters, and we always beat the sample complexity resulting from running the χ^2-test with noisy counts, or standard approaches such as repetition for endowing non-private χ^2-style statistics with differential privacy guarantees. We experimentally compare the sample complexity of our method to that of recently proposed methods for private hypothesis testing (Gaboardi et al., 2016; Kifer & Rogers, 2017).

1. Introduction

Hypothesis testing is the age-old problem of deciding whether observations from an unknown phenomenon p conform to a model q. Often p can be viewed as a distribution over some alphabet Σ, and the goal is to determine, using samples from p, whether it is equal to some model distribution q or not. This type of test is the lifeblood of the scientific method and has received tremendous study in statistics since its very beginnings. Naturally, the focus has been on minimizing the number of observations from the unknown distribution p that are needed to determine, with confidence, whether $p = q$ or $p \neq q$.

In several fields of research and application, however, samples may contain sensitive information about individuals; consider for example, individuals participating in some clinical study of a disease that carries social stigma. It may thus be crucial to guarantee that operating on the samples needed to test a statistical hypothesis protects sensitive information about the samples. This is not at odds with the goal of hypothesis testing itself, since the latter is about verifying a property of the population p from which the samples are drawn, and not of the samples themselves.

Without care, however, sensitive information about the sample might actually be divulged by statistical processing that is improperly designed. As recently exhibited, for example, it may be possible to determine whether individuals participated in a study from data that would typically be published in genome-wide association studies (Homer et al., 2008). Motivated in part by this realization, there has been increased recent interest in developing data sharing techniques which are private (Johnson & Shmatikov, 2013; Uhler et al., 2013; Yu et al., 2014; Simmons et al., 2016).

Protecting privacy when computing on data has been extensively studied in several fields ranging from statistics to diverse branches of computer science including algorithms, cryptography, database theory, and machine learning; see, e.g., (Dalenius, 1977; Adam & Worthmann, 1989; Agrawal & Aggarwal, 2001; Dinur & Nissim, 2003; Dwork, 2008; Dwork & Roth, 2014) and their references. A notion of privacy proposed by theoretical computer scientists which has found a lot of traction is that of *differential privacy* (Dwork et al., 2006). Roughly speaking, it requires that the output of an algorithm on two neighboring datasets D and D' that differ in the value of one element be statistically close. For a formal definition see Section 2.

Our goal in this paper is to develop tools for privately performing statistical hypothesis testing. In particular, we are interested in studying the tradeoffs between statistical accuracy, power, significance, and privacy in the sample size. To be precise, given samples from a categorical distribution p over some domain Σ, an explicitly described distribution q over Σ, some privacy parameter ε, accuracy parameter α, and requirements β_{I} and β_{II} for the type I and type II errors of our test, the goal is to distinguish between $p = q$

[*] Equal contribution [1] Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Correspondence to: Bryan Cai <bcai@mit.edu>, Constantinos Daskalakis <costis@csail.mit.edu>, Gautam Kamath <g@csail.mit.edu>.

Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

and $d_{\mathrm{TV}}(p,q) \geq \alpha$. We want that the output of our test be $(\varepsilon,0)$-differentially private, and that the probability we make a type I or type II error be β_{I} and β_{II} respectively. Treating these as hard constraints, we want to *minimize the number of samples that we draw from* p.

Notice that the *correctness* constraint on our test pertains to whether we draw the right conclusion about how p compares to q, while the *privacy* constraint pertains to whether we respect the privacy of the samples that we draw from p. The pertinent question is how much the privacy constraint increases the number of samples that are needed to guarantee correctness. Our main result is that privacy may come for free in certain regimes of parameters, and has a mild cost for all regimes of parameters.

To be precise, *without privacy constraints*, it is well known that identity testing can be performed from $O(\frac{\sqrt{n}}{\alpha^2} \cdot \log \frac{1}{\beta})$ samples, where n is the size of Σ and $\beta = \min\{\beta_{\mathrm{I}}, \beta_{\mathrm{II}}\}$, and that this is tight (Batu et al., 2001; Paninski, 2008; Valiant & Valiant, 2014; Acharya et al., 2015). Our main theoretical result is that, *with privacy constraints*, the number of samples that are needed is

$$\tilde{O}\left(\max\left\{\frac{\sqrt{n}}{\alpha^2}, \frac{\sqrt{n}}{\alpha^{3/2}\varepsilon}, \frac{n^{1/3}}{\alpha^{5/3}\varepsilon^{2/3}}\right\} \cdot \log(1/\beta)\right). \quad (1)$$

Our statistical test is provided in Section 5 where the above upper bound on the number of samples that it requires is proven as Theorem 3. Notice that privacy comes for free when the privacy requirement ε is $\Omega(\sqrt{\alpha})$ – for example when $\varepsilon = 10\%$ and the required statistical accuracy is 3%.

The precise constants sitting in the $O(\cdot)$ notation of Eq. (1) are given in the proof of Theorem 3. We experimentally verify the sample efficiency of our tests by comparing them to recently proposed private statistical tests (Gaboardi et al., 2016; Kifer & Rogers, 2017), discussed in more detail shortly. Fixing a differential privacy and type I, type II error constraints, we compare how many samples are required by our and their methods to distinguish between hypotheses that are $\alpha = 0.1$ apart in total variation distance. We find that different algorithms are more efficient depending on the regime and properties desired by the analyst. Our experiments and further discussion of the tradeoffs are presented in Section 6.

