Locally Private Hypothesis Testing

Or Sheffet [1]

Abstract

We initiate the study of differentially private hypothesis testing in the local-model, under both the standard (symmetric) randomized-response mechanism (Warner, 1965; Kasiviswanathan et al., 2008) and the newer (non-symmetric) mechanisms (Bassily & Smith, 2015; Bassily et al., 2017). First, we study the general framework of mapping each user's type into a signal and show that the problem of finding the maximum-likelihood distribution over the signals is feasible. Then we discuss the randomized-response mechanism and show that, in essence, it maps the null- and alternative-hypotheses onto new sets, an affine translation of the original sets. We then give sample complexity bounds for identity and independence testing under randomized-response. We then move to the newer non-symmetric mechanisms and show that there too the problem of finding the maximum-likelihood distribution is feasible. Under the mechanism of Bassily et al (2017) we give identity and independence testers with better sample complexity than the testers in the symmetric case, and we also propose a χ^2-based identity tester which we investigate empirically.

1. Introduction

Differential privacy is a mathematically rigorous notion of privacy that has become the de-facto gold-standard of privacy preserving data analysis. Informally, ϵ-differential privacy bounds the affect of a single datapoint on any result of the computation by ϵ. In recent years the subject of private hypothesis testing has been receiving increasing attention (see Related Work below). However, by and large, the focus of private hypothesis testing is in the centralized model (or the curated model), where a single trusted entity holds the sensitive details of n users and runs the private hypothesis tester on the actual data.

[1]Dept. of Computing Science, University of Alberta.. Correspondence to: Or Sheffet <osheffet@ualberta.ca>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

In contrast, the subject of this work is private hypothesis testing in the *local*-model (or the distributed model), where a ϵ-differentially private mechanism is applied *independently* to each datum. This model, which alleviates trust (each user can run the mechanism independently on her own and release the noisy signal from the mechanism), has gained much popularity in recent years, especially since it was adopted by Google's Rappor (Erlingsson et al., 2014) and Apple (Apple, 2017). And yet, despite its popularity, and the fact that recent works (Bassily & Smith, 2015; Bassily et al., 2017) have shown the space of possible locally-private mechanism is richer than what was originally thought, little is known about private hypothesis testing in the local-model.

1.1. Background: Local Differential Privacy

We view the local differentially private model as a signaling scheme. Each datum / user has a type x taken from a predefined and publicly known set of possible types $\mathcal{X}$ whose size is $T = |\mathcal{X}|$. The differentially private mechanism is merely a randomized function $\mathcal{M} : ([n], \mathcal{X}) \to \mathcal{S}$, mapping each possible type $\mathcal{X}$ of the i-th datum to some set of possible signals $\mathcal{S}$, which we assume to be ϵ-differentially private: for any index i, any pair of types $x, x' \in \mathcal{X}$ and any signal $s \in \mathcal{S}$ it holds that $\Pr[\mathcal{M}(i,x) = s] \leq e^{\epsilon} \Pr[\mathcal{M}(i,x') = s]$.[1] In our most general results (Theorems 1 and 9), we ignore the fact that $\mathcal{M}$ is ϵ-differentially private, and just refer to any signaling scheme that transforms one domain (namely, $\mathcal{X}$) into another ($\mathcal{S}$). For example, a surveyer might unify rarely occurring types under the category of "other", or perhaps users report their types over noisy channels, etc.

We differentiate between two types of signaling schemes: the *symmetric* (or index-oblivious) variety, and the *non-symmetric* (index-aware) type. A local signaling mechanism is called *symmetric* if it is independent of the index of the datum. Namely, if for any $i \neq j$ we have that $\mathcal{M}(i,x) = \mathcal{M}(j,x) \stackrel{\text{def}}{=} \mathcal{M}(x)$. A classic exam-

[1]For simplicity, we assume $\mathcal{S}$, the set of possible signals, is discrete. Note that this doesn't exclude mechanisms such as adding Gaussian/Gamma noise to a point in $\mathbb{R}^d$ — such mechanisms require $\mathcal{X}$ to be some bounded subset of $\mathbb{R}^d$ and use the bound to set the noise appropriately. Therefore, the standard approach of discretizing $\mathcal{X}$ and projecting the noisy point to the closest point in the grid yields a finite set of signals $\mathcal{S}$.

ple of such a mechanism is *randomized-response* — that actually dates back to before differential privacy was defined (Warner, 1965) and was first put to use in differential privacy in (Kasiviswanathan et al., 2008) — where each user / datum x draws her own signal from the set $\mathcal{S} = \mathcal{X}$ skewing the probability ever-so-slightly in favor of the original type. I.e. if the user's type is x then $\mathcal{M}(x) = \begin{cases} x, & \text{w.p. } \frac{e^\epsilon}{T-1+e^\epsilon} \\ x', & \text{for any other } x' \text{ w.p. } \frac{1}{T-1+e^\epsilon} \end{cases}$.

The utility of the above-mentioned symmetric mechanism scales polynomially with T (or rather, with $|\mathcal{S}|$), which motivated the question of designing local mechanisms with error scaling logarithmically in T. This question was recently answered in the affirmative by the works of Bassily and Smith (2015) and Bassily et al (2017), whose mechanisms are *not* symmetric. In fact, both of them work by presenting each user i with a mapping $f_i : \mathcal{X} \to \mathcal{S}$ (the mapping itself is chosen randomly, but it is public, so we treat it as a given), and the user then runs the standard randomized response mechanism *on the signals* using $f_i(x)$ as the more-likely signal. (In fact, in both schemes, $\mathcal{S} = \{1, -1\}$: in (Bassily & Smith, 2015) f_i is merely the j-th coordinate of a hashing of the types where j and the hashing function are publicly known, and in (Bassily et al., 2017) f_i maps a u.a.r chosen subset of $\mathcal{X}$ to 1 and its complementary to -1.[2]) So given f_i, the user then tosses her own private random coins to determine what signal she broadcasts. Therefore, each user's mechanism can be summarized in a $|\mathcal{S}| \times |\mathcal{X}|$-matrix, where $\mathcal{M}_i(s, x)$ is the probability a user of type x sends the signal s. For example, using the mechanism of (Bassily et al., 2017), each user whose type maps to 1 sends "signal 1" with probability $\frac{e^\epsilon}{1+e^\epsilon}$ and "signal -1" with probability $\frac{1}{1+e^\epsilon}$. Namely, $\mathcal{M}_i(f_i(x), x) = \frac{e^\epsilon}{1+e^\epsilon}$ and $\mathcal{M}_i(-f_i(x), x) = \frac{1}{1+e^\epsilon}$, where f_i is the mapping $\mathcal{X} \to \{1, -1\}$ set for user i.

1.2. Our Contribution and Organization

This work initiates (to the best of our knowledge) the *theory* of differentially private hypothesis testing in the local model. First we survey related work and preliminaries. Then, in Section 3, we examine the symmetric case and show that any mechanism (not necessarily a differentially private one) yields a distribution on the signals for which finding a maximum-likelihood hypothesis is feasible, assuming the set of possible hypotheses is convex. Then, focusing on the classic randomized-response mechanism, we show that the problem of maximizing the likelihood of the observed signals is strongly-convex and thus simpler than the original problem. More importantly, in essence we give a characterization of hypothesis testing under randomized response: the symmetric locally-private mechanism translates the original null hypothesis H_0 (and the alternative H_1) by a known affine translation into a different set $\varphi(H_0)$ (and resp. $\varphi(H_1)$). Hence, hypothesis testing under randomized-response boils to discerning between two different (and considerably closer in total-variation distance) sets, but in *the exact same model* as in standard hypothesis testing as all signals were drawn from the same hypothesis in $\varphi(H_0)$. As an immediate corollary we give bounds on identity-testing (Corollary 5) and independence-testing (Theorem 6) under randomized-response. (The latter requires some manipulations and far less straightforward than the former.) The sample complexity (under certain simplifying assumptions) of both problems is proportional to $T^{2.5}$.

