Journal of Privacy and Confidentiality
Vol. 10 (2) 2020 TPDP 2018
Submitted March 2019
Published June 2020

INSPECTRE: PRIVATELY ESTIMATING THE UNSEEN

JAYADEV ACHARYA, GAUTAM KAMATH, ZITENG SUN, AND HUANYU ZHANG

Cornell University
e-mail address: acharya@cornell.edu

Simons Institute for the Theory of Computing and University of Waterloo
e-mail address: g@csail.mit.edu

Cornell University
e-mail address: zs335@cornell.edu

Cornell University
e-mail address: hz388@cornell.edu

ABSTRACT. We develop differentially private methods for estimating various distributional properties. Given a sample from a discrete distribution p, some functional f and accuracy and privacy parameters α and ε, the goal is to estimate $f(p)$ up to accuracy α, while maintaining ε-differential privacy of the sample. We prove almost-tight bounds on the sample size required for this problem for several functionals of interest, including support size, support coverage, and entropy. We show that the cost of privacy is negligible in a variety of settings, both theoretically and experimentally. Our methods are based on a sensitivity analysis of several state-of-the-art methods for estimating these properties with sublinear sample complexities.

Key words and phrases: differential privacy, statistics, property estimation.

An extended abstract of this paper previously appeared in ICML 2018 (Acharya et al., 2018a). Authors are listed in alphabetical order.

Acharya was supported by NSF CCF-1657471 and a Cornell University startup grant.

Work done while Kamath was supported as a Microsoft Research Fellow, as part of the Simons-Berkeley Research Fellowship program, and by ONR N00014-12-1-0999, NSF CCF-1617730, CCF-1650733, and CCF-1741137. Work partially done while Kamath was an intern at Microsoft Research, New England, and while employed at the Simons Institute for the Theory of Computing. Kamath is currently at the University of Waterloo.

Sun was supported by NSF CCF-1657471 and a Cornell University startup grant.

Zhang was supported by NSF CCF-1657471 and a Cornell University startup grant.

JPrivacy Confidentiality
www.journalprivacyconfidentiality.org
DOI:10.29012/jpc.724

Introduction

How can we infer properties of a distribution given a sample from it? If data is in abundance, the solution may be simple – the empirical distribution will approximate the true distribution. However, challenges arise when data is scarce in comparison to the size of the domain, and especially when we wish to quantify "rare events." This is frequently the case: for example, it has recently been observed that there are several very rare genetic mutations which occur in humans, and we wish to know how many such mutations exist (Keinan and Clark, 2012; Tennessen et al., 2012; Nelson et al., 2012). Many of these mutations have only been seen once, and we can infer that there are many which have not been seen at all – therefore, the empirical distribution will completely omit such unobserved events. Over the last decade, a large body of work has focused on developing theoretically sound and effective tools for such settings (see, e.g., (Orlitsky et al., 2016) and references therein), including the problem of estimating the frequency distribution of rare genetic variations (Zou et al., 2016).

However, in many settings where one wishes to perform statistical inference, data may contain sensitive information about individuals. For example, in medical studies, where the data may contain individuals' health records which could be used by insurance companies to raise premiums. Alternatively, one can consider a map application which suggests routes based on aggregate positions of individuals, which contains delicate information including users' residence data. In these settings, it is critical that our methods protect sensitive information contained in the dataset. This does not preclude our overall goals of statistical analysis, as we are trying to infer properties of the population p, and not the samples which are drawn from said population.

That said, without careful experimental design, published statistical findings may be prone to leaking sensitive information about the sample. As a notable example, it was recently shown that one can determine the identity of some individuals who participated in genome-wide association studies (Homer et al., 2008). This realization has motivated a surge of interest in developing data sharing techniques with an explicit focus on maintaining privacy of the data (Vu and Slavković, 2009; Johnson and Shmatikov, 2013; Uhler et al., 2013; Yu et al., 2014; Simmons et al., 2016).

Privacy-preserving computation has enjoyed significant study in a number of fields, including statistics and almost every branch of computer science, including cryptography, machine learning, algorithms, and database theory – see, e.g., Dalenius (1977); Adam and Worthmann (1989); Agrawal and Aggarwal (2001); Dinur and Nissim (2003); Dwork (2008); Dwork and Roth (2014) and references therein. Perhaps the most celebrated notion of privacy, proposed by theoretical computer scientists, is *differential privacy* (Dwork, McSherry, Nissim and Smith, 2006). Informally, an algorithm is differentially private if its outputs on similar datasets are statistically close (for a more precise definition, see Section 1). Differential privacy has become the standard for theoretically-sound data privacy, leading to its adoption by several large technology companies, including Google, Apple and Microsoft (Erlingsson et al., 2014; Differential Privacy Team, Apple, 2017; Ding et al., 2017).

Our focus in this paper is to develop tools for private distribution property estimation. In particular, we study the tradeoff between statistical accuracy, privacy, and error rate in the sample size. Our model is that we are given sample access to some unknown discrete distribution p, over a domain of size k, which is possibly unknown in some tasks. We wish to estimate the following properties:

- **Support Coverage**: If we take m samples from the distribution, what is the number of unique elements we expect to see?
- **Support Size**: How many elements of the support have non-zero probability?
- **Entropy**: What is the Shannon entropy of the distribution?

For more formal statements of these problems, see Section 1.1. We require that our output is accurate, differentially private, and correct with high probability. The goal is to give an algorithm with minimal sample complexity n, while simultaneously being computationally efficient.

Theoretical Results. Our main results show that privacy can be achieved for all these problems at a very low cost. For example, if one wishes to privately estimate entropy, this incurs an additional additive cost in the sample complexity which is very close to linear in $1/\alpha\varepsilon$. We draw attention to two features of this bound. First, this is independent of k. All the problems we consider have sample complexity $\Theta(k/\log k)$, so in the primary regime of study where $k \gg 1/\alpha\varepsilon$, this small additive cost is dwarfed by the inherent sample complexity of the non-private problem. Second, note that performing even the most basic statistical task privately, estimating the bias of a coin, incurs this linear dependence. Surprisingly, we show that much more sophisticated inference tasks can be privatized at almost no cost. In particular, these properties imply that the additive cost of privacy is $o(1)$ in the most studied regime where the support size is large. In general, this is not true – for many other problems, including distribution estimation and hypothesis testing, the additional cost of privacy depends significantly on the support size or dimension (Diakonikolas et al., 2015; Cai et al., 2017; Acharya et al., 2018b; Aliakbarpour et al., 2018; Acharya et al., 2019c). We also provide lower bounds, showing that our upper bounds are almost tight. A more formal statement of our results appears in Section 2.