Approach. A standard approach to turn an algorithm differentially private is to use repetition. As already mentioned above, absent differential privacy constraints, statistical tests have been provided that use an optimal $m = O(\frac{\sqrt{n}}{\alpha^2} \cdot \log \frac{1}{\beta})$ number of samples. A trivial way to get $(\varepsilon,0)$-differential privacy using such a non-private test is to create $O(1/\varepsilon)$ datasets, each comprising m samples from p, and run the non-private test on one of these datasets, chosen randomly. It is clear that changing the value of a single element in the combined dataset may only affect the output of the test with probability at most ε. Thus the output is $(\varepsilon,0)$-differentially private; see Section 3 for a proof. The issue with this approach is that the total number of samples that it draws is $m/\varepsilon = O(\frac{\sqrt{n}}{\varepsilon\alpha^2} \cdot \log \frac{1}{\beta})$, which is higher than our target. See Corollary 1.

A different approach towards private hypothesis testing is to look deeper into the non-private tests and try to "privatize" them. The most sample-efficient tests are variations of the classical χ^2-test. They compute the number of times, N_i, that element $i \in \Sigma$ appears in the sample and aggregate those counts using a statistic that equals, or is close to, the χ^2-divergence between the empirical distribution defined by these counts and the hypothesis distribution q. They accept q if the statistic is low and reject q if it is high, using some threshold.

A reasonable approach to privatize such a test is to add noise, e.g. Laplace$(1/\varepsilon)$ noise, to each count N_i, before running the test. It is well known that adding Laplace$(1/\varepsilon)$ noise to a set of counts makes them differentially private, see Theorem 1. However, it also increases the variance of the statistic. This has been noticed empirically in recent work of (Gaboardi et al., 2016) for the χ^2-test. We show that the variance of the optimal χ^2-style test statistic significantly increases if we add Laplace noise to the counts, in Section 4.1, thus increasing the sample complexity from $O(\sqrt{n})$ to $\Omega(n^{3/4})$. So this route, too, seems problematic.

A last approach towards designing differentially private tests is to exploit the distance beween the null and the alternative hypotheses. A correct test should accept the null with probability close to 1, and reject an alternative that is α-far from the null with probability close to 1, but there are no requirements for correctness when the alternative is very close to the null. We could thus try to interpolate smoothly between datasets that we expect to see when sampling the null and datasets that we expect to see when sampling an alternative that is far from the null. Rather than outputting "accept" or "reject" by merely thresholding our statistic, we would like to tune the probability that we output "reject" based on the value of our statistic, and make it so that the "reject" probability is ε-Lipschitz as a function of the dataset. Moreover, the probability should be close to 0 on datasets that we expect to see under the null and close to 1 on datasets that we expect to see under an alternative that is α-far. As we show in Section 4.2, χ^2-style statistics have high sensitivity, requiring $\omega(\sqrt{n})$ samples to be made appropriately Lipschitz.

While both the approach of adding noise to the counts, and that of turning the output of the test Lipschitz fail in isolation, our test actually goes through by intricately combining these two approaches. It has two steps:

1. A *filtering step,* whose goal is to "reject" when p is blatantly far from q. This step is performed by comparing the counts N_i with their expectations under q, after having added Laplace$(1/\varepsilon)$ noise to these counts. If the noisy counts deviate from their expectation, taking into account the extra variance introduced by the noise, then we can safely "reject." Moreover, because noise was added, this step is differentially private.

2. If the filtering step fails to reject, we perform a *statistical step.* This step just computes the χ^2-style statistic from (Acharya et al., 2015), *without adding noise to the counts.* The crucial observation is that if the filtering step does not reject, then the statistic is actually ε-Lipschitz with respect to the counts, and thus the value of the statistic is still differentially private. We use the value of the statistic to determine the bias of a coin that outputs "reject."

Details of our test are given in Section 5.

Related Work. Identity testing is one of the most classical problems in statistics, where it is traditionally called hypothesis or goodness-of-fit testing, see (Pearson, 1900; Fisher, 1935; Rao & Scott, 1981; Agresti, 2012) for some classical and contemporary references. In this field, the focus is often on asymptotic analysis, where the number of samples goes to infinity, and we wish to get a grasp on their asymptotic distributions and error exponents (Agresti, 2012; Tan et al., 2010). In the past twenty years, this problem has enjoyed significant interest in the theoretical computer science community (see, i.e., (Batu et al., 2001; Paninski, 2008; Levi et al., 2013; Valiant & Valiant, 2014; Acharya et al., 2015; Canonne et al., 2016; Diakonikolas & Kane, 2016; Daskalakis et al., 2016), and (Canonne, 2015) for a survey), where the focus has instead been on the finite sample regime, rather than asymptotics. Specifically, the goal is to minimize the number of samples required, while still remaining computationally tractable.

A number of recent works (Wang et al., 2015; Gaboardi et al., 2016; Kifer & Rogers, 2017) (and a simultaneous work, focused on independence testing (Kakizaki et al., 2017)) investigate differential privacy with the former set of goals. In particular, their algorithms focus on fixing a desired significance (type I error) and privacy requirement, and study the asymptotic distribution of the test statistics. On the other hand, we are the first work to apply differential privacy to the latter line of inquiry, where our goal is to minimize the number of samples required to ensure the desired significance, power and privacy. As a point of comparison between these two worlds, we provide an empirical evaluation of our method versus their methods.

The problem of distribution *estimation* (rather than testing) has also recently been studied under the lens of differential privacy (Diakonikolas et al., 2015). This is another classical statistics problem which has recently piqued the interest of the theoretical computer science community. We note that the techniques required for this setting are quite different from ours, as we must deal with issues that arise from very sparsely sampled data.

2. Preliminaries

In this paper, we will focus on discrete probability distributions over $[n]$. For a distribution p, we will use the notation p_i to denote the mass p places on symbol i.