In Section 4 we move to the non-symmetric local-model. Again, we start with a general result showing that in this case too, finding an hypothesis that maximizes the likelihood of the observed signals is feasible when the hypothesis-set is convex. We then focus on the mechanism of Bassily et al (2017) and show that it also makes the problem of finding a maximum-likelihood hypothesis strongly-convex. We then give a simple identity tester under this scheme whose sample complexity is proportional to T^2, and is thus more efficient than *any* tester under standard randomized-response. Similarly, we also give an independence-tester with a similar sample complexity. In Section 4.2 we empirically investigate alternative identity-testing and independence-testing based on Pearson's χ^2-test in this non-symmetric scheme, and identify a couple of open problems in this regime.

1.3. Related Work

Several works have looked at the intersection of differential privacy and statistics (Dwork & Lei, 2009; Smith, 2011; Chaudhuri & Hsu, 2012; Duchi et al., 2013a; Dwork et al., 2015) mostly focusing on robust statistics; but only a handful of works study rigorously the significance and power of hypotheses testing under differential privacy. Vu and Slavkovic (2009) looked at the sample size for privately testing the bias of a coin. Johnson and Shmatikov (2013), Uhler et al (2013) and Yu et al (2014) focused on the Pearson χ^2-test (the simplest goodness of fit test), showing that the noise added by differential privacy vanishes asymptotically as the number of datapoints goes to infinity, and propose a private χ^2-based test which they study empirically. Wang et al (2015) and Gaboardi et al (2016) who have noticed the issues with both of these approaches, have revised the statistical tests themselves to incorporate also the added noise in the private computation. Cai et al (2017) give a private identity tester based on noisy χ^2-test over large bins, Sheffet (2017) studies private Ordinary Least Squares using the JL transform, and Karwa and Vadhan (2018) give

[2] In both works, much effort is put to first reducing T to the most frequent $\sqrt{n}$ types, and then run the counting algorithm. Regardless, the end-counts / collection of users' signals are the ones we care for the sake of hypothesis testing.

matching upper- and lower-bounds on the confidence intervals for the mean of a population. All of these works however deal with the centralized-model of differential privacy.

Perhaps the closest to our work are the works of Duchi et al (2013a; 2013b) who give matching upper- and lower-bound on robust estimators in the local model. And while their lower bounds do inform as to the sample complexity's dependency on ϵ^{-2}, they do not ascertain the sample complexity dependency on the size of the domain (T) we get in Section 3. Moreover, these works disregard independence testing (and in fact (Duchi et al., 2013b) focus on mean estimation so they apply randomized-response to each feature independently generating a product-distribution even when the input isn't sampled from a product-distribution). And so, to the best of our knowledge, no work has focused on hypothesis testing in the local model, let alone in the (relatively new) non-symmetric local model. Lastly, developed concurrently to our work, Gaboardi and Rogers (2018) study the asymptotic power of a variety chi-squared based hypothesis testing in the local model.

2. Preliminaries, Notation and Background

Notation. We user $lower$-case letters to denote scalars, ***bold*** characters to denote vectors and $CAPITAL$ letters to denote matrices. So 1 denotes the number, $\mathbf{1}$ denotes the all-1 vector, and $1_{\mathcal{X}\times\mathcal{X}}$ denotes the all-1 matrix over a domain $\mathcal{X}$. We use $\boldsymbol{e}_x$ to denote the standard basis vector with a single 1 in coordinate corresponding to x. To denote the x-coordinate of a vector $\boldsymbol{v}$ we use $v(x)$, and to denote the (x,x')-coordinate of a matrix M we use $M(x,x')$. For a given vector $\boldsymbol{v}$, we use $\mathrm{diag}(\boldsymbol{v})$ to denote the matrix whose diagonal entries are the coordinates of $\boldsymbol{v}$. For any natural n, we use $[n]$ to denote the set $\{1,2,...,n\}$.

Distances and norms. Unless specified otherwise $\|\boldsymbol{v}\|$ refers to the L_2-norm of $\boldsymbol{v}$, whereas $\|\boldsymbol{v}\|_1$ refers to the L_1-norm. We also denote $\|\boldsymbol{v}\|_{\frac{2}{3}} = \left(\sum_i |v_i|^{\frac{2}{3}}\right)^{\frac{3}{2}}$. For a matrix, $\|M\|_1$ denotes (as usual) the maximum absolute column sum. We identify a distribution $\boldsymbol{p}$ over a domain $\mathcal{X}$ as a T-dimensional vector with non-negative entries that sum to 1. This defines the *total variation* distance between two distributions: $d_{\mathrm{TV}}(\boldsymbol{p},\boldsymbol{q}) = \frac{1}{2}\|\boldsymbol{p}-\boldsymbol{q}\|_1$. (On occasion, we will apply d_{TV} to vectors that aren't distributions, but rather nearby estimations; in those cases we use the same definition: the half of the L_1-norm.) It is known that the TV-distance is a metric overs distributions. We also use the χ^2-divergence to measure difference between two distributions: $d_{\chi^2}(\boldsymbol{p},\boldsymbol{q}) = \sum_x \frac{(p(x)-q(x))^2}{p(x)} = \left(\sum_x \frac{(q(x))^2}{p(x)}\right) - 1$. The χ^2-divergence is not symmetric and can be infinite, however it is non-negative and zeros only when $\boldsymbol{p}=\boldsymbol{q}$. We refer the reader to (Sason & Verdú, 2016) for more properties of the total-variance distance the χ^2-divergence.

Differential Privacy. An algorithm $\mathcal{A}$ is called ϵ-differentially private, if for any two datasets D and D' that differ only on the details of a single user and any set of outputs $\mathcal{O}$, we have that $\Pr[\mathcal{A}(D)\in\mathcal{O}] \le e^{\epsilon}\Pr[\mathcal{A}(D')\in\mathcal{O}]$. The unacquainted reader is referred to the Dwork-Roth monograph (Dwork & Roth, 2014) as an introduction to the rapidly-growing field of differential privacy.

Hypothesis testing. The problem of hypothesis testing is to test whether a given set of samples was drawn from a distribution satisfying the null-hypothesis or the alternative-hypothesis. Thus, the null-hypothesis is merely a set of possible distributions H_0 and the alternative is disjoint set H_1. Hypothesis tests boils down to estimating a test-statistic θ whose distribution has been estimated under the null-hypothesis. We can thus `reject` the null-hypothesis if the value of θ is highly unlikely, or `accept` the null-hypothesis otherwise. We call an algorithm a *tester* if the acceptance (in the completeness case) or rejection (in the soundness case) happen with probability $\ge 2/3$. Standard amplification techniques (return the median of independent tests) reduce the error probability from $1/3$ to any $\beta > 0$ at the expense of increasing the sample complexity by a factor of $O(\log(1/\beta))$; hence we focus on achieving a constant error probability. One of the most prevalent and basic tests is the *identity*-testing, where the null-hypothesis is composed of a single distribution $H_0 = \{\boldsymbol{p}\}$ and our goal is to accept if the samples are drawn from $\boldsymbol{p}$ and reject if they were drawn from any other α-far (in d_{TV}) distribution. Another extremely common tester is for *independence* when $\mathcal{X}$ is composed of several features (i.e., $\mathcal{X} = \mathcal{X}^1\times\mathcal{X}^2\times...\times\mathcal{X}^d$) and the null-hypothesis is composed of all product distributions $H_0 = \{\boldsymbol{p}^1\times...\times\boldsymbol{p}^d\}$ where each $\boldsymbol{p}^j$ is a distribution on the jth feature $\mathcal{X}^j$.

Miscellaneous. We use $M \succeq 0$ to denote that M is a positive semi-definite (PSD) matrix, and $M \succeq N$ to denote that $(M-N)\succeq 0$. We use $M^{\dagger}$ to denote M's pseudo-inverse. We emphasize that we made no effort to minimize constants in our proofs, and only strived to obtain asymptotic bounds ($O(\cdot),\Omega(\cdot)$).

3. Symmetric Signaling Scheme

Recall, in the symmetric signaling scheme, each user's type is mapped through a random function $\mathcal{M}$ into a set of signals $\mathcal{S}$. This mapping is index-oblivious — each user of type $x\in\mathcal{X}$, sends the signal s with the same probability $\Pr[\mathcal{M}(x)=s]$. We denote the matrix G as the $(|\mathcal{S}|\times|\mathcal{X}|)$-matrix whose entries are $\Pr[\mathcal{M}(x)=s]$, and its sth-row by $\boldsymbol{g}_s$. Note that all entries of G are non negative and that for each x we have $\|G\boldsymbol{e}_x\|_1 = 1$. By garbling each datum *i.i.d*, we observe the new dataset $(y_1,y_2,...,y_n)\in\mathcal{S}^n$.