Experimental Results. We demonstrate the efficacy of our method with experimental evaluations. As a baseline, we compare with the non-private sample-optimal algorithms of Orlitsky et al. (2016) and Wu and Yang (2019). Overall, we find that our algorithms' performance is nearly identical, showing that, in many cases, privacy comes (essentially) for free. We begin with an evaluation on synthetic data. Then, inspired by Valiant and Valiant (2013); Orlitsky et al. (2016), we analyze a text corpus consisting of words from Hamlet, in order to estimate the number of unique words which occur. Finally, we investigate name frequencies in the US census data. This setting has been previously considered by Orlitsky et al. (2016), but we emphasize that this is an application where private statistical analysis is critical. This is evidenced by efforts of the US Census Bureau to incorporate differential privacy into the 2020 US census (Dajani et al., 2017).

Techniques. Our approach works by choosing statistics for these tasks which possess bounded sensitivity, which is well-known to imply privacy under the Laplace or Gaussian mechanism. We note that bounded sensitivity of statistics is not always something that can be taken for granted. Indeed, for many fundamental tasks, optimal algorithms for the non-private setting may be highly sensitive, thus necessitating crucial modifications to obtain differential privacy (Acharya et al., 2015; Cai et al., 2017). Thus, careful choice and design of statistics must be a priority when performing inference with privacy considerations.

To this end, we leverage recent results of Acharya et al. (2017a), which studies estimators for non-private versions of the problems we consider. The main technical work in their paper exploits bounded sensitivity to show sharp cutoff-style concentration bounds for certain estimators, which operate using the principle of best-polynomial approximation. They use these results to show that a single algorithm, the Profile Maximum Likelihood (PML), can

estimate all these properties simultaneously. On the other hand, we consider the sensitivity of these estimators for purposes of privacy – the same property is utilized by both works for very different purposes, a connection which may be of independent interest. Specifically, we hope that by drawing this connection, researchers will revisit statistics which use similar analysis and investigate their amenability to privacy.

We note that bounded sensitivity of a statistic may be exploited for purposes other than privacy. For instance, by McDiarmid's inequality, any such statistic also enjoys very sharp concentration of measure, implying that one can boost the success probability of the test at an additive cost which is logarithmic in the inverse of the failure probability. One may naturally conjecture that, if a statistical task is based on a primitive which concentrates in this sense, then it may also be privatized at a low cost. However, this is not true – estimating a discrete distribution in ℓ_1 distance is such a task, but the cost of privatization depends significantly on the support size (Acharya et al., 2019b).

One can observe that, algorithmically, our method is quite simple: compute the non-private statistic, and add a relatively small amount of Laplace noise. The non-private statistics have recently been demonstrated to be practical (Orlitsky et al., 2016; Wu and Yang, 2019), and the additional cost of the Laplace mechanism is minimal. This is in contrast to several differentially private algorithms which invoke significant overhead in the quest for privacy. Our algorithms attain almost-optimal rates (which are optimal up to constant factors for most parameter regimes of interest), while simultaneously operating effectively in practice, as demonstrated in our experimental results.

Related Work. Over the last decade, there have been a flurry of works on the problems we study in this paper by the computer science and information theory communities, including Shannon and Rényi entropy estimation (Paninski, 2003; Valiant and Valiant, 2017; Jiao et al., 2017; Acharya et al., 2017b; Obremski and Skorski, 2017; Wu and Yang, 2019), support coverage and support size estimation (Valiant and Valiant, 2017, 2016; Raghunathan et al., 2017; Orlitsky et al., 2016; Wu and Yang, 2019). A recent paper studies the general problem of estimating functionals of discrete distribution from samples in terms of the smoothness of the functional (Fukuchi and Sakuma, 2017). These have culminated in a nearly-complete understanding of the sample complexity of these properties, with optimal sample complexities (up to constant factors) for most parameter regimes.

Recently, there has been significant interest in performing statistical tasks under differential privacy constraints. Perhaps most relevant to this work are Cai et al. (2017); Acharya et al. (2018b); Aliakbarpour et al. (2018), which study the sample complexity of differentially privately performing classical distribution testing problems, including identity and closeness testing. Some recent work focuses on the testing of simple hypotheses: Canonne et al. (2019) studies the sample complexity of this problem, while Awan and Slavkovic (2018) provides a uniformly most powerful (UMP) test for binomial data (though Brenner and Nissim (2014) shows that UMP tests can not exist in general). Other works investigating private hypothesis testing include Wang et al. (2015a); Gaboardi et al. (2016); Kifer and Rogers (2017); Kakizaki et al. (2017); Rogers (2017); Campbell et al. (2018); Swanberg et al. (2019); Couch et al. (2019), which focus less on characterizing the finite-sample guarantees of such tests, and more on understanding their asymptotic properties and applications to computing p-values. There has also been study on private distribution learning (Diakonikolas et al., 2015; Duchi et al., 2018; Karwa and Vadhan, 2018; Kamath et al., 2018), in which we wish to estimate parameters of the distribution, rather than just a particular property of interest. Private hypothesis testing and learning have also been considered in the *local*

model of differential privacy Sheffet (2018); Gaboardi and Rogers (2018); Acharya et al. (2019a,c). Similar to our work, Smith (2011) shows that the cost of privacy in statistical estimation can be a lower order term – roughly, he shows that this is the case for any statistic which is asymptotically normal. A number of other problems have been studied with privacy requirements, including clustering (Wang et al., 2015b; Balcan et al., 2017), principal component analysis (Chaudhuri et al., 2013; Kapralov and Talwar, 2013; Hardt and Price, 2014), ordinary least squares (Sheffet, 2017), and much more.

1. Preliminaries

Let $\Delta \stackrel{\text{def}}{=} \left\{(p(1), \ldots, p(k)) : p(i) \geq 0, \sum_{i=1}^{k} p(i) = 1, 1 \leq k \leq \infty\right\}$ be the set of discrete distributions over a countable support. Let Δ_k be the set of distributions in Δ with at most k non-zero probability values. A *property* $f(p)$ is a mapping from Δ to $\mathbb{R}$. We now describe the classical distribution property estimation problem, and then state the problem under differential privacy.