Definition 1. *The* total variation distance *between p and q is defined as*

$$d_{\mathrm{TV}}(p,q) = \frac{1}{2}\sum_{i\in[n]} |p_i - q_i| \,.$$

Definition 2. *A randomized algorithm M with domain $\mathbb{N}^n$ is (ε,δ)*-differentially private *if for all $S \subseteq \mathrm{Range}(M)$ and for all pairs of inputs D, D' such that $\|D - D'\|_1 \le 1$:*

$$\Pr\left[M(D)\in S\right] \le e^{\varepsilon}\Pr\left[M(D')\in S\right] + \delta.$$

If $\delta = 0$, the guarantee is called pure *differential privacy.*

In the context of distribution testing, the neighboring dataset definition corresponds to two datasets where one dataset is generated from the other by removing one sample. Up to a factor of 2, this is equivalent to the alternative definition where one dataset is generated from the other by arbitrarily changing one sample.

Definition 3. *An algorithm for the $(\alpha, \beta_{\mathrm{I}}, \beta_{\mathrm{II}})$*-identity testing *problem with respect to a (known) distribution q takes m samples from an (unknown) distribution p and has the following guarantees:*

- *If $p = q$, then with probability at least $1-\beta_{\mathrm{I}}$ it outputs "$p = q$;"*
- *If $d_{\mathrm{TV}}(p,q) \ge \alpha$, then with probability at least $1-\beta_{\mathrm{II}}$ it outputs "$p \ne q$."*

In particular, β_{I} and β_{II} are the type I and type II errors of the test. Parameter α is the radius of distinguishing accuracy. Notice that, when p satisfies neither of cases above, the algorithm's output may be arbitrary.

We note that if an algorithm is to satisfy both these definitions, the latter condition (the *correctness* property) need only be satisfied when p falls into one of the two cases, while the former condition (the *privacy* property) must be satisfied for *all realizations* of the samples from p (and in particular, for p which do not fall into the two cases above).

We recall the classical Laplace mechanism, which states that applying independent Laplace noise to a set of counts is differentially private.

Theorem 1 (Theorem 3.6 of (Dwork & Roth, 2014))**.** *Given a set of counts $N_1, \ldots, N_n$, the noised counts $(N_1 + Y_1, \ldots, N_n + Y_n)$ are $(\varepsilon, 0)$-differentially private when the Y_i's are i.i.d. random variables drawn from $Laplace(1/\varepsilon)$.*

Finally, we recall the definition of zero-concentrated differential privacy from (Bun & Steinke, 2016) and its relationship to differential privacy.

Definition 4. *A randomized algorithm M with domain $\mathbb{N}^n$ is ρ*-zero-concentrated differentially private *(ρ-zCDP) if for all pairs of inputs D, D' such that $\|D - D'\|_1 \le 1$ and all $\alpha \in (1, \infty)$:*

$$\mathrm{D}_\alpha(M(D)||M(D')) \le \rho\alpha,$$

where D_α is the α-Rényi divergence between the distribution of $M(D)$ and $M(D')$.

Proposition 1 (Propositions 1.3 and 1.4 of (Bun & Steinke, 2016))**.** *If a mechanism M_1 satisfies $(\varepsilon, 0)$-differential privacy, then M_1 satisfies $\frac{\varepsilon^2}{2}$-zCDP. If a mechanism M_2 satisfies ρ-zCDP, then M_2 satisfies $(\rho + 2\sqrt{\rho\log(1/\delta)}, \delta)$-differential privacy for any $\delta > 0$.*

3. A Simple Upper Bound

In this section, we provide an $O\left(\frac{\sqrt{n}}{\alpha^2\varepsilon}\right)$ upper bound for the differentially private identity testing problem. More generally, we show that if an algorithm requires a dataset of size m for a decision problem, then it can be made $(\varepsilon, 0)$-differentially private at a multiplicative cost of $1/\varepsilon$ in the sample size. This is a folklore result, but we include and prove it here for completeness.

Theorem 2. *Suppose there exists an algorithm for a decision problem P which succeeds with probability at least $1 - \beta$ and requires a dataset of size m. Then there exists an $(\varepsilon, 0)$-differentially private algorithm for P which succeeds with probability at least $\frac{4}{5}(1 - \beta) + 1/10$ and requires a dataset of size $O(m/\varepsilon)$.*

Proof. First, with probability $1/5$, we flip a coin and output yes or no with equal probability. This guarantees that we have probability at least $1/10$ of either outcome, which will allow us to satisfy the multiplicative guarantee of differential privacy.

We then draw $10/\varepsilon$ datasets of size m, and solve the decision problem (non-privately) for each of them. Finally, we select a random one of these computations and output its outcome.

The correctness follows, since we randomly choose the right answer with probability $1/10$, or with probability $4/5$, we solve the problem correctly with probability $1 - \beta$. As for privacy, we note that, if we remove a single element of the dataset, we may only change the outcome of one of these computations. Since we pick a random computation, this is selected with probability $\varepsilon/10$, and thus the probability of any outcome is additively shifted by at most $\varepsilon/10$. Since we know the minimum probability of any output is $1/10$, this gives the desired multiplicative guarantee required for $(\varepsilon, 0)$-differential privacy. □

We obtain the following corollary by noting that the tester of (Acharya et al., 2015) (among others) requires $O(\sqrt{n}/\alpha^2)$ samples for identity testing.

Corollary 1. *There exists an $(\varepsilon, 0)$-differentially private testing algorithm for the $(\alpha, \beta_{\mathrm{I}}, \beta_{\mathrm{II}})$-identity testing problem for any distribution q which requires*

$$m = O\left(\frac{\sqrt{n}}{\varepsilon\alpha^2} \cdot \log(1/\beta)\right)$$

samples, where $\beta = \min(\beta_{\mathrm{I}}, \beta_{\mathrm{II}})$.