Theorem 1. *For any convex set H of hypotheses, the problem of finding the max-likelihood $\boldsymbol{p}\in H$ generating the observed signals $(y_1,..,y_n)$ is poly-time solvable.*

Proof. Since $G(s,x)$ describes the probability that a user of type x sends the signal s, any distribution $\boldsymbol{p} \in H$ over the types in $\mathcal{X}$ yields a distribution on $\mathcal{S}$ where $\Pr[\text{user sends } s] = \sum_{x\in\mathcal{X}} G(s,x)\cdot p(x) = \boldsymbol{g}_s^\mathsf{T}\boldsymbol{p}$. Therefore, given signals $(y_1,...,y_n)$ summarized as a signals-histogram $\langle n_s\rangle_{s\in\mathcal{S}}$, the likelihood of these signals is given by: $L(\boldsymbol{p}; y_1,...,y_n) = \prod_i \boldsymbol{g}_{y_i}^\mathsf{T}\boldsymbol{p} = \prod_{s\in\mathcal{S}}(\boldsymbol{g}_s^\mathsf{T}p)^{n_s} = \exp\left(\sum_s n_s \log(\boldsymbol{g}_s^\mathsf{T}\boldsymbol{p})\right)$. Thus, the gradient of the negative log-loss function is $\nabla f = -\frac{1}{n}\sum_{s\in\mathcal{S}} \frac{n_s}{\boldsymbol{g}_s^\mathsf{T}\boldsymbol{p}}\cdot \boldsymbol{g}_s$, and its Hessian is given by the matrix $\frac{1}{n}\sum_{s\in S}\frac{n_s}{(\boldsymbol{g}_s^\mathsf{T}\boldsymbol{p})^2}\boldsymbol{g}_s\boldsymbol{g}_s^\mathsf{T}$. Clearly, as a non-negative sum of rank-1 matrices, the Hessian is a PSD matrix.so our loss-function is convex. Known poly-time algorithms for minimizing a convex function over a convex set (e.g. (Zinkevich, 2003)) conclude the proof. □

Unfortunately, in general the solution to this problem has no closed form (to the best of our knowledge). However, we can find a close-form solution under the assumption that G isn't just any linear transformation but rather one that induces probability distribution over $\mathcal{S}$, the assumption that $|\mathcal{S}| \le |\mathcal{X}|$ (in all applications we are aware of use fewer signals than user-types) and one extra-condition.

Corollary 2. *Let $\boldsymbol{q}^*$ be the $|\mathcal{S}|$-dimensional vector given by $\langle\frac{n_s}{n}\rangle$. Given that $|\mathcal{S}| \le |\mathcal{X}|$, that G is a full-rank matrix satisfying $\|G\|_1 = 1$ and assuming that $\left(G^\dagger\boldsymbol{q}^* + \ker(G)\right)\cap H \neq \emptyset$, then any vector in H of the form $\boldsymbol{p}^* + \boldsymbol{u}$ where $\boldsymbol{p}^* = G^\dagger\boldsymbol{q}^*$ and $\boldsymbol{u} \in \ker(G)$ is an hypothesis that maximizes the likelihood of the given signals $(y_1,...,y_n)$.*

Proof deferred to the supplementary material, Section B.

3.1. Hypothesis Testing under Randomized-Response

We now aim to check the affect of a particular G, the one given by the randomized-response mechanism. In this case $\mathcal{S} = \mathcal{X}$ and we denote G as the matrix whose entries are $G(x,x') = \begin{cases}\rho+\gamma & \text{, if } x'=x\\ \rho & \text{, otherwise}\end{cases}$ where $\rho \overset{\text{def}}{=} \frac{1}{T-1+e^\epsilon}$ and $\gamma \overset{\text{def}}{=} \frac{e^\epsilon-1}{T-1+e^\epsilon}$. We get that $G = \rho\cdot 1_{\mathcal{X}\times\mathcal{X}} + \gamma I$ (where $1_{\mathcal{X}\times X}$ is the all-1 matrix). In particular, all vectors $\boldsymbol{g}_s = \boldsymbol{g}_x$, which correspond to the rows of G, are of the form: $\boldsymbol{g}_x = \rho\mathbf{1} + \gamma\boldsymbol{e}_x$. It follows that for any probability distribution $\boldsymbol{p}\in H$ we have that $\Pr[\text{seeing signal } x] = \boldsymbol{g}_x^\mathsf{T}\boldsymbol{p} = \rho + \gamma p(x)$. We have therefore translated any $\boldsymbol{p}\in H$ (over $\mathcal{X}$) to an hypothesis $\boldsymbol{q}$ over $\mathcal{S}$ (which in this case $\mathcal{S}=\mathcal{X}$), using the affine transformation $\varphi(\boldsymbol{p}) = \rho\mathbf{1} + \gamma\boldsymbol{p} = T\rho\boldsymbol{u}_\mathcal{X} + \gamma\boldsymbol{p}$ when $\boldsymbol{u}_\mathcal{X}$ denotes the uniform distribution over $\mathcal{X}$. (Indeed, $\gamma = 1 - T\rho$, an identity we will often apply.) At the risk of overburdening notation, we use φ to denote the same transformation over scalars, vectors and even sets (applying φ to each vector in the set). Since φ is injective, we have therefore discovered the following theorem.

Theorem 3. *Under the classic randomized response mechanism, testing for any hypothesis H_0 (or for comparing H_0 against the alternative H_1) of the original distribution, translates into testing for hypothesis $\varphi(H_0)$ (or $\varphi(H_0)$ against $\varphi(H_1)$) for generating the signals $y_1,...,y_n$.*

Theorem 3 seems very natural and simple, and yet (to the best of our knowledge) it was never put to words.

Moreover, it is simple to see that under standard-randomized response, our log-loss function is in fact strongly-convex, and therefore finding $\boldsymbol{p}^*$ becomes drastically more efficient (see, for example (Hazan et al., 2006)).

Claim 4. *Given signals $y_1,...,y_n$ generated using standard randomized response with parameter $\epsilon < 1$, we have that our log-loss function is $\Theta(\epsilon^2\cdot\frac{\min_x\{n_x\}}{n})$-strongly convex.*

Note that in expectation $n_x \ge \rho n$, hence with overwhelming probability we have $\min_x n_x \ge n/(2T)$. The proof is fairly straight-forward and is deferred to the supplementary material, Section B.

A variety of corollaries follow from Theorem 3. In particular, a variety of detailing matching sample complexity upper- and lower-bounds translate automatically into the realm of making such hypothesis-tests over the outcomes of the randomized-response mechanism. We focus here on two of the most prevalent tests: identity testing and independence testing.

Identity Testing. Perhaps the simplest of the all hypothesis testing is to test whether a given sample was generated according to a given distribution or not. Namely, the null hypothesis is a single hypothesis $H_0 = \{\boldsymbol{p}\}$, and the alternative is $H_1 = \{\boldsymbol{q} : d_{\text{TV}}(\boldsymbol{p},\boldsymbol{q}) \ge \alpha\}$ for a given parameter α. The seminal work of Valiant and Valiant (2014) discerns that (roughly) $\Theta(\|\boldsymbol{p}\|_{\frac{2}{3}}/\alpha^2)$ samples are sufficient and are necessary for correctly rejecting or accepting the null-hypothesis w.p.$\ge 2/3$.[3]

Here, the problem of identity testing under standard randomized response reduces to the problem of hypothesis testing between $\varphi(H_0) = \{\rho\mathbf{1} + \gamma\boldsymbol{p} : \boldsymbol{p}\in H_0\}$ and $\varphi(H_1) = \{\varphi(\boldsymbol{q}) : \boldsymbol{q} \text{ satisfying } d_{\text{TV}}(\boldsymbol{p},\boldsymbol{q}) \ge \alpha\}$.

Corollary 5. *In order to do identity testing under standard randomized response with confidence and power $\ge 2/3$, it is necessary and sufficient that we get $\Theta(\frac{T^{2.5}}{\epsilon^2\alpha^2})$ samples.*

The proof uses the results of (Valiant & Valiant, 2014) as a black-box and is mainly composed of calculations, so it is deferred to supplementary material, Section B.