The Hamming distance between two sequences X^n and Y^n is defined as $d_{ham}(X^n, Y^n) = \sum_{i=1}^{n} \mathbb{I}(X_i \neq Y_i)$, the number of positions X^n and Y^n differ. An estimator $\hat{f}$ is ε-differentially private (DP) (Dwork et al., 2006) if for any X^n and Y^n, with $d_{ham}(X^n, Y^n) \leq 1$, $\frac{\Pr\left(\hat{f}(X^n) \in S\right)}{\Pr\left(\hat{f}(Y^n) \in S\right)} \leq e^{\varepsilon}$, for all measurable S.

Property Estimation. Given α, β, f, and independent samples X^n from an unknown distribution p, design an estimator[1] $\hat{f} : X^n \to \mathbb{R}$ such that with probability at least $1 - \beta$, $\left|\hat{f}(X^n) - f(p)\right| < \alpha$. The *sample complexity* of $\hat{f}$,

$$C_{\hat{f}}(f, \alpha, \beta) \stackrel{\text{def}}{=} \min\left\{n : \Pr\left(\left|\hat{f}(X^n) - f(p)\right| > \alpha\right) < \beta\right\}$$

is the smallest number of samples required to estimate f to accuracy α, with probability of error β. We study the problem for $\beta = 1/3$, and by the median trick, we can boost the success probability to $1 - \beta$ with an additional multiplicative $\log(1/\beta)$ more samples.[2] Therefore, focusing on $\beta = 1/3$, we define $C_{\hat{f}}(f, \alpha) \stackrel{\text{def}}{=} C_{\hat{f}}(f, \alpha, 1/3)$. The sample complexity of estimating a property $f(p)$ is the minimum sample complexity over all estimators: $C(f, \alpha) = \min_{\hat{f}} C_{\hat{f}}(f, \alpha)$.

Private Property Estimation. Given $\alpha, \varepsilon, \beta, f$, and independent samples X^n from an unknown distribution p, design an ε-differentially private estimator $\hat{f} : X^n \to \mathbb{R}$ such that with probability at least $1 - \beta$, $\left|\hat{f}(X^n) - f(p)\right| < \alpha$. Similar to the non-private setting, the *sample complexity* of ε-differentially private estimation problem is $C(f, \alpha, \varepsilon) = \min_{\hat{f}:\hat{f} \text{ is } \varepsilon\text{-DP}} C_{\hat{f}}(f, \alpha, 1/3)$, the smallest number of samples n for which there exists such an ε-DP $\pm\alpha$ estimator with error probability at most $1/3$.

[1] Technically, we design a family of estimators, one for each $n > 0$. For readability, we slightly abuse notation and refer to this entire family of estimators as $\hat{f}$.

[2] The median trick involves repeating the estimator $O(\log(1/\beta))$ times on disjoint sets of samples, then taking the median of the estimates. If the original estimator has success probability $2/3$, the boosted estimator will have success probability $1 - \beta$. This also preserves privacy by post processing since each estimate is differentially private.

In their original paper, (Dwork et al., 2006) provide a scheme for differential privacy, known as the Laplace mechanism. This method adds Laplace noise to a non-private scheme in order to make it private. We first define the sensitivity of an estimator, and then state their result in our setting.

Definition 1.1. The *sensitivity* of an estimator $\hat{f} : [k]^n \to \mathbb{R}$ is

$$\Delta_{n,\hat{f}} \stackrel{\text{def}}{=} \max_{d_{ham}(X^n,Y^n)\le 1} \left|\hat{f}(X^n) - \hat{f}(Y^n)\right|.$$

Lemma 1.2. *Let* $D_{\hat{f}}(\alpha,\varepsilon) = \min\{n : \Delta_{n,\hat{f}} \le \alpha\varepsilon\}$.

$$C(f,\alpha,\varepsilon) = O\left(\min_{\hat{f}} \left\{C_{\hat{f}}(f,\alpha/2) + D_{\hat{f}}\left(\alpha/4,\varepsilon\right)\right\}\right).^{3}$$

Proof. Dwork et al. (2006) showed that for a function with sensitivity $\Delta_{n,\hat{f}}$, adding Laplace noise $X \sim Lap(\Delta_{n,\hat{f}}/\varepsilon)$ makes the output an ε-differentially private. By the definition of $D_{\hat{f}}(\frac{\alpha}{4},\varepsilon)$, the Laplace noise we add has parameter at most $\frac{\alpha}{4}$. Recall that the probability density function of $Lap(b)$ is $\frac{1}{2b}e^{-\frac{|x|}{b}}$, hence we have $\Pr\left(|X| > \alpha/2\right) < \frac{1}{e^2}$. By the union bound, we get an additive error larger than $\alpha = \frac{\alpha}{2} + \frac{\alpha}{2}$ with probability at most $1/3 + \frac{1}{e^2} < 0.5$. Hence, with the aforementioned median trick, we can boost the error probability to $1/3$, at the cost of a constant factor in the number of samples. □

To prove sample complexity lower bounds for differentially private estimators, we observe any estimator which α-approximates a property can be used to perform hypothesis testing between two distributions which are α-separated in said property: in short, estimation is a harder problem than testing. To derive lower bounds on the sample complexity of private hypothesis testing, Acharya et al. (2018b) gives the following argument based on coupling:

Lemma 1.3. *Suppose there is a coupling between distributions* $X_1^n \sim p$ *and* $Y_1^n \sim q$ *over* $\mathcal{X}^n$*, such that* $\mathbb{E}\left[d_{ham}(X^n,Y^n)\right] \le D$*. Then, any* ε*-differentially private algorithm that distinguishes between* p *and* q *with error probability at most* $1/3$ *must satisfy* $D = \Omega\left(\frac{1}{\varepsilon}\right)$.

1.1. Problems of Interest.

Support Coverage. For a distribution p, and an integer m, let $S_m(p) = \sum_x (1-(1-p(x))^m)$, be the expected number of symbols that appear when we obtain m independent samples from the distribution p. The objective is to find the least number of samples n in order to estimate $S_m(p)$ to an additive $\pm\alpha m$.

Support coverage arises in many ecological and biological studies (Colwell et al., 2012) to quantify the number of *new* elements (gene mutations, species, words, etc) that can be expected to be seen in the future. Good and Toulmin (1956) proposed an estimator that, for any constant α, requires $m/2$ samples to estimate $S_m(p)$.

[3] Asymptotics in this paper are with respect to α and ε tending to 0, and the parameters n, k, m, when appearing in an expression, tending to infinity.