4. Roadblocks to Differentially Private Testing

In this section, we describe roadblocks which prevent two natural approaches to differentially private testing from working.

In Section 4.1, we show that if one simply adds Laplace noise to the empirical counts of a dataset (i.e., runs the Laplace mechanism of Theorem 1) and then attempts to run an optimal identity tester, the variance of the statistic increases dramatically, and thus results in a much larger sample complexity, even for the case of uniformity testing. The intuition behind this phenomenon is as follows. When performing uniformity testing in the small sample regime (when the number of samples m is the square root of the domain size n), we will see a $(1 - o(1))n$ elements 0 times, $O(\sqrt{n})$ elements 1 time, and $O(1)$ elements 2 times. If we add $Laplace(10)$ noise to guarantee $(0.1, 0)$-differential privacy, this obliterates the signal provided by these collision statistics, and thus many more samples are required before the signal prevails.

In Section 4.2, we demonstrate that χ^2 statistics have high sensitivity, and thus are not naturally differentially private. In other words, if we consider a χ^2 statistic Z on two datasets D and D' which differ in one record, $|Z(D) - Z(D')|$ may be quite large. This implies that methods such as rescaling this statistic and interpreting it as a probability, or applying noise to the statistic, will not be differentially private until we have taken a large number of samples.

4.1. A Laplaced χ^2-statistic has large variance

Proposition 2. *Applying the Laplace mechanism to a dataset before applying the identity tester of (Acharya et al., 2015) results in a significant increase in the variance, even when considering the case of uniformity. More precisely, if we consider the statistic*

$$Z'(D) = \sum_{i \in [n]} \frac{(N_i + Y_i - m/n)^2 - (N_i + Y_i)}{m/n}$$

where N_i is the number of occurrences of symbol i in the dataset D (which is of size $Poisson(m)$) and $Y_i \sim Laplace(1/\varepsilon)$, then

- *If p is uniform, then $\mathbf{E}[Z'] = \frac{2n^2}{\varepsilon^2 m}$ and $\mathbf{Var}[Z'] \geq \frac{20n^3}{\varepsilon^4 m^2}$.*
- *If p is a particular distribution which is α-far in total variation distance from uniform, then $\mathbf{E}[Z'] = 4m\alpha^2 + \frac{2n^2}{\varepsilon^2 m}$.*

The variance of the statistic can be compared to that of the unnoised statistic, which is upper bounded by $m^2\alpha^4$. We can see that the noised statistic has larger variance until $m = \Omega(n^{3/4})$.

Proof. First, we compute the mean of Z'. Note that since $|D| \sim Poisson(m)$, the N_i's will be independently distributed as $Poisson(mp_i)$ (see, i.e., (Acharya et al., 2015) for additional discussion).

$$\begin{aligned}
\mathbf{E}[Z'] &= \mathbf{E}\left[\sum_{i \in [n]} \frac{(N_i + Y_i - m/n)^2 - (N_i + Y_i)}{m/n}\right] \\
&= \mathbf{E}\left[\sum_{i \in [n]} \frac{(N_i - m/n)^2 - N_i}{m/n}\right. \\
&\quad \left. + \sum_{i \in [n]} \frac{Y_i^2 + 2Y_i(N_i - m/n) - Y_i}{m/n}\right] \\
&= m \cdot \chi^2(p, q) + \sum_{i \in [n]} \frac{\frac{2}{\varepsilon^2}}{m/n} \\
&= m \cdot \chi^2(p, q) + \frac{2n^2}{\varepsilon^2 m}
\end{aligned}$$

In other words, the mean is a rescaling of the χ^2 distance between p and q, shifted by some constant amount. When $p = q$, the χ^2-distance between p and q is 0, and the expectation is just the second term. Focus on the case where n is even, and consider p such that $p_i = (1 + 2\alpha)/n$ if i is even, and $(1 - 2\alpha)/n$ otherwise. This is α-far from uniform in total variation distance. Furthermore, by direct calculation, $\chi^2(p, q) = 4\alpha^2$, and thus the expectation of Z' in this case is $4m\alpha^2 + \frac{2n^2}{\varepsilon^2 m}$.

Next, we examine the variance of Z'. Let $\lambda_i = mp_i$ and $\lambda'_i = mq_i = m/n$. By a similar computation as before, we have that

$$\begin{aligned}
\mathbf{Var}[Z'] = \sum_{i \in [n]} \frac{1}{\lambda_i'^2} \Big[& 2\lambda_i^2 + 4\lambda_i(\lambda_i - \lambda'_i)^2 \\
& + \frac{1}{\varepsilon^2}(8\lambda_i + 2(2\lambda_i - 2\lambda'_i - 1)^2) + \frac{20}{\varepsilon^4} \Big].
\end{aligned}$$

Since all four summands of this expression are non-negative, we have that

$$\mathbf{Var}[Z'] \geq \frac{20}{\varepsilon^4} \sum_{i \in [n]} \frac{1}{\lambda_i'^2} = \frac{20n^3}{\varepsilon^4 m^2}.$$

If we wish to use Chebyshev's inequality to separate these two cases, we require that $\mathbf{Var}[Z']$ is at most the square of the mean separation. In other words, we require that $\frac{20n^3}{\varepsilon^4 m^2} \leq m^2\alpha^4$, or that $m = \Omega\left(\frac{n^{3/4}}{\varepsilon\alpha}\right)$. □

4.2. A χ^2-statistic has high sensitivity

Consider the primary statistic which we use in Algorithm 1:

$$Z(D) = \frac{1}{m\alpha^2} \sum_{i \in [n]} \frac{(N_i - mq_i)^2 - N_i}{mq_i}.$$