Independence Testing. Another prevalent hypothesis testing over a domain $\mathcal{X}$ where each type is composed of multiple feature is independence testing. Denoting $\mathcal{X} = \mathcal{X}^1\times\mathcal{X}^2\times...\times\mathcal{X}^d$ as a domain with d possible features (hence $T = |\mathcal{X}| = \prod_j|\mathcal{X}^j| \overset{\text{def}}{=} \prod_j T^j$), our goal is to discern whether an observed sample is drawn from a product distribution or a distribution α-far from any product distri-

[3] For the sake of brevity, we ignore pathological examples where by removing α probability mass from $\boldsymbol{p}$ we obtain a vector of significantly smaller $\frac{2}{3}$-norm.

bution. In particular, the null-hypothesis in this case is a complex one: $H_0 = \{\bar{\boldsymbol{p}} = \boldsymbol{p}^1 \times \boldsymbol{p}^2 \times ... \times \boldsymbol{p}^d\}$ and the alternative is $H_1 = \{\boldsymbol{q} : \min_{\bar{\boldsymbol{p}} \in H_0} d_{\mathrm{TV}}(\boldsymbol{q}, \bar{\boldsymbol{p}}) \geq \alpha\}$. To the best of our knowledge, the (current) tester with smallest sample complexity is of Acharya et al (2015), which requires $\Omega\left((\sqrt{T} + \sum_j T^j)/\alpha^2\right)$ iid samples.

We now consider the problem of testing for independence under standard randomized response. Our goal is to prove the following theorem.

Theorem 6. *There exists an algorithm that takes $n = \tilde{\Omega}(\frac{T^2}{\alpha^2\epsilon^2}\left(d^2(\max_j\{T^j\})^2 + \sqrt{T}\right))$ signals generated by applying standard randomized response (with $\epsilon < 1$) on n samples drawn from a distribution $\boldsymbol{p}$ and with probability $\geq 2/3$ accepts if $\boldsymbol{p} \in H_0$, or rejects if $\boldsymbol{p} \in H_1$. Moreover, no algorithm can achieve such guarantee using $n = o(T^{5/2}/(\alpha^2\epsilon^2))$ signals.*

Note there are at least two types per feature, so $d \leq \log_2(T)$. Should all T^js be equal we have $(T^j)^2 \leq T^{\frac{2}{d}}$, making $T^{2.5}/(\alpha^2\epsilon^2)$ the leading term in the above bound.

Proof. Theorem 3 implies we are comparing $\varphi(H_0) = \{\rho\mathbf{1}_{\mathcal{X}} + \gamma(\boldsymbol{p}^1 \times ... \times \boldsymbol{p}^d)\}$ to $\varphi(H_1) = \{\rho\mathbf{1}_{\mathcal{X}} + \gamma\boldsymbol{q} : q \in H_1\}$. Note that $\varphi(H_0)$ is not a subset of product-distributions over $\mathcal{X}$ but rather a convex combination (with publicly known weights) of the uniform distribution and H_0; so we cannot run the independence tester of Acharya et al on the signals as a black-box. Luckily it holds that $\varphi(H_1)$ is far from all distributions in $\varphi(H_0)$: for each $\boldsymbol{q} \in H_1$ and $\bar{\boldsymbol{p}} \in H_0$ we have $d_{\mathrm{TV}}(\varphi(\boldsymbol{q}), \varphi(\bar{\boldsymbol{p}})) \geq \gamma d_{\mathrm{TV}}(\boldsymbol{q}, \bar{\boldsymbol{p}}) \geq \gamma\alpha$. And so we leverage on the main result of Acharya et al ((2015), Theorem 2): we first find a distribution $\rho\mathbf{1} + \gamma\bar{\boldsymbol{z}} \in \varphi(H_0)$ such that if the signals were generated by some $\rho\mathbf{1}_{\mathcal{X}} + \gamma\bar{\boldsymbol{p}} \in \varphi(H_0)$ then $d_{\chi^2}(\varphi(\bar{\boldsymbol{z}}), \varphi(\bar{\boldsymbol{p}})) \leq \gamma^2\alpha^2/500$, and then test if indeed the signals are likely to be generated by a distribution close to $\varphi(\bar{\boldsymbol{z}})$ using Acharya et al's algorithm. We now give our procedure for finding the product-distribution $\bar{\boldsymbol{z}}$.

Per feature j, given the jth feature of the signals $y_1^j, ..., y_n^j$ where each $x^j \in \mathcal{X}^j$ appears n_{x^j} times, our procedure for finding $\boldsymbol{z}^j$ is as follows.

0. (Preprocessing:) Denote $\tau = \alpha/(10d \cdot T^j)$. We call any type x^j where $\frac{n_{x^j}}{n} \leq \frac{1-\gamma}{T^j} + \gamma\tau$ as *small* and otherwise we say type x^j is *large*. Ignore all small types, and learn $\boldsymbol{z}^j$ only over large types. (For brevity, we refer to n as the number of signals on large types and T^j as the number of large types.)
1. Set the distribution $\tilde{\boldsymbol{z}}^j$ as the "add-1" estimator of Kamath et al (2015) for the signals: $\tilde{\boldsymbol{z}}^j(x^j) = \frac{1+n_{x^j}}{T^j+n}$.
2. Compute $\boldsymbol{z}^j = \frac{1}{\gamma}\left(I - \frac{1-\gamma}{T^j}1_{\mathcal{X}^j}\right)\tilde{\boldsymbol{z}}^j$.

Once $\boldsymbol{z}^j$ is found for each feature j, set $\bar{\boldsymbol{z}} = \boldsymbol{z}^1 \times ... \times \boldsymbol{z}^d$ run the test of Acharya et al (2015) (Theorem 2) with $\varphi(\bar{\boldsymbol{z}})$ looking only at the large types from each feature, setting the distance parameter to $\frac{\alpha\gamma}{2}$ and confidence $\frac{1}{9}$, to decide whether to accept or reject.

In order to successfully apply the Acharya et al's test, a few conditions need to hold. First, the provided distribution $\varphi(\bar{\boldsymbol{z}})$ should be close to $\varphi(H_0)$. This however hold trivially, as $\bar{\boldsymbol{z}}$ is a product-distribution. Secondly, we need that $\varphi(\bar{\boldsymbol{z}})$ and $\varphi(\bar{\boldsymbol{p}})$ to be close in χ^2-divergence, as we argue next.

Lemma 7. *Suppose that n, the number of signals, is at least $\Omega(\frac{d^2}{\alpha^2\gamma^2}\max_j\{T^j\})$. Then the above procedure creates distributions $\boldsymbol{z}^j$ such that the product distribution $\bar{\boldsymbol{z}} = \boldsymbol{z}^1 \times \boldsymbol{z}^2 \times ... \times \boldsymbol{z}^d$ satisfies the following property. If the signals $y_1, ..., y_n$ were generated by $\varphi(\bar{\boldsymbol{p}})$ for some product-distribution $\bar{\boldsymbol{p}} = \boldsymbol{p}^1 \times ... \times \boldsymbol{p}^d$, then w.p. $\geq 8/9$ we have that $d_{\chi^2}(\varphi(\bar{\boldsymbol{z}}), \varphi(\bar{\boldsymbol{p}})) \leq \gamma^2\alpha^2/1000$.*

We table the proof of Lemma 7 to Section B in the supplementary material. Next, either completeness or soundness must happen: either the signals were taken from randomized-response on a product distribution, or they were generated by a distribution $\gamma\alpha/2$-far from $\varphi(H_0)$. If no type of any feature was deemed as "small", this condition clearly holds; but we need to argue this continues to hold even when we run our tester on a strict subset of $\mathcal{X}$ composed only of large types in each feature. Completeness is straight-forward: since we remove types feature by feature, the types now come from a product distribution $\bar{\boldsymbol{p}}_{\mathrm{large}} = \boldsymbol{p}^1_{\mathrm{large}} \times ... \times \boldsymbol{p}^d_{\mathrm{large}}$ where each $\boldsymbol{p}^j_{\mathrm{large}}$ is a restriction of $\boldsymbol{p}^j$ to the large types of feature j. Soundness however is more intricate. We partition $\mathcal{X}$ into two subsets: $\mathrm{AllLarge} = \{(x^1, x^2, ..., x^d) \in \mathcal{X} : \forall j,\ x^j \text{ is large}\}$ and $\mathrm{Rest} = \mathcal{X} \setminus \mathrm{AllLarge}$; and break $\boldsymbol{q}$ into $\boldsymbol{q} = \eta\boldsymbol{q}_{\mathrm{Rest}} + (1-\eta)\boldsymbol{q}_{\mathrm{AllLarge}}$, with $\eta = \Pr_{\boldsymbol{q}}[\mathrm{Rest}]$. Claim 8 (proof deferred to the supplementary material) argues that $\eta < \frac{\alpha}{2}$. Therefore, $d_{\mathrm{TV}}(\boldsymbol{q}, \boldsymbol{q}_{\mathrm{AllLarge}}) \leq \frac{\alpha}{2}$, implying that $d_{\mathrm{TV}}(\varphi(\boldsymbol{q}_{\mathrm{AllLarge}}), \varphi(H_0)) > \alpha \cdot \gamma - \frac{\alpha\gamma}{2} = \frac{\alpha\gamma}{2}$.