Support Size. The support size of a distribution p is $S(p) = |\{x : p(x) > 0\}|$, the number of symbols with non-zero probability values. However, notice that estimating $S(p)$ from samples can be hard due to the presence of symbols with negligible, yet non-zero probabilities. To circumvent this issue, Raskhodnikova et al. (2009) proposed to study the problem when the smallest probability is bounded. Let $\Delta_{\geq \frac{1}{k}} \stackrel{\text{def}}{=} \{p \in \Delta : \forall x \in [k], p(x) \in \{0\} \cup [1/k, 1]\}$ be the set of all distributions where all non-zero probabilities have value at least $1/k$. For $p \in \Delta_{\geq \frac{1}{k}}$, our goal is to estimate $S(p)$ up to $\pm\alpha k$ with the least number of samples from p.

Entropy. The Shannon entropy of a distribution p is $H(p) = \sum_x p(x) \log \frac{1}{p(x)}$, $H(p)$ is a central object in information theory (Cover and Thomas, 2006), and also arises in many fields such as machine learning (Nowozin, 2012), neuroscience (Berry et al., 1997; Nemenman et al., 2004), and others. Estimating $H(p)$ is hard with any finite number of samples due to the possibility of infinite support. To circumvent this, a natural approach is to consider distributions in Δ_k. The goal is to estimate the entropy of a distribution to an additive $\pm\alpha$.

2. Statement of Results

Our theoretical results for estimating support coverage, support size, and entropy are given below. All results are with $0 < \alpha < 1$ and $0 < \varepsilon < 1$. Algorithms for these problems and proofs of these statements are provided in Section 3. Our experimental results are described and discussed in Section 4.

Theorem 2.1. *The sample complexity of support coverage estimation is*

$$C(S_m, \alpha, \varepsilon) = \begin{cases} O\Big(\frac{m \log(1/\alpha)}{\log m} + \frac{m \log(1/\alpha)}{\log(2+\varepsilon m)}\Big), & \text{when } m \geq \frac{1}{\alpha\varepsilon}, \\ O\big(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon}\big), & \text{when } m \leq \frac{1}{\alpha\varepsilon}. \end{cases}$$

Furthermore,

$$C(S_m, \alpha, \varepsilon) = \Omega\left(\frac{m \log(1/\alpha)}{\log m} + \frac{1}{\alpha\varepsilon}\right).$$

Theorem 2.2. *The sample complexity of support size estimation is*

$$C(S, \alpha, \varepsilon) = \begin{cases} O\Big(\frac{k \log^2(1/\alpha)}{\log k} + \frac{k \log^2(1/\alpha)}{\log(2+\varepsilon k)}\Big), & \text{when } k \geq \frac{1}{\alpha\varepsilon}, \\ O\big(k \log(1/\alpha) + \frac{1}{\alpha\varepsilon}\big), & \text{when } \frac{1}{\alpha} \leq k \leq \frac{1}{\alpha\varepsilon}, \\ O\big(k \log k + \frac{k}{\varepsilon}\big), & \text{when } k \leq \frac{1}{\alpha}. \end{cases}$$

Furthermore,

$$C(S, \alpha, \varepsilon) = \begin{cases} \Omega\Big(\frac{k \log^2(1/\alpha)}{\log k} + \frac{1}{\alpha\varepsilon}\Big), & \text{when } k \geq \frac{1}{\alpha}, \\ \Omega\big(k \log k + \frac{k}{\varepsilon}\big), & \text{when } k \leq \frac{1}{\alpha}. \end{cases}$$

Theorem 2.3. *Let $\lambda > 0$ be* any *small fixed constant. To be concrete, λ can be chosen to be any constant between* 0.01 *and* 1*. We have the following upper bounds on the sample complexity of entropy estimation:*

$$C(H, \alpha, \varepsilon) = O\left(\frac{k}{\alpha} + \frac{\log^2(\min\{k, n\})}{\alpha^2} + \frac{1}{\alpha\varepsilon} \log\left(\frac{1}{\alpha\varepsilon}\right)\right),$$

and

$$C(H,\alpha,\varepsilon) = O\left(\frac{k}{\lambda^2 \alpha \log k} + \frac{\log^2(\min\{k,n\})}{\alpha^2} + \left(\frac{1}{\alpha\varepsilon}\right)^{1+\lambda}\right).$$

Furthermore, [4]

$$C(H,\alpha,\varepsilon) = \Omega\left(\frac{k}{\alpha \log k} + \frac{\log^2(\min\{k,n\})}{\alpha^2} + \frac{\log k}{\alpha\varepsilon}\right).$$

These results can all be implemented in near-linear time in the number of samples. This is straightforward for most of our algorithms, though one must use the Remez algorithm to achieve this running time for entropy estimation.

We provide some discussion of our results. At a high level, we wish to emphasize the following two points:

(1) Our upper bounds show that the cost of privacy in these settings is often negligible compared to the sample complexity of the non-private statistical task, especially when for distributions with a large support. Furthermore, our upper bounds are almost tight in all parameters.
(2) The algorithmic complexity introduced by the requirement of privacy is minimal, consisting only of a single step which noises the output of an estimator. Therefore, our methods are realizable in practice, and we demonstrate their effectiveness on several synthetic and real-data examples.

Before we continue, we emphasize that, in Theorems 2.1 and 2.2, we consider the "sublinear" regime to be of primary interest (when $m \geq \frac{1}{\alpha\varepsilon}$ or $k \geq \frac{1}{\alpha\varepsilon}$, respectively), both technically, and, due to settings as discussed before where many symbols may be unseen, in terms of parameter regimes which may be of greatest interest in practice. We include results for other regimes mostly for completeness.

First, we examine our results on support coverage and support size estimation in the sublinear regime, when $m \geq \frac{1}{\alpha\varepsilon}$ (focusing on support coverage for simplicity, but support size is similar). In this regime, if $\varepsilon = \Omega(1/m^{1-\gamma})$ for any constant $\gamma > 0$, then up to constant factors, our upper bound is within a constant factor of the optimal sample complexity without privacy constraints. In other words, for most meaningful values of ε, privacy comes for free. In the non-sublinear regime for these problems, we provide upper and lower bounds which match in a number of cases. We note that in this regime, the cost of privacy may not be a lower order term – however, this regime only occurs when one requires very high accuracy, or unreasonably strict privacy, which we consider to be of somewhat less interest. For instance, most deployments of differential privacy we are aware of choose ε to be a small constant; choosing it to be inversely linear in m for large m would be rather unusual.