As shown in Section 5, $\mathbf{E}[Z] = 0$ if $p = q$ and $\mathbf{E}[Z] \geq 1$ if $d_{\mathrm{TV}}(p, q) \geq \alpha$, and the variance of Z is such that these two cases can be separated with constant probability. A natural approach is to truncate this statistic to the range $[0, 1]$, interpret it as a probability and output the result of $Bernoulli(Z)$ – if $p = q$, the result is likely to be 0, and if $d_{\mathrm{TV}}(p, q) \geq \alpha$, the result is likely to be 1. One might hope that this statistic is naturally private. More specifically, we would like that the statistic Z has low sensitivity, and does not change much if we remove a single individual. Unfortunately, this is not the case. We consider datasets D, D', where D' is identical to D, but with one fewer occurrence of symbol i. It can be shown that the difference in Z is

$$|Z(D) - Z(D')| = \frac{2|N_i - mq_i - 1|}{m^2\alpha^2 q_i}$$

Letting q be the uniform distribution and requiring that this is at most ε (for the sake of privacy), we have a constraint which is roughly of the form $\frac{2N_i n}{m^2\alpha^2} \leq \varepsilon$, or that $m = \Omega\left(\frac{\sqrt{N_i}\sqrt{n}}{\varepsilon^{0.5}\alpha}\right)$.

In particular, if $N_i = n^c$ for any $c > 0$, this does not achieve the desired $O(\sqrt{n})$ sample complexity. One may observe that, if N_i is this large, looking at symbol i alone is sufficient to conclude p is not uniform, even if the count N_i had Laplace noise added. Indeed, our main algorithm of Section 5 works in part due to our formalization and quantification of this intuition.

5. Priv'IT: A Differentially Private Identity Tester

In this section, we sketch the proof of our main testing upper bound:

Theorem 3. *There exists an $(\varepsilon, 0)$-differentially private testing algorithm for the $(\alpha, \beta_{\mathrm{I}}, \beta_{\mathrm{II}})$-identity testing problem for any distribution q which requires*

$$m = \tilde{O}\left(\max\left\{\frac{\sqrt{n}}{\alpha^2}, \frac{\sqrt{n}}{\alpha^{3/2}\varepsilon}, \frac{n^{1/3}}{\alpha^{5/3}\varepsilon^{2/3}}\right\} \cdot \log(1/\beta)\right)$$

samples, where $\beta = \min(\beta_{\mathrm{I}}, \beta_{\mathrm{II}})$.

The full details of the proof are provided in the supplementary materials.

The pseudocode for this algorithm is provided in Algorithm 1. We fix the constants $c_1 = 1/4$ and $c_2 = 3/40$. For a high-level overview of our algorithm's approach, we refer the reader to the Approach paragraph in Section 1.

Algorithm 1 Priv'IT: A differentially private identity tester

```
1: Input: $\varepsilon$; an explicit distribution $q$; sample access to a distribution $p$
2: Define $\mathcal{A} \leftarrow \{i : q_i \geq c_1\alpha/n\}$, $\bar{\mathcal{A}} \leftarrow [n] \setminus \mathcal{A}$
3: Sample $Y_i \sim Laplace(2/c_2\varepsilon)$ for all $i \in \mathcal{A}$
4: if there exists $i \in \mathcal{A}$ such that $|Y_i| \geq \frac{2}{c_2\varepsilon}\log\left(\frac{1}{1-(1-c_2)^{1/|\mathcal{A}|}}\right)$ then
5:     return either "$p \neq q$" or "$p = q$" with equal probability
6: end if
7: Draw a multiset $S$ of $Poisson(m)$ samples from $p$
8: Let $N_i$ be the number of occurrences of the $i$th domain element in $S$
9: for $i \in \mathcal{A}$ do
10:     if $|N_i + Y_i - mq_i| \geq \frac{2}{c_2\varepsilon}\log\left(\frac{1}{1-(1-c_2)^{1/|\mathcal{A}|}}\right) + \max\left\{4\sqrt{mq_i \log n}, \log n\right\}$ then
11:         return "$p \neq q$"
12:     end if
13: end for
14: $Z \leftarrow \frac{2}{m\alpha^2}\sum_{i\in\mathcal{A}} \frac{(N_i - mq_i)^2 - N_i}{mq_i}$
15: Let $T$ be the closest value to $Z$ which is contained in the interval $[0, 1]$
16: Sample $b \sim Bernoulli(T)$
17: if $b = 1$ then
18:     return "$p \neq q$"
19: else
20:     return "$p = q$"
21: end if
```

Proof of Theorem 3 (sketch): We focus on the case where $\beta = 1/3$, the general case follows at the cost of a multiplicative $\log(1/\beta)$ in the sample complexity from a standard amplification argument. We will require the following tail bounds on N_i and Y_i.

Claim 1. $|Y_i| \leq \frac{2}{c_2\varepsilon}\log\left(\frac{1}{1-(1-c_2)^{1/|\mathcal{A}|}}\right)$ *simultaneously for all $i \in \mathcal{A}$ with probability exactly $1 - c_2$.*

Claim 2. $|N_i - mp_i| \leq \max\left\{4\sqrt{mp_i \log n}, \log n\right\}$ *simultaneously for all $i \in \mathcal{A}$ with probability at least $1 - \frac{2}{n^{0.84}} - \frac{1.1}{n}$.*

Correctness. Correctness can be shown in a similar way to (Acharya et al., 2015) – in short, if $m = \Omega(\sqrt{n}/\alpha^2)$, then the expectations are separated in the two cases, and the variance is bounded. A careful combination of the previous claims and Chebyshev's inequality guarantee correctness.

Privacy. We will prove $(0, c_2\varepsilon/2)$-differential privacy, which in our setting, will imply $(\varepsilon, 0)$-differential privacy (due to Claim 1).