Claim 8. *Assume the underlying distribution of the samples is $\boldsymbol{q}$ and that the number of signals is at least $n = \Omega(\frac{d^2(\max_j T^j)^2}{\alpha^2\gamma^2}\log(d\max_j T^j))$. Then w.p. $\geq 8/9$ our preprocessing step marks certain types each feature as "small" such that the probability (under $\boldsymbol{q}$) of sampling a type $(x^1, x^2, ..., x^d)$ such that $\exists j, x^j$ is small is $\leq \alpha/2$.*

So, given that both Lemma 7 and Claim 8 hold, we can use the test of Acharya et al, which requires a sample of size $n = \Omega(\sqrt{T}/(\alpha\gamma)^2)$. Recall that $\epsilon < 1$ so $\gamma = \Theta(\epsilon/T)$, and we get that the sample size required for the last test is $n = \Omega(\frac{T^{2.5}}{\alpha^2\epsilon^2})$. Moreover, for this last part, the lower bound in Acharya et al (2015) still holds (for the same reason it holds in the identity-testing case): the lower bound is derived from the counter example of testing whether the signals were generated from the uniform distribution (which clearly lies in $\varphi(H_0)$) or any distribution from a collection of perturbations which all belong to $\varphi(H_1)$ (See (Paninski, 2008) for more details). Each of distribution is thus $\gamma\alpha$-far from $\varphi(H_0)$ and so any tester for this particular construc-

tion requires $\sqrt{T}/(\alpha\gamma)^2$-many samples. This proves both the upper- and lower-bounds of Theorem 6. □

4. Non-Symmetric Signaling Schemes

Let us recall the non-symmetric signaling schemes in (Bassily & Smith, 2015; Bassily et al., 2017). Each user, with true type $x \in \mathcal{X}$, is assigned her own mapping (the mapping is broadcast and publicly known) $f_i : \mathcal{X} \to \mathcal{S}$. This sets her inherent signal to $f_i(x)$, and then she runs standard (symmetric) randomized response *on the signals*, making the probability of sending her true signal $f_i(x)$ to be e^ϵ-times greater than any other signal $s \neq f_i(x)$.

In fact, let us allow an even broader look. Each user is given a mapping $f_i : \mathcal{X} \to \mathcal{S}$, and denoting (like before) $T = |\mathcal{X}|$ and $S = |\mathcal{S}|$, we identify this mapping with a $(S \times T)$-matrix G_i. The column $\boldsymbol{g}_i^x = G_i \boldsymbol{e}_x$ is the probability distribution that a user of type x is going to use to pick which signal she broadcasts. (And so the guarantee of differential privacy is that for any signal $s \in \mathcal{S}$ and any two types $x \neq x'$ we have that $g_i^x(s) \leq e^\epsilon g_i^{x'}(s)$.) Therefore, all entries in G_i are non-negative and $\|G_i\|_1 = 1$ for all is.

Similarly to the symmetric case, we first exhibit the feasibility of finding a maximum-likelihood hypothesis given the signals from the non-symmetric scheme. Since we view which signal in $\mathcal{S}$ was sent, our likelihood mainly depends on the *row* vectors $\boldsymbol{g}_i^s$. We prove the following theorem, proof deferred to Section C in the supplementary material.

Theorem 9. *For any convex set H, the problem of finding the max-likelihood $\boldsymbol{p} \in H$ generating the observed non-symmetric signals $(y_1, .., y_n)$ is poly-time solvable.*

4.1. Hypothesis Testing under Non-Symmetric Locally-Private Mechanisms

Let us recap the differentially private scheme of Bassily et al (2017). It this scheme, the mechanism uses solely two signals $\mathcal{S} = \{1, -1\}$ (so $S = 2$). For every i the mechanism sets G_i by picking u.a.r for each $x \in \mathcal{X}$ which of the two signals in $\mathcal{S}$ is more likely; the chosen signal gets a probability mass of $\frac{e^\epsilon}{1+e^\epsilon}$ and the other get probability mass of $\frac{1}{1+e^\epsilon}$. We denote η as the constant such that $\frac{1}{2} + \eta = \frac{e^\epsilon}{1+e^\epsilon}$ and $\frac{1}{2} - \eta = \frac{1}{1+e^\epsilon}$; namely $\eta = \frac{e^\epsilon - 1}{2(e^\epsilon+1)} = \Theta(\epsilon)$ when $\epsilon < 1$. Thus, for every $s \in \{1, -1\}$ the row vector $\boldsymbol{g}_i^s$ is chosen such that each coordinate is chosen iid and uniformly from $\{\frac{1}{2} + \eta, \frac{1}{2} - \eta\}$. (Obviously, there's dependence between $\boldsymbol{g}_i^1$ and $\boldsymbol{g}_i^{-1}$, as $\boldsymbol{g}_i^1 + \boldsymbol{g}_i^{-1} = \mathbf{1}$, but the distribution of $\boldsymbol{g}_i^1$ is identical to the one of $\boldsymbol{g}_i^{-1}$.)

First we argue that for any distribution $\boldsymbol{p}$, if n is sufficiently large then w.h.p over the generation of the G_is and over the signals we view from each user, then finding $\hat{\boldsymbol{p}}$ which maximizes the likelihood of the observed signals yields a good approximation to $\boldsymbol{p}$. To that end, it suffices to argue that the function we optimize is Lipfshitz and strongly-convex.

Lemma 10. *Fix $\delta > 0$ and assume that the number of signals we observe is $n = \Omega(T^3 \log(1/\delta))$. Then w.p.$\geq 1 - \delta$ it holds that the function $f(\boldsymbol{p})$ we optimize (as given in Equation (1)) is $\left(3\sqrt{T}\right)$-Lipfshitz and $\left(\frac{\eta^2}{2}\right)$-strongly convex over the subspace $\{\boldsymbol{x} : \boldsymbol{x}^\mathsf{T}\mathbf{1} = 0\}$ (all vectors orthogonal to the all-1 vector).*

The proof of Lemma 10 — which (in part) is hairy due to the dependency between the matrix G_i and the signal y_i — is deferred to Section C in the supplementary material.

Identity Testing. Designing an Identity Test based solely on the maximum-likelihood is feasible, due to results like Cesa-Binachi et al (2002) which allow us to compare between the risk of the result $\tilde{\boldsymbol{p}}$ of a online gradient descent algorithm to the original distribution $\boldsymbol{p}$ which generated the signals. Through some manipulations one can (eventually) infer that $|f(\boldsymbol{p}) - f(\tilde{\boldsymbol{p}})| = O(1/\sqrt{n})$. However, since strong-convexity refers to the L_2-norm *squared* of $\|\boldsymbol{p} - \tilde{\boldsymbol{p}}\|$, we derive the resulting bound is $\|\boldsymbol{p} - \tilde{\boldsymbol{p}}\|_1^2 \leq T\|\boldsymbol{p} - \tilde{\boldsymbol{p}}\|_2^2 = O(\frac{1}{\eta^2\sqrt{n}})$, which leads to a sample complexity bound proportional to $T^3/(\alpha\eta)^4$. This bound is worse than the bounds in Section 3.

We therefore design a different, simple, identity tester in the local non-symmetric scheme, based on the estimator given in (Bassily et al., 2017). The tester itself — which takes as input a given distribution $\boldsymbol{p}$, a distance parameter $\alpha > 0$ and the n signals — is quite simple.