Next, we turn our attention to entropy estimation. We note that the second upper bound in Theorem 2.3 has a parameter λ that indicates a tradeoff between the sample complexity incurred in the first and third term. This parameter determines the degree of a polynomial to be used for entropy estimation. As the degree becomes smaller (corresponding to a large λ), accuracy of the polynomial estimator decreases, however, at the same time, low-degree polynomials have a small sensitivity, allowing us to privatize the outcome.

[4] A brief discussion on why the upper bounds are larger than the lower bound: (1) For the first upper bound, it is enough to show $k + 1/(\alpha\varepsilon)\log(1/(\alpha\varepsilon)) > \log k/(\alpha\varepsilon)$. This is equivalent to showing $k\alpha\varepsilon > \log(k\alpha\varepsilon)$, which is easy to see. (2) For the second upper bound, we only need to consider the case when $\log k/(\alpha\varepsilon) > k/(\alpha \log k)$, which implies $1/\varepsilon > k/(\log k)^2$. Then we have$(1/(\alpha\varepsilon))^{1+\lambda} \geq 1/(\alpha\varepsilon)(k/(\log k)^2)^{\lambda} = \Omega(\log k/(\alpha\varepsilon))$

In terms of our theoretical results, one can think of $\lambda = 0.01$. With this parameter setting, it can be observed that our upper bounds are almost tight. For example, one can see that the upper and lower bounds match to either logarithmic factors (when looking at the first upper bound), or a very small polynomial factor in $1/\alpha\varepsilon$ (when looking at the second upper bound). For our experimental results, we empirically determined an effective value for the parameter λ on a single synthetic instance. We then show that this choice of parameter generalizes, giving highly-accurate private estimation in other instances, on both synthetic and real-world data.

3. Algorithms and Analysis

We will prove our results for support coverage in Section 3.1, support size in Section 3.2, and entropy in Section 3.3. For each problem, we analyze our algorithms, proving an upper bound, and then describe a construction and prove a corresponding lower bound.

The general methodology of our algorithms is the following:

(1) Compute a non-private estimate of the property;
(2) Privatize this estimate by adding a Laplace noise, whose parameter is determined by analyzing the sensitivity of the estimator.

3.1. Support Coverage Estimation.

In this section, we prove Theorem 2.1, restated below. We prove the upper bound in Section 3.1.1, and the lower bound in Section 3.1.2.

Theorem 2.1. *The sample complexity of support coverage estimation is*

$$C(S_m, \alpha, \varepsilon) = \begin{cases} O\Big(\frac{m\log(1/\alpha)}{\log m} + \frac{m\log(1/\alpha)}{\log(2+\varepsilon m)}\Big), & \text{when } m \geq \frac{1}{\alpha\varepsilon}, \\ O\big(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon}\big), & \text{when } m \leq \frac{1}{\alpha\varepsilon}. \end{cases}$$

Furthermore,

$$C(S_m, \alpha, \varepsilon) = \Omega\left(\frac{m\log(1/\alpha)}{\log m} + \frac{1}{\alpha\varepsilon}\right).$$

3.1.1. *Upper Bound for Support Coverage Estimation.* We split the analysis into two regimes of α. First, we focus on the case where $m \leq \frac{2}{\alpha\varepsilon}$, and we prove the upper bound $O\left(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)$. The algorithm in this case is simple: assuming $n \geq m$, we group the dataset into n/m batches of size m.[5] Let Y_j be the number of unique symbols observed in batch j. Our estimator is

$$\hat{S}_m(X^n) = \frac{m}{n}\sum_{j=1}^{n/m} Y_j.$$

Observe that $\mathbb{E}[Y_j] = S_m(p)$, and that $\mathrm{Var}[Y_j] \leq m$. The latter can be seen by observing that Y_j is the sum of m negatively correlated indicator random variables, each one being the indicator of whether that sample in the batch is the first time the symbol is observed. This gives that $\hat{S}_m(X^n)$ is an unbiased estimator of $S_m(p)$, with variance $O(m^2/n)$. By Chebyshev's inequality, since we want an estimate which is accurate up to $\pm\alpha m$, this gives

[5] This will only affect the sample complexity by a constant since $m \leq \frac{2}{\alpha\varepsilon}$. Similarly, when $m > \frac{2}{\alpha\varepsilon}$, we assume $n \leq m$.

us that $C_{\hat{S}_m}(S_m(p), \alpha/2) = O\left(\frac{1}{\alpha^2}\right)$. Furthermore, we can see that the sensitivity of $\hat{S}_m(X^n)$ is at most $2m/n$. By Lemma 1.2, there is a private algorithm for support coverage estimation as long as

$$\Delta\left(\frac{\hat{S}_m(X^n)}{m}\right) \leq \alpha\varepsilon.$$

With the above bound on sensitivity, this is true with $n = O(1/\alpha\varepsilon)$, giving the desired upper bound.

Now, we turn our attention to the case where $m \geq \frac{2}{\alpha\varepsilon}$, and we prove the upper bound $O\left(\frac{m\log(1/\alpha)}{\log m} + \frac{m\log(1/\alpha)}{\log(2+\varepsilon m)}\right)$. Let φ_i be the number of symbols that appear i times in X^n. We will use the following non-private support coverage estimator from Orlitsky et al. (2016):

$$\hat{S}_m(X^n) = \sum_{i=1}^{n} \varphi_i\big(1 - (-t)^i \cdot \Pr\left(Z \geq i\right)\big), \tag{3.1}$$

where Z is a Poisson random variable with mean r (which is a parameter to be instantiated later), and $t = (m-n)/n$.

Our private estimator of support coverage is derived by adding Laplace noise to this non-private estimator with the appropriate noise parameter, and thus the performance of our private estimator, is analyzed by bounding the sensitivity and the bias of this non-private estimator according to Lemma 1.2.