We first consider the possibility of rejecting in line 11. Noising our counts by the random variables Y_i ensures that this step is $(0, c_2\varepsilon/4)$-differentially private.

Consider the difference in value of Z for two neighboring datasets D and D', differing in i: $Z(D) - Z(D') = \frac{2(N_i - mq_i - 1)}{m^2\alpha^2 q_i}$. Conditioning on the event that we did not return in line 11, we can show

$$|N_i - mq_i| \leq \frac{4\log(n/c_2)}{c_2\varepsilon} + \max\left\{4\sqrt{mq_i \log n}, \log n\right\}.$$

This implies that $|Z(D) - Z(D')| \leq \frac{2}{m^2\alpha^2 q_i}\left(\frac{6\log(n/c_2)}{c_2\varepsilon} + 4\sqrt{mq_i \log n}\right)$. Enforcing that each of these terms are at most $c_2\varepsilon/8$ gives the condition $m \geq \max\left\{\sqrt{\frac{96}{c_2^2 c_1}}\frac{\sqrt{n\log(n/c_2)}}{\alpha^{1.5}\varepsilon}, \left(\frac{64}{c_2\sqrt{c_1}}\right)^{2/3}\frac{(n\log n)^{1/3}}{\alpha^{5/3}\varepsilon^{2/3}}\right\}$. Since both terms are at most $c_2\varepsilon/8$, this step is $(0, c_2\varepsilon/4)$-differentially private. By composition of differential privacy, this gives the desired overall $(0, c_2\varepsilon/2)$-differential privacy and thus ε-pure differential privacy. □

6. Experiments

We performed an empirical evaluation of our algorithm, `Priv'IT`, on synthetic datasets. All experiments were performed on a laptop computer with a 2.6 GHz Intel Core i7-6700HQ CPU and 8 GB of RAM. Significant discussion is required to provide a full comparison with prior work in this area, since performance of the algorithms varies depending on the regime.

We compared our algorithm with two recent algorithms for differentially private hypothesis testing:

1. The Monte Carlo Goodness of fit test with Laplace noise from (Gaboardi et al., 2016), `MCGOF`;

2. The projected Goodness of Fit test from (Kifer & Rogers, 2017), `zCDP-GOF`.

We note that we implemented a modified version of `Priv'IT`, which differs from Algorithm 1 in lines 14 to 21. In particular, we instead consider a statistic

$$Z = \sum_{i \in \mathcal{A}} \frac{(N_i - mq_i)^2 - N_i}{mq_i}.$$

We add Laplace noise to Z, with scale parameter $\Theta(\Delta/\varepsilon)$, where Δ is the sensitivity of Z, which guarantees $(\varepsilon/2, 0)$-differential privacy. Then, similar to the other algorithms, we choose a threshold for this noised statistic such that we have the desired type I error. This algorithm can be analyzed to provide identical theoretical guarantees as Algorithm 1, but with the practical advantage that there are fewer parameters to tune.

To begin our experimental evaluation, we started with uniformity testing. Our experimental setup was as follows. The algorithms were provided q as the uniform distribution over $[n]$. The algorithms were also provided with samples from some distribution p. This (unknown) p was q for the case $p = q$, or a distribution which we call the "Paninski construction" for the case $d_{\mathrm{TV}}(p, q) \geq \alpha$. The Paninski construction is a distribution where half the elements of the support have mass $(1+\alpha)/n$ and half have mass $(1-\alpha)/n$. We use this name for the construction as (Paninski, 2008) showed that this example is one of the hardest to distinguish from uniform: one requires $\Omega(\sqrt{n}/\alpha^2)$ samples to (non-privately) distinguish a random permutation of this construction from the uniform distribution. We fixed parameters $\varepsilon = 0.1$ and $\alpha = 0.1$. In addition, recall that Proposition 1 implies that pure differential privacy (the privacy guaranteed by `Priv'IT`) is stronger than zCDP (the privacy guaranteed by `zCDP-GOF`). In particular, our guarantee of ε-pure differential privacy implies $\varepsilon^2/2$-zCDP. As a result, we ran `zCDP-GOF` with a privacy parameter of 0.005-zCDP, which is equivalent to the amount of zCDP our algorithm provides. Our experiments were conducted on a number of different support sizes n, ranging from 10 to 10600. For each n, we ran the testing algorithms with increasing sample sizes m in order to discover the minimum sample size when the type I and type II errors were both empirically below $1/3$. To determine these empirical error rates, we ran all algorithms 1000 times for each n and m, and recorded the fraction of the time each algorithm was correct. As the other algorithms take a parameter β_{I} as a target type I error, we input $1/3$ as this parameter.

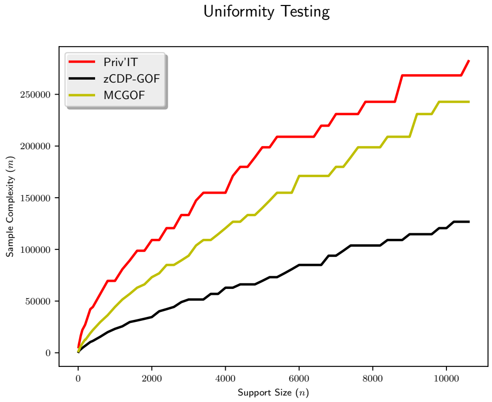

Figure 1. The sample complexities of `Priv'IT`, `MCGOF`, and `zCDP-GOF` for uniformity testing

The results of our first test are provided in Figure 1. The x-axis indicates the support size, and the y-axis indicates the minimum number of samples required. We plot three lines, which demonstrate the empirical number of samples required to obtain $1/3$ type I and type II error for the different algorithms. We can see that in this case, `zCDP-GOF` is the most statistically efficient, followed by `MCGOF` and `Priv'IT`.