1. Given the n matrices $G_1, ..., G_n$ and the n observed signals $y_1, ..., y_n$, compute the estimator $\boldsymbol{\theta} = \frac{1}{n}\sum_i \frac{1}{\eta}\left(\boldsymbol{g}_i^{y_i} - \frac{1}{2}\mathbf{1}\right)$.
2. If $d_{\mathrm{TV}}(\frac{1}{2\eta}\boldsymbol{\theta}, \boldsymbol{p}) \leq \frac{\alpha}{2}$ then `accept`, else `reject`.

Theorem 11. *Assume $\epsilon < 1$. If we observe $n = \Omega(\left(\frac{T}{\alpha\epsilon}\right)^2)$ signals generated by a distribution $\boldsymbol{q}$ then w.p. $\geq 2/3$ over the matrices G_i we generate and the signals we observe, it holds that $d_{\mathrm{TV}}(\frac{1}{2\eta}\boldsymbol{\theta}, \boldsymbol{q}) \leq \alpha/2$.*

The correctness of the tester now follows from checking for the two cases where either $\boldsymbol{p} = \boldsymbol{q}$ or $d_{\mathrm{TV}}(\boldsymbol{p}, \boldsymbol{q}) > \alpha$.

Proof. In the first part of the proof we assume the types of the n users were already drawn and are now fixed. We denote x_i as the type of user i. We denote the frequency vector $\boldsymbol{f} = \langle \frac{n_x}{n} \rangle_{x \in \mathcal{X}}$, generated by counting the number of users of type x and normalizing it by n.

Given $\boldsymbol{f}$, we examine the estimator $\boldsymbol{\theta}$. For each user i we have that $\frac{1}{\eta}(\boldsymbol{g}_i^{y_i} - \frac{1}{2}\mathbf{1}) \in \{-1, 1\}^T$. Because x_i, the type of user i, is fixed, then for each coordinate $x' \neq x_i$, the signal y_i is *independent* of the x'-column in G_i (y_i depends solely on the entries in the x_i-column). We thus have that $g_i^{y_i}(x')$ is distributed uniformly among $\{\frac{1}{2} \pm \eta\}$ and so $\mathrm{E}[\frac{1}{\eta}(g_i^{y_i}(x') - \frac{1}{2})] = 0$. In contrast, $\Pr[\, \frac{1}{\eta}(g_i^{y_i}(x_i) - \frac{1}{2}) = 1\,]$ $= \sum_{s \in \{-1,1\}} \Pr[\frac{1}{\eta}(g_i^s(x_i) - \frac{1}{2}) = 1 \text{ and } y_i = s]$ $= 2 \cdot \frac{1}{2} \cdot (\frac{1}{2} + \eta) = \frac{1}{2} + \eta$. Therefore, $\mathrm{E}[\frac{1}{\eta}(g_i^{y_i}(x_i) - \frac{1}{2})] =$

$(\frac{1}{2}+\eta)-(\frac{1}{2}-\eta)=2\eta$. It follows that $\mathrm{E}[\frac{1}{\eta}(\boldsymbol{g}_i^{y_i}-\frac{1}{2}\mathbf{1})]=2\eta\boldsymbol{e}_{x_i}$ and so $\mathrm{E}[\boldsymbol{\theta}]=2\eta\boldsymbol{f}$.

Next we examine the variance of $\boldsymbol{\theta}$, and argue the following (proof deferred to supplementary material).

Proposition 12. $\mathrm{E}[(\boldsymbol{\theta}-2\eta\boldsymbol{f})(\boldsymbol{\theta}-2\eta\boldsymbol{f})^{\mathsf{T}}]\preceq\frac{1}{n}I$

So as a result, the expected L_2-difference $\mathrm{E}[\|\boldsymbol{\theta}-2\eta\boldsymbol{f}\|^2]=\mathrm{E}[\mathrm{trace}((\boldsymbol{\theta}-2\eta\boldsymbol{f})(\boldsymbol{\theta}-2\eta\boldsymbol{f})^{\mathsf{T}})]=\mathrm{trace}(\mathrm{E}[(\boldsymbol{\theta}-2\eta\boldsymbol{f})(\boldsymbol{\theta}-2\eta\boldsymbol{f})^{\mathsf{T}}])\leq\frac{T}{n}$. Chesbyshev's inequality assures us that therefore $\Pr[\frac{1}{2\eta}\|\boldsymbol{\theta}-2\eta\boldsymbol{f}\|>\frac{\sqrt{6T}}{2\eta\sqrt{n}}]\leq\frac{T/n}{6T/n}=\frac{1}{6}$.

So far we have assumed $\boldsymbol{f}$ is fixed, and only looked at the event that the coin-tosses of the mechanism yielded an estimator far from its expected value. We now turn to bounding the distance between $\boldsymbol{f}$ and its expected value $\boldsymbol{q}$ (the distribution that generated the types). Indeed, it is clear to see that the expected value of $\boldsymbol{f}=\frac{1}{n}\sum_i\boldsymbol{e}_{x_i}$ is $\mathrm{E}[\boldsymbol{f}]=\boldsymbol{q}$. Moreover, it isn't hard (and has been computed before many times, e.g. Agresti (2003)) to see that $\mathrm{E}[(\boldsymbol{f}-\boldsymbol{q})(\boldsymbol{f}-\boldsymbol{q})^{\mathsf{T}}]=\frac{1}{n}\left(\mathrm{diag}(\boldsymbol{q})-\boldsymbol{q}\boldsymbol{q}^{\mathsf{T}}\right)$. Thus $\mathrm{E}[\|\boldsymbol{f}-\boldsymbol{q}\|^2]=\mathrm{trace}(\frac{1}{n}\left(\mathrm{diag}(\boldsymbol{q})-\boldsymbol{q}\boldsymbol{q}^{\mathsf{T}}\right))=\frac{1}{n}(1-\|\boldsymbol{q}\|^2)$. Therefore, applying Chebyshev again, we get that w.p. at most $1/6$ over the choice of types by $\boldsymbol{q}$, we have that $\Pr[\|\boldsymbol{f}-\boldsymbol{q}\|>\sqrt{6/n}]\leq\frac{1/n}{6/n}=\frac{1}{6}$.

Combining both results we get that w.p. $\geq 2/3$ we have that $\|\frac{1}{2\eta}\boldsymbol{\theta}-\boldsymbol{q}\|_1\leq\sqrt{T}\|\frac{1}{2\eta}\boldsymbol{\theta}-\boldsymbol{q}\|\leq\sqrt{T}\left(\|\frac{1}{2\eta}\boldsymbol{\theta}-\boldsymbol{f}\|+\|\boldsymbol{f}-\boldsymbol{q}\|\right)\leq\sqrt{\frac{6T^2}{4\eta^2n}}+\sqrt{\frac{6T}{n}}\leq\alpha$ since we have $n=\Omega(\frac{T^2}{\eta^2\alpha^2})$. Recall that $\eta=\Theta(\epsilon)$ and that $d_{\mathrm{TV}}(\boldsymbol{x},\boldsymbol{y})=\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_1$, and the bound of $\frac{\alpha}{2}$ is proven. □

Independence Testing. Similarly to the identity tester, we propose a similar tester for independence. Recall that in this case, $\mathcal{X}$ is composed of d features, hence $\mathcal{X}=\mathcal{X}^1\times\mathcal{X}^2\times...\times\mathcal{X}^d$, with our notation of $T^j=|\mathcal{X}^j|$ for each j. Our tester should `accept` when the underlying distribution over the types is some product distribution $\boldsymbol{p}$, and should `reject` when the underlying distribution over the types is α-far from any product distribution. The tester, whose input is the n signals and a distance parameter $\alpha>0$, is as follows.

1. Given the n matrices $G_1,...,G_n$ and the n observed signals $y_1,...,y_n$, compute the estimator $\boldsymbol{\theta}=\frac{1}{n}\sum_i\frac{1}{\eta}\left(\boldsymbol{g}_i^{y_i}-\frac{1}{2}\mathbf{1}\right)$.
2. For each feature j compute $\boldsymbol{\theta}^j$ — the jth marginal of $\frac{1}{2\eta}\boldsymbol{\theta}$ (namely, for each $x^j\in\mathcal{X}^j$ sum all types whose jth feature is x^j). Denote $\bar{\boldsymbol{\theta}}=\boldsymbol{\theta}^1\times...\times\boldsymbol{\theta}^d$.
3. If $d_{\mathrm{TV}}(\frac{1}{2\eta}\boldsymbol{\theta},\bar{\boldsymbol{\theta}})\leq\frac{\alpha}{2}$ then `accept`, else `reject`.