Setting $r = \log(3/\alpha)$, Orlitsky et al. (2016) showed that there is a constant C, such that with $n = C\frac{m}{\log m}\log(3/\alpha)$ samples, with probability at least 0.9,

$$\left|\frac{\hat{S}_m(X^n)}{m} - \frac{S_m(p)}{m}\right| \leq \alpha.$$

If we change one sample in X^n, at most two of the φ_j's change. Orlitsky et al. (2016) also showed that the sensitivity of the estimator can be bounded by:

$$\Delta\left(\frac{\hat{S}_m(X^n)}{m}\right) \leq \frac{2}{m} \cdot \max_{i\in[n]}\left(1 - (-t)^i \cdot \Pr\left(Z \geq i\right)\right) \leq \frac{2}{m} \cdot \left(1 + e^{r(t-1)}\right). \tag{3.2}$$

By Lemma 1.2, there is a private algorithm for support coverage estimation as long as

$$\Delta\left(\frac{\hat{S}_m(X^n)}{m}\right) \leq \alpha\varepsilon,$$

which by (3.2) holds if

$$2(1 + \exp(r(t-1))) \leq \alpha\varepsilon m.$$

Let $r = \log(3/\alpha)$, note that $t - 1 = \frac{m}{n} - 2$.[6] Since $\alpha\varepsilon m > 2$, the condition above reduces to

$$\log\left(\frac{3}{\alpha}\right) \cdot \left(\frac{m}{n} - 2\right) \leq \log\left(\frac{1}{2}\alpha\varepsilon m - 1\right).$$

[6]Here we assume $n < m/2$. Else we can simply ignore samples other than the first $m/2$.

This is equivalent to

$$n \geq \frac{m\log(3/\alpha)}{\log(\frac{1}{2}\alpha\varepsilon m - 1) + 2\log(3/\alpha)} = \frac{m\log(3/\alpha)}{\log(\frac{3}{2}\varepsilon m - 3/\alpha) + \log(3/\alpha)}.$$

Then the condition above reduces to the requirement that

$$n = \Omega\left(\frac{m\log(1/\alpha)}{\log(2+\varepsilon m)}\right).$$

3.1.2. *Lower Bound for Support Coverage Estimation.* We now prove the following lower bound on the sample complexity of support coverage:

$$C(S_m, \alpha, \varepsilon) = \Omega\left(\frac{m\log(1/\alpha)}{\log m} + \frac{1}{\alpha\varepsilon}\right).$$

The first term is the sample complexity of non-private support coverage estimation, shown in Orlitsky et al. (2016). We therefore only prove the second term as a lower bound.

We first consider the case when $m\alpha \leq 1$. We construct the following two distributions: U_1 is uniform over $[m]$; U_2 is distributed over $m+1$ elements $[m] \cup \{\nabla\}$ where $U_2[i] = \frac{1-\alpha}{m}, \forall i \in [m]$ and $U_2[\nabla] = \alpha$. Then we have

$$S_m(U_1) = m\left(1 - \left(1 - \frac{1}{m}\right)^m\right), \text{ and } S_m(U_2) = m\left(1 - \left(1 - \frac{1-\alpha}{m}\right)^m\right) + 1 - (1-\alpha)^m.$$

Now we look at the difference between $S_m(U_1)$ and $S_m(U_2)$.

$$\begin{aligned}
S_m(U_2) - S_m(U_1) &= m\left(1 - \left(1 - \frac{1-\alpha}{m}\right)^m\right) - m\left(1 - \left(1 - \frac{1}{m}\right)^m\right) + (1 - (1-\alpha)^m) \\
&= (1 - (1-\alpha)^m) - m\left(\left(1 - \frac{1-\alpha}{m}\right)^m - \left(1 - \frac{1}{m}\right)^m\right) \\
&= (1 - (1-\alpha)^m) - m\frac{\alpha}{m}\left(\sum_{i=0}^{m-1}\left(1 - \frac{1-\alpha}{m}\right)^i\left(1 - \frac{1}{m}\right)^{m-1-i}\right) \\
&\geq (1 - (1-\alpha)^m) - \frac{m\alpha}{m} \cdot m\left(1 - \frac{1-\alpha}{m}\right)^{m-1} \\
&= (1 - (1-\alpha)^m) - m\alpha\left(1 - \frac{1-\alpha}{m}\right)^{m-1}. \qquad (3.3)
\end{aligned}$$

where the third equation comes from the fact that $\forall a, b \in \mathbb{R}, m \in \mathbb{N}^+, a^m - b^m = (a-b)\left(\sum_{i=0}^{m-1} a^i \cdot b^{m-1-i}\right)$.

When $m\alpha \leq 1$, the first term in (3.3) can be bounded by

$$1 - (1-\alpha)^m \geq 0.5m\alpha.$$

In order to bound the second term, we recall the folklore that $(1-\frac{c}{m})^m \le e^{-c}$ when $m \ge c \ge 0$. Then we can upper bound $\left(1-\frac{1-\alpha}{m}\right)^{m-1}$ by

$$\left(1-\frac{1-\alpha}{m}\right)^{m-1} \le \frac{1}{\left(1-\frac{1-\alpha}{m}\right)\cdot e^{1-\alpha}} \le \frac{1}{\left(1-\frac{1}{m}\right)\cdot e^{1-\alpha}}.$$

Suppose $m \ge 10$ and $\alpha \le \frac{1}{10}$,

$$\left(1-\frac{1-\alpha}{m}\right)^{m-1} \le \frac{1}{\left(1-\frac{1}{m}\right)\cdot e^{1-\alpha}} \le \frac{1}{0.9e^{0.9}} < 0.5.$$

Hence we conclude

$$S_m(U_2) - S_m(U_1) \ge \left(0.5 - \frac{1}{(0.9e)^{0.9}}\right)\cdot m\alpha = \Omega(m\alpha).$$

Now we move to the case when $m\alpha \ge 1$. In this situation, we slightly change the construction of U_2. Now U_2 is distributed over $m+1$ elements $[m]\cup\{\nabla\}$ where $U_2[i] = \frac{1-6\alpha}{m}, \forall i \in [m]$ and $U_2[\nabla] = 6\alpha$.

$$\begin{aligned}
S_m(U_1) - S_m(U_2) &= m\left(1-\left(1-\frac{1}{m}\right)^m\right) - (1-(1-6\alpha)^m) - m\left(1-\left(1-\frac{1-6\alpha}{m}\right)^m\right)\\
&= m\left(\left(1-\frac{1-6\alpha}{m}\right)^m - \left(1-\frac{1}{m}\right)^m\right) - (1-(1-6\alpha)^m)\\
&\ge \frac{6m\alpha}{m}\cdot m\left(1-\frac{1}{m}\right)^{m-1} - (1-(1-6\alpha)^m)\\
&= 6m\alpha\left(1-\frac{1}{m}\right)^{m-1} - (1-(1-6\alpha)^m)\\
&\ge 6m\alpha\left(1-\frac{1}{m}\right)^{m-1} - 1.
\end{aligned}$$

Since when $m > 0$, $\left(1-\frac{1}{m}\right)^{m-1} \ge \frac{1}{e}$, we have $S_m(U_1) - S_m(U_2) = \Omega(m\alpha)$.