To explain this difference in statistical efficiency, we note that the theoretical guarantees of `Priv'IT` imply that it performs well even when data is sparsely sampled. More precisely, one of the benefits of our tester is that it can reduce the variance induced by elements whose expected number of occurrences is less than 1. Since none of these testers reach this regime (i.e., even `zCDP-GOF` at $n = 10000$ expects to see each element 10 times), we do not reap the benefits of `Priv'IT`. Ideally, we would run these algorithms on the uniform distribution at sufficiently large support sizes. However, since this is prohibitively expensive to do with thousands of repetitions (for any of these methods), we instead demonstrate the advantages of our tester on a different distribution.

Our second test is conducted with q being a 2-histogram[1], where all but a vanishing fraction of the probability mass is concentrated on a small, constant fraction of the support[2]. This serves as our proxy for a very large support, since now we will have elements which have a sub-constant expected number of occurrences. The algorithms are provided with samples from a distribution p, which is either q or a similar Paninski construction as before, where the total variation distance from q is placed on the support elements containing non-negligible mass. We ran the test on support sizes n ranging from 10 to 6800. All other parameters are the same

[1] A k-histogram is a distribution where the domain can be partitioned into k intervals such that the distribution is uniform over each interval.

[2] In particular, in Figure 3, $n/200$ support elements contained $1 - 10/n$ probability mass, but similar trends hold with modifications of these parameters.

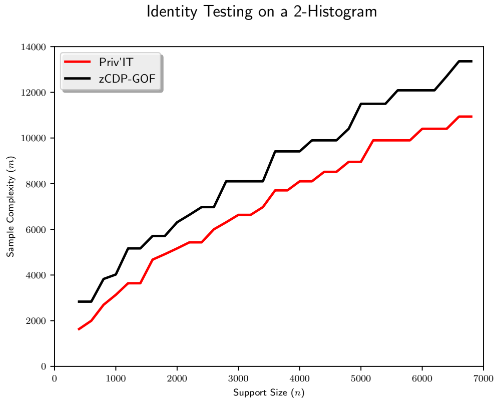

Figure 2. The sample complexities of `Priv'IT` and `zCDP-GOF` for identity testing on a 2-histogram

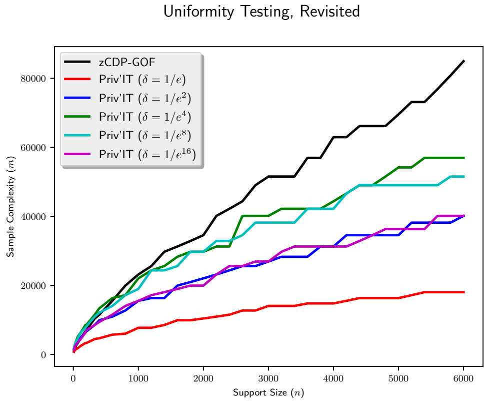

Figure 3. The sample complexities of `Priv'IT` and `zCDP-GOF` for uniformity testing, with approximate differential privacy

as in the previous test.

The results of our second test are provided in Figure 2. In this case, we compare `Priv'IT` and `zCDP-GOF`, and note that our test is slightly better for all support sizes n, though the difference can be pronounced or diminished depending on the construction of the distribution q. We found that `MCGOF` was incredibly inefficient on this construction – even for $n = 400$ it required 130000 samples, which is a factor of 10 worse than `zCDP-GOF` on a support of size $n = 6800$. To explain this phenomenon, we can inspect the contribution of a single domain element i to their statistic:

$$\frac{(N_i + Y_i - mq_i)^2}{mq_i}.$$

In the case where $mq_i \ll 1$ and $p = q$, this is approximately equal to $\frac{Y_i^2}{mq_i}$. The standard deviation of this term will be of the order $\frac{1}{mq_i\varepsilon^2}$, which can be made arbitrarily large as $mq_i \to 0$. While `zCDP-GOF` may naively seem susceptible to this same pitfall, their projection method appears to elegantly avoid it.

As a final test, we note that `zCDP-GOF` guarantees zCDP, while `Priv'IT` guarantees (vanilla) differential privacy. In our previous tests, our guarantee was ε-differential privacy, while theirs was $\frac{\varepsilon^2}{2}$-zCDP: by Proposition 1, our guarantees imply theirs. In the third test, we revisit uniformity testing, but when *their guarantees imply ours*. More specifically, again with $\varepsilon = 0.1$, we ran `zCDP-GOF` with the guarantee of $\frac{\varepsilon^2}{2}$-zCDP and `Priv'IT` with the guarantee of $(\frac{\varepsilon^2}{2} + \varepsilon\sqrt{2\log(1/\delta)}, \delta)$ for various $\delta > 0$. We note that δ is often thought in theory to be "cryptographically small" (such as 2^{-100}), but we compare with a wide range of δ, both large and small: $\delta = 1/e^t$ for $t \in \{1, 2, 4, 8, 16\}$. This test was conducted on support sizes n ranging from 10 to 6000.

The results of our third test are provided in Figure 3. We found that, for all δ tested, `Priv'IT` required fewer samples than `zCDP-GOF`. This is unsurprising for δ very large and small, since the differential privacy guarantees become very easy to satisfy, but we found it to be true for even "moderate" values of δ. This implies that if an analyst is satisfied with approximate differential privacy, she might be better off using `Priv'IT`, rather than an algorithm which guarantees zCDP.