Theorem 13. *Assume $\epsilon<1$. Given $n=\Omega(\frac{T}{\alpha^2\epsilon^2}\left(T+d^2\sum_j T^j\right))$ iid drawn signals from the non-symmetric locally-private mechanism under a dataset whose types were drawn iid from some distribution $\boldsymbol{q}$, then w.p. $\geq 2/3$ over the matrices G_i we generate and the types in the dataset we have the following guarantee. If $\boldsymbol{q}$ is a product distribution, then $d_{\mathrm{TV}}(\frac{1}{2\eta}\boldsymbol{\theta},\bar{\boldsymbol{\theta}})\leq\frac{\alpha}{2}$, and if $\boldsymbol{q}$ is α-far from any product distribution then $d_{\mathrm{TV}}(\frac{1}{2\eta}\boldsymbol{\theta},\bar{\boldsymbol{\theta}})>\frac{\alpha}{2}$.*
(Proof deferred to the supplementary material, Section C.)

Open Problems. (1) Is there a tester with a better sample complexity? The experiment in Section 4.2 leads us to conjecture that there exists a tester with sample complexity of $T^{1.5}/(\eta\alpha)^2$. There could exist better testers, of smaller sample complexity, which leads to the second question.
(2) Can one derive lower bounds for identity/independence testing in this model, where each sample has its own distribution, related to the original distribution over types? In Section D in the supplementary material we give more details as to possible venues to tackle both problems, relating them to the problem of learning a mixture-model of product distributions.

4.2. Experiment: Proposed χ^2-Based Testers

Following the derivations in the proof of Theorem 11, we can see that $\mathrm{Var}(\boldsymbol{\theta})=\frac{1}{n}\left(I-4\eta^2\mathrm{diag}(\boldsymbol{f}^2)\right)$. As ever, we assume ϵ is a small constant and as a result the variance in $2\eta\boldsymbol{f}$ (which is approximately $\frac{4\eta^2}{n}\mathrm{diag}(\boldsymbol{p})$) is significantly smaller than the variance of $\boldsymbol{\theta}$. This allows us to use the handwavey approximation $\boldsymbol{f}\approx\boldsymbol{p}$, and argue that we have the approximation $\mathrm{Var}(\boldsymbol{\theta})\approx\frac{1}{n}\left(I-4\eta^2\mathrm{diag}(\boldsymbol{p}^2)\right)\stackrel{\mathrm{def}}{=}\frac{1}{n}M$. Central Limit Theorem thus give that $\sqrt{n}M^{-1/2}(\boldsymbol{\theta}-2\eta\boldsymbol{p})\stackrel{n\to\infty}{\to}\mathcal{N}(0,I)$. Therefore, it stands to reason that the norm of the LHS is distributed like a χ^2-distribution, namely,

$$P(\boldsymbol{\theta})\stackrel{\mathrm{def}}{=}n\sum_{x\in\mathcal{X}}\frac{(\theta(x)-2\eta\cdot p(x))^2}{1-4\eta^2p(x)^2}\stackrel{n\to\infty}{\to}\chi^2_T$$

Our experiment is aimed at determining whether $P(\boldsymbol{\theta})$ can serve as a test statistic and assessing its sample complexity.

Setting and Default Values. We set a true ground distribution on T possible types, $\boldsymbol{p}$. We then pick a distribution $\boldsymbol{q}$ which is α-far from $\boldsymbol{p}$ using the counter example of Paninski (2008): we pair the types and *randomly* move $\frac{2\alpha}{T}$ probability mess between each pair of matched types. We then generate n samples according to $\boldsymbol{q}$, and apply the non-symmetric ϵ-differentially private mechanism of (Bassily et al., 2017). Finally, we aggregate the suitable vectors to obtain our estimator $\boldsymbol{\theta}$ and compute $P(\boldsymbol{\theta})$. If we decide to `accept/reject` we do so based on comparison of P to the $\frac{2}{3}$-quantile of the χ^2_T-distribution, so that in the limit we `reject` only w.p. $1/3$ under the null-hypothesis. We repeat this *entire* process t times. We have set the default values $T=10$, $\boldsymbol{p}=\boldsymbol{u}_T$ (uniform on $[T]$), $\alpha=0.2$, $n=1000$, $\epsilon=0.25$ and therefore $\eta=\frac{1}{2}\frac{e^\epsilon-1}{e^\epsilon+1}$, and $t=10000$.

Experiment 1: Convergence to the χ^2-distribution in the null case. First we ask ourself whether our approximation, denoting $P(\boldsymbol{\theta})\approx\chi^2_T$ is correct when indeed $\boldsymbol{p}$ is

the distribution generating the signals. To that end, we set $\alpha = 0$ (so the types are distributed according to $\boldsymbol{p}$) and plot the t empirical values of P we in our experiment, varying both the sample size $n \in \{10, 100, 1000, 10000\}$ and the domain size $T \in \{10, 25, 50, 100\}$.
The *results* are consistent — P is distributed like a χ^2_T-distribution. Indeed, the mean of the t sample points is $\approx T$ (the mean of a χ^2_T-distribution). The results themselves appear in Figure 2 in the supplementary material, Section D.

Experiment 2: Divergence from the χ^2-distribution in the alternate case. Secondly, we asked whether P can serve as a good way to differentiate between the null hypothesis (the distribution over the types is derived from $\boldsymbol{p}$) and the alternative hypothesis (the distribution over the types if $\geq \alpha$-far from $\boldsymbol{p}$). We therefore ran our experiment while varying α (between 0.25 and 0.05) and increasing n. Again, the *results* show that the distribution does shift towards higher values as n increases. The results are given in Figure 3 in the supplementary material, Section D.

Experiment 3: Sample Complexity. Next, we set to find the required sample complexity for rejection. We fix the α-far distribution from $\boldsymbol{p}$, and first do binary search to hone on an interval $[n_L, n_U]$ where the empirical rejection probability is between $30\% - 35\%$; then we equipartition this interval and return the n for which the empirical rejection probability is the closest to 33%. We repeat this experiment multiple times, each time varying just one of the 3 most important parameters, T, α and ϵ. We maintain two parameters at default values, and vary just one parameter: $T \in \{5, 10, 15, .., 100\}, \alpha \in \{0.05, 0.1, 0.15, ..., 0.5\}$, $\epsilon \in \{0.05, 0.1, 0.15, ..., 0.5\}$.
The *results* are shown in Figure 1, where next to each curve we plot the curve of our conjecture in a dotted line.[4] We conjecture initially that $n \propto T^{c_T} \cdot \alpha^{c_\alpha} \cdot \epsilon^{c_\epsilon}$. And so, for any parameter $\xi \in \{T, \alpha, \epsilon\}$, if we compare two experiments i, j that differ only on the value of this parameter and resulted in two empirical estimations N_i, N_j of the sample complexity, then we get that $c_\xi \approx \frac{\log(N_i/N_j)}{\log(\xi_i/\xi_j)}$. And so for any $\xi \in \{T, \alpha, \epsilon\}$ we take the median over of all pairs of i and j and we get the empirical estimations of $c_\epsilon = -1.900793, c_\alpha = -1.930947$ and $c_T = 1.486957$. This leads us to the conjecture that the actual sample complexity according to this test is $\frac{T^{1.5}}{\alpha^2\epsilon^2}$.