Hence we know in both settings, their support coverage differs by $\Omega(\alpha m)$. Moreover, their total variation distance is $O\left(\frac{\alpha}{1+\alpha}\right)$. The following lemma is folklore, based on the coupling interpretation of total variation distance and the fact that total variation distance is subadditive for product measures.

Lemma 3.1. *For any two distributions p and q, there is a coupling between n i.i.d. samples from the two distributions with an expected Hamming distance $d_{\mathrm{TV}}(p,q)\cdot n$.*

Using Lemma 3.1 and $d_{\mathrm{TV}}(u_1,u_2) = O\left(\frac{\alpha}{1+\alpha}\right)$, we get the following.

Lemma 3.2. *Suppose u_1 and u_2 are as defined before, there is a coupling between u_1^n and u_2^n with expected Hamming distance equal to $O\left(\frac{n\alpha}{1+\alpha}\right)$.*

Moreover, given n samples, we must be able to privately distinguish between u_1 and u_2 given an α accurate estimator of support coverage with privacy considerations. Thus, according to Lemma 1.3 and 3.2, we must have $n = \Omega\left(\frac{1}{\alpha\varepsilon}\right)$.

3.2. **Support Size Estimation.** We will now prove Theorem 2.2, restated below. We prove the upper bound in Section 3.2.1 and the lower bound in Section 3.2.2.

Theorem 2.2. *The sample complexity of support size estimation is*

$$C(S,\alpha,\varepsilon)=\begin{cases}O\Big(\frac{k\log^2(1/\alpha)}{\log k}+\frac{k\log^2(1/\alpha)}{\log(2+\varepsilon k)}\Big), & \text{when } k\geq\frac{1}{\alpha\varepsilon},\\ O\big(k\log(1/\alpha)+\frac{1}{\alpha\varepsilon}\big), & \text{when } \frac{1}{\alpha}\leq k\leq\frac{1}{\alpha\varepsilon},\\ O\big(k\log k+\frac{k}{\varepsilon}\big), & \text{when } k\leq\frac{1}{\alpha}.\end{cases}$$

Furthermore,

$$C(S,\alpha,\varepsilon)=\begin{cases}\Omega\Big(\frac{k\log^2(1/\alpha)}{\log k}+\frac{1}{\alpha\varepsilon}\Big), & \text{when } k\geq\frac{1}{\alpha},\\ \Omega\big(k\log k+\frac{k}{\varepsilon}\big), & \text{when } k\leq\frac{1}{\alpha}.\end{cases}$$

3.2.1. *Upper Bound for Support Size Estimation.* We split the analysis into two regimes. First we consider the case when $k\geq\frac{2}{\alpha\varepsilon}$. In this case we show an upper bound of $O\Big(\frac{k\log^2(1/\alpha)}{\log k}+\frac{k\log^2(1/\alpha)}{\log(2+\varepsilon k)}\Big)$. This bound is $O\Big(\frac{k\log(3/\alpha)}{2}\Big)$ when $k\geq\frac{2}{\alpha\varepsilon}$. Hence we denote it by the "sparse" case.

Sparse case. In the sparse case, $k\geq\frac{2}{\alpha\varepsilon}$. In Orlitsky et al. (2016), it is shown that the support coverage estimator can be used to obtain optimal results for support size estimation. In particular, their result is based on the following observation which states that if $m=k\log(3/\alpha)$, then $S_m(p)$ is a *good estimate* for $S(p)$.

Lemma 3.3. *Suppose $m\geq k\log(3/\alpha)$, then for any $p\in\Delta_{\geq\frac{1}{k}}$,*

$$|S_m(p)-S(p)|\leq\frac{\alpha k}{3}.$$

Proof. Note that $S_m(p)\leq S(p)$ from its definition. For the other side, using $p(x)\geq 1/k$,

$$\begin{aligned}S(p)-S_m(p)&=\sum_{x:p(x)>0}(1-p(x))^m\leq\sum_{x:p(x)>0}e^{-mp(x)}\\&\leq k\cdot e^{-\log(3/\alpha)}=\frac{k\alpha}{3}.\end{aligned}\tag{3.4}$$

□

Further, we have:

$$\Big|\hat{S}_m(X^n)-S(p)\Big|\leq\Big|\hat{S}_m(X^n)-S_m(p)\Big|+|S_m(p)-S(p)|$$

Therefore, estimating $S_m(p)$ for $m=k\log(3/\alpha)$, up to $\pm\alpha k/3$, also estimates $S(p)$ up to $\pm\alpha k$. Using the same estimator in (3.1) with $t=\frac{k\log(3/\alpha)}{n}-1$, according to Orlitsky et al. (2016), $n=O\Big(\frac{m}{\log m}\log(1/\alpha)\Big)=O\Big(\frac{k}{\log k}\log^2(1/\alpha)\Big)$ samples are enough with success probability at least 0.9.

The computations for sensitivity are similar. From Lemma 1.2, we need to find the value of n such that

$$2+2e^{r(t-1)}\leq\alpha\varepsilon k,$$

where $r = \log(3/\alpha), t = \frac{k\log(3/\alpha)}{n} - 1$ and we assume that $n \le \frac{1}{2}k\log(3/\alpha)$, else we just keep the first $\frac{1}{2}k\log(3/\alpha)$ samples and discard the other ones. By computations similar to the previous case, this reduces to

$$n \ge \frac{k\log^2(3/\alpha)}{\log\frac{\alpha\varepsilon k}{2} + \log\frac{3}{\alpha}}.$$

Therefore

$$n = O\left(\frac{k\log^2(1/\alpha)}{\log(2+\varepsilon k)}\right) \tag{3.5}$$

for the sensitivity result to hold.

Dense case. Now, we consider $k \le \frac{2}{\alpha\varepsilon}$. Let N_x denote the number of times x appears in the sample X^n, and

$$W(X^n) \stackrel{\text{def}}{=} \{x : N_x > 0\}$$

be the set of symbols that appear in X^n. Our non-private estimator for support size is

$$\hat{S}(X^n) = \sum_{x\in W(X^n)} \min\left\{1, \frac{N_x}{\frac{n}{3k}}\right\}.$$

We analyze this algorithm for $k \le \frac{1}{\alpha}$ and $\frac{1}{\alpha} \le k \le \frac{2}{\alpha\varepsilon}$ separately.

When $k \le \frac{1}{\alpha}$, then $k\alpha \le 1$, and since the support size is an integer, we need to output the exact support size. $\hat{S}(X^n)$ is equal to $S(p)$ when all the symbols in the support appear at least $\frac{n}{3k}$ times. For any symbol x, since $p(x) \ge \frac{1}{k}$, by the Chernoff bound,

$$\Pr\left(N_x < \frac{n}{3k}\right) \le \exp\left(-\frac{2n^2/9k^2}{n\cdot\frac{1}{k}}\right) = \exp\left(-\frac{2n}{9k}\right). \tag{3.6}$$

When $n \ge 18k\log k$, we have $\Pr\left(N_x < \frac{n}{3k}\right) \le \frac{1}{k^4}$. Then by the union bound, the probability of all the symbols appearing at least $\frac{n}{3k}$ is greater than $1 - \frac{1}{k^3} > 2/3$ for $k \ge 2$, showing that our algorithm is correct with probability at least $\frac{2}{3}$.

Now, note that the sensitivity of $\hat{S}(X^n)$ is at most $3k/n$. By Lemma 1.2, we can privatize the non-private estimator as long as

$$\Delta\left(\hat{S}(X^n)\right) \le \varepsilon.$$

Since the sensitivity is at most $3k/n$, $n = O(k/\varepsilon)$, is enough to privatize the estimator. Combining the two claims above, $n = O\left(k\log k + \frac{k}{\varepsilon}\right)$ is enough to estimate the support size for the case when $k < 1/\alpha$.

Next, consider the case $\frac{1}{\alpha} \le k \le \frac{2}{\alpha\varepsilon}$ Assume without loss of generality that $\alpha < 0.5$. Let

$$Y(X^n) \stackrel{\text{def}}{=} \sum_{x\in S(p)} \mathbf{1}\{N_x \ge n/3k\},$$

be the number of symbols appearing at least $\frac{n}{3k}$ times in X^n. For any x with $p(x) \ge \frac{1}{k}$ from (3.6), $\Pr\left(N_x < \frac{n}{3k}\right) \le \alpha^2 \le 0.5\alpha$, since $\alpha < 0.5$. Therefore, by the linearity of expectations, $\mathbb{E}\left[Y(X^n)\right] > S(p)(1 - 0.5\alpha)$. Moreover, $\mathrm{Var}\left(Y(X^n)\right) < 0.5\alpha\cdot S(p)$ since it is the sum of $S(p)$ negatively correlated Bernoulli random variables with bias less than 0.5α.

By Chebyshev's inequality, and noting that $S(p) \le k$,

$$\Pr\left(Y(X^n) \geq S(p) + \alpha k\right) \leq \frac{0.5\alpha S(p)}{\alpha^2 k^2} \leq \frac{1}{2k\alpha}, \tag{3.7}$$

Similarly,

$$\Pr\left(Y(X^n) \leq S(p) - \alpha k\right) \leq \frac{0.5\alpha S(p)}{\left(\alpha k - 0.5\alpha S(p)\right)^2} \leq \frac{2}{k\alpha}, \tag{3.8}$$

Therefore,

$$\Pr\left(S(p) - \alpha k < Y(X^n) < S(p) + \alpha k\right) \geq 1 - \frac{5}{2k\alpha} \geq \frac{2}{3},$$

where the last inequality uses the fact that $k\alpha \geq c$, where c is some constant.

Furthermore, we can see that the sensitivity of $\hat{S}(X^n)$ is the same, which is at most $3k/n$.

By Lemma 1.2, there is a private algorithm for support coverage estimation as long as

$$\Delta\Big(\hat{S}(X^n)\Big) \leq k\alpha\varepsilon.$$

With the above bound on sensitivity, this is true with $n = O(\frac{1}{\alpha\varepsilon})$, giving the desired upper bound.

3.2.2. *Lower Bound for Support Size Estimation.* We will prove the following lower bound

$$C(S, \alpha, \varepsilon) = \begin{cases} \Omega\Big(\frac{k \log^2(1/\alpha)}{\log k} + \frac{1}{\alpha\varepsilon}\Big), & \text{when } k \geq \frac{1}{\alpha} \\ \Omega\big(k \log k + \frac{k}{\varepsilon}\big). & \text{when } k \leq \frac{1}{\alpha} \end{cases}$$

given in Theorem 2.2.

First we consider the case $k \geq \frac{1}{\alpha}$. The first term of the complexity is the lower bound for the non-private setting, which follows by combining the lower bound of Orlitsky et al. (2016) for support coverage, with the equivalence between estimation of support size and coverage as implied by Lemma 3.3. We prove the final term as a lower bound.

Let u_1 be the uniform distribution over $[k]$ and u_2 be the uniform distribution over $[(1-\alpha)k]$. Then, $S(u_1) - S(u_2) = \alpha k$, and $d_{\mathrm{TV}}(u_1, u_2) = \alpha$. Hence by Lemma 3.1, we know the following:

Lemma 3.4. *Suppose $u_1 \sim U[k]$ and $u_2 \sim U[(1-\alpha)k]$, there is a coupling between u_1^n and u_2^n with expected Hamming distance αn.*

If we can estimate the support size, then we can use it to obtain a tester that can distinguish between u_1, u_2. However, we know from Lemma 1.3 and Lemma 3.4, that to distinguish between any two distributions with total variation distance α, using an ε-DP algorithm, we must have

$$\alpha n \geq \frac{1}{\varepsilon} \Rightarrow n = \Omega\left(\frac{1}{\varepsilon\alpha}\right).$$

This proves the second term when $k \geq \frac{1}{\alpha}$.

Then we move to the second case when $k \leq \frac{1}{\alpha}$. Because $k\alpha < 1$, we need to recover the support size exactly. The first term of the complexity is the lower bound for the non-private setting which can be proved using a coupon-collector-style argument, so here we focus on the second term.

We consider the following two distributions: u_1 is a uniform distribution over $[k]$ and u_2 is a uniform distribution over $[k-1]$. We must distinguish between these two distributions, for which $d_{\mathrm{TV}}(u_1, u_2) = \frac{1}{k}$. This is equivalent to choosing $\alpha = 1/k$ in the argument above, and therefore we obtain

$$\frac{n}{k} \geq \frac{1}{\varepsilon} \Rightarrow n = \Omega\left(\frac{k}{\varepsilon}