While the main focus of our evaluation was statistical in nature, we will note that `Priv'IT` was more efficient in runtime than our implementation of `MCGOF`, and more efficient in memory usage than our implementation of `zCDP-GOF`. The former point was observed by noting that, in the same amount of time, `Priv'IT` was able to reach a trial corresponding to a support size of 20000, while `MCGOF` was only able to reach 10000. The latter point was observed by noting that `zCDP-GOF` ran out of memory at a support size of 11800. This is likely because `zCDP-GOF` requires matrix computations on a matrix of size $O(n^2)$. It is plausible that all of these implementations could be made more time and memory efficient, but we found our implementations to be sufficient for the sake of our comparison.

Acknowledgments

The authors would like to thank Jon Ullman for helpful discussions in the early stages of this work. The authors were supported by NSF CCF-1551875, CCF-1617730, CCF-1650733, and ONR N00014-12-1-0999.

References

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

Adam, Nabil R. and Worthmann, John C. Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys (CSUR)*, 21(4):515–556, 1989.

Agrawal, Dakshi and Aggarwal, Charu C. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '01, pp. 247–255, New York, NY, USA, 2001. ACM.

Agresti, Alan. *Categorical Data Analysis*. Wiley, 2012.

Batu, Tugkan, Fischer, Eldar, Fortnow, Lance, Kumar, Ravi, Rubinfeld, Ronitt, and White, Patrick. Testing random variables for independence and identity. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, FOCS '01, pp. 442–451, Washington, DC, USA, 2001. IEEE Computer Society.

Bun, Mark and Steinke, Thomas. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Proceedings of the 14th Conference on Theory of Cryptography*, TCC '16-B, pp. 635–658, Berlin, Heidelberg, 2016. Springer.

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

Canonne, Clément L. A short note on Poisson tail bounds. http://www.cs.columbia.edu/~ccanonne/files/misc/2017-poissonconcentration.pdf, 2017.

Canonne, Clément L., Diakonikolas, Ilias, Gouleakis, Themis, and Rubinfeld, Ronitt. Testing shape restrictions of discrete distributions. In *Proceedings of the 33rd Symposium on Theoretical Aspects of Computer Science*, STACS '16, pp. 25:1–25:14, 2016.

Dalenius, Tore. Towards a methodology for statistical disclosure control. *Statistisk Tidskrift*, 15:429–444, 1977.

Daskalakis, Constantinos, Dikkala, Nishanth, and Kamath, Gautam. Testing Ising models. *arXiv preprint arXiv:1612.03147*, 2016.

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

Diakonikolas, Ilias, Hardt, Moritz, and Schmidt, Ludwig. Differentially private learning of structured discrete distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pp. 2566–2574. Curran Associates, Inc., 2015.

Dinur, Irit and Nissim, Kobbi. Revealing information while preserving privacy. In *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, pp. 202–210, New York, NY, USA, 2003. ACM.

Dwork, Cynthia. Differential privacy: A survey of results. In *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation*, TAMC '08, pp. 1–19, Berlin, Heidelberg, 2008. Springer.

Dwork, Cynthia and Roth, Aaron. *The Algorithmic Foundations of Differential Privacy*. Now Publishers, Inc., 2014.

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

Fisher, Ronald A. *The Design of Experiments*. Macmillan, 1935.

Gaboardi, Marco, Lim, Hyun-Woo, Rogers, Ryan M., and Vadhan, Salil P. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *Proceedings of the 33rd International Conference on Machine Learning*, ICML '16, pp. 1395–1403. JMLR, Inc., 2016.

Homer, Nils, Szelinger, Szabolcs, Redman, Margot, Duggan, David, Tembe, Waibhav, Muehling, Jill, Pearson, John V., Stephan, Dietrich A., Nelson, Stanley F., and Craig, David W. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genetics*, 4 (8):1–9, 2008.

Johnson, Aaron and Shmatikov, Vitaly. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pp. 1079–1087, New York, NY, USA, 2013. ACM.

Kakizaki, Kazuya, Sakuma, Jun, and Fukuchi, Kazuto. Differentially private chi-squared test by unit circle mechanism. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17. JMLR, Inc., 2017.

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

Klar, Bernhard. Bounds on tail probabilities of discrete distributions. *Probability in the Engineering and Informational Sciences*, 14(02):161–171, 2000.

Levi, Reut, Ron, Dana, and Rubinfeld, Ronitt. Testing properties of collections of distributions. *Theory of Computing*, 9(8):295–347, 2013.

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

Pearson, Karl. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, 1900.

Pollard, David. A few good inequalities. http://www.stat.yale.edu/~pollard/Books/Mini/Basic.pdf, 2015.

Rao, Jon N.K. and Scott, Alastair J. The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the Americal Statistical Association*, 76 (374):221–230, 1981.

Simmons, Sean, Sahinalp, Cenk, and Berger, Bonnie. Enabling privacy-preserving gwass in heterogeneous human populations. *Cell Systems*, 3(1):54–61, 2016.

Tan, Vincent Y.F., Anandkumar, Animashree, and Willsky, Alan S. Error exponents for composite hypothesis testing of Markov forest distributions. In *Proceedings of the 2010 IEEE International Symposium on Information Theory*, ISIT '10, pp. 1613–1617, Washington, DC, USA, 2010. IEEE Computer Society.

Uhler, Caroline, Slavković, Aleksandra, and Fienberg, Stephen E. Privacy-preserving data sharing for genome-wide association studies. *The Journal of Privacy and Confidentiality*, 5(1):137–166, 2013.

Valiant, Gregory and Valiant, Paul. An automatic inequality prover and instance optimal identity testing. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pp. 51–60, Washington, DC, USA, 2014. IEEE Computer Society.

Wang, Yue, Lee, Jaewoo, and Kifer, Daniel. Differentially private hypothesis testing, revisited. *arXiv preprint arXiv:1511.03376*, 2015.

Yu, Fei, Fienberg, Stephen E., Slavković, Aleksandra B., and Uhler, Caroline. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50:133–141, 2014.