Open Problem. Perhaps even more interesting, is the experiment we wish we could have run: a χ^2-based independence testing. Assuming the distribution of the type is a product distribution $\bar{\boldsymbol{p}} = \boldsymbol{p}^1 \times ... \times \boldsymbol{p}^d$, the proof of Theorem 13 shows that for each feature j we have $\mathrm{Var}(\boldsymbol{\theta}^j - \boldsymbol{p}^j) \approx \frac{1}{4\eta^2 n}\frac{T}{T^j} I_{\mathcal{X}^j}$. Thus $4\eta^2 n \frac{T^j}{T}\|\boldsymbol{\theta}^j - \boldsymbol{p}^j\|^2 \overset{n\to\infty}{\to} \chi^2_{T^j}$. However, the d estimators $\boldsymbol{\theta}^j$ are not independent, so it is *not true* that $\sum_j 4\eta^2 n \frac{T^j}{T}\|\boldsymbol{\theta}^j - \boldsymbol{p}^j\|^2 \overset{n\to\infty}{\to} \chi^2_{\sum_j T^j}$. Moreover, even if the estimators of the marginals were independent,[5] we are still unable to determine the asymptotic distribution of $\|\bar{\boldsymbol{\theta}} - \bar{\boldsymbol{p}}\|^2$ (only a bound, scaled by $O(\max_j T_j)$, using Proposition 17 in the supplementary material), let alone the asymptotic distribution of $\|\frac{1}{2\eta}\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|^2$.

Nonetheless, we did empirically measure the quantity $Q(\boldsymbol{\theta}) \overset{\text{def}}{=} n\sum_x \frac{(\frac{1}{2\eta}\theta(x) - \bar{\theta}(x))^2}{\bar{\theta}(x)}$ under the null ($\alpha = 0$) and the alternative ($\alpha = 0.25$) hypothesis with $n = 25,000$ samples in each experiment. The results (given in Figure 4 in the supplementary material) show that the distribution of Q — albeit not resembling a χ^2-distribution — is different under the null- and the alternative-hypothesis, so we suspect that there's merit to using this quantity as a tester. We thus leave the design of a χ^2-based statistics for independence in this model as an open problem.

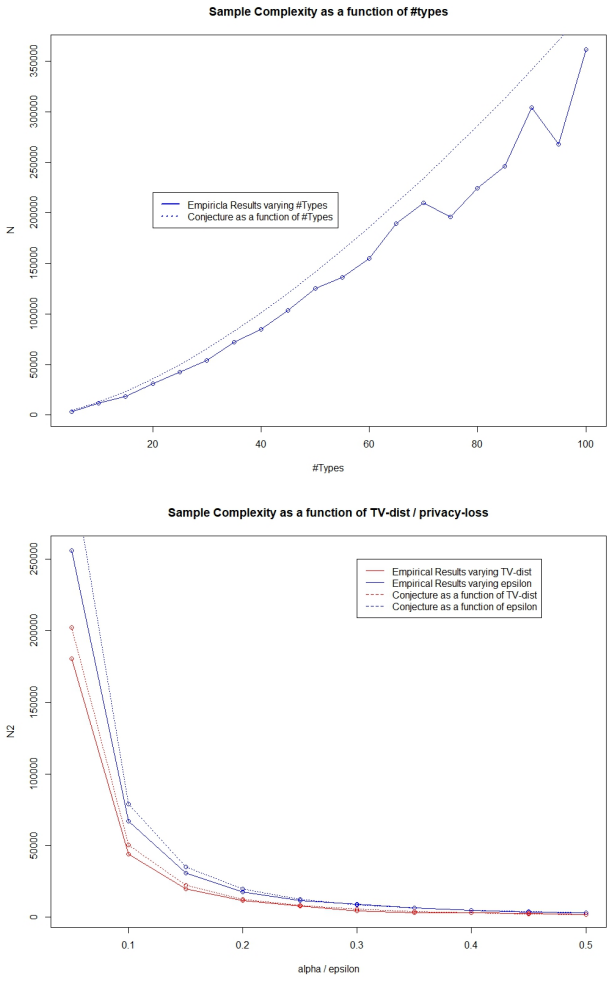

Figure 1: Empirical sample-complexity to have the tester reject w.p.$\sim 2/3$ under the alternative hypothesis.

[4] We plot the dependency on α and ϵ on the same plot, as both took the same empirical values.

[5] E.g. by assigning each example i to one of the d estimators, costing only $d = \log(T)$ factor in sample complexity

Acknowledgments

This work was supported by the Natural Sciences and Engineering Council of Canada, Grant #2017-06701. The author is also an unpaid collaborator on NSF grant 1565387. The authors thanks the anonymous reviewers for many helpful suggestions and ideas, as well as Marco Gaboardi and Ryan Rogers for helpful discussions illustrating the similarities and differences between our two papers.

References

Acharya, J., Daskalakis, C., and Kamath, G. Optimal testing for properties of distributions. In *NIPS*, pp. 3591–3599. 2015.

Agresti, A. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. 2003.

Apple, D. P. T. Learning with privacy at scale. *Apple Machine Learning Journal*, 1(8), 2017. available on `http://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html`.

Bassily, R. and Smith, A. D. Local, private, efficient protocols for succinct histograms. In *STOC*, pp. 127–135, 2015.

Bassily, R., Nissim, K., Stemmer, U., and Thakurta, A. G. Practical locally private heavy hitters. In *NIPS*, pp. 2285–2293, 2017.

Cai, B., Daskalakis, C., and Kamath, G. Priv'it: Private and sample efficient identity testing. In *ICML*, pp. 635–644, 2017.

Cesa-bianchi, N., Conconi, A., and Gentile, C. On the generalization ability of on-line learning algorithms. In *NIPS*, pp. 359–366. 2002.

Chaudhuri, K. and Hsu, D. Convergence rates for differentially private statistical estimation. In *ICML*, 2012.

Duchi, J., Jordan, M., and Wainwright, M. Local privacy and statistical minimax rates. In *FOCS*, 2013a.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Local privacy and minimax bounds: Sharp rates for probability estimation. In *NIPS*, pp. 1529–1537, 2013b.

Dwork, C. and Lei, J. Differential privacy and robust statistics. In *STOC*, 2009.

Dwork, C. and Roth, A. *The Algorithmic Foundations of Differential Privacy*. Foundations and Trends in Theoretical Computer Science, NOW Publishers, 2014.

Dwork, C., Su, W., and Zhang, L. Private false discovery rate control. *CoRR*, abs/1511.03803, 2015.

Erlingsson, Ú., Pihur, V., and Korolova, A. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *CCS*, 2014.

Gaboardi, M. and Rogers, R. M. Local private hypothesis testing: Chi-square tests. In *ICML (to appear)*, 2018.

Gaboardi, M., Lim, H. W., Rogers, R., and Vadhan, S. P. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *ICML*, pp. 2111–2120, 2016.

Hazan, E., Kalai, A., Kale, S., and Agarwal, A. Logarithmic regret algorithms for online convex optimization. In *COLT*, pp. 499–513, 2006.

Johnson, A. and Shmatikov, V. Privacy-preserving data exploration in genome-wide association studies. In *KDD*, pp. 1079–1087, 2013.

Kamath, S., Orlitsky, A., Pichapati, D., and Suresh, A. T. On learning distributions from their samples. In *COLT*, pp. 1066–1100, 2015.

Karwa, V. and Vadhan, S. Finite sample differentially private confidence intervals, 2018.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? In *FOCS*, 2008.

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

Reiss, R. *Approximate distributions of order statistics: with applications to nonparametric statistics*. Springer series in statistics. 1989.

Sason, I. and Verdú, S. f-divergence inequalities. *IEEE Trans. Information Theory*, 62(11):5973–6006, 2016.

Sheffet, O. Differentially private ordinary least squares. In *ICML*, 2017.

Smith, A. Privacy-preserving statistical estimation with optimal convergence rates. In *STOC*, 2011.

Uhler, C., Slavkovic, A. B., and Fienberg, S. E. Privacy-preserving data sharing for genome-wide association studies. *Journal of Privacy and Confidentiality*, 5(1), 2013.

Valiant, G. and Valiant, P. An automatic inequality prover and instance optimal identity testing. In *FOCS*, pp. 51–60, 2014.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. 2010. URL `http://arxiv.org/abs/1011.3027`.

Vu, D. and Slavkovic, A. Differential privacy for clinical trial data: Preliminary evaluations. In *ICDM*, pp. 138–143, 2009.

Wang, Y., Lee, J., and Kifer, D. Differentially private hypothesis testing, revisited. *CoRR*, abs/1511.03376, 2015.

Warner, S. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, 60(309), March 1965.

Yu, F., Fienberg, S., Slavkovic, A., and Uhler, C. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50:133–141, 2014.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pp. 928–936, 2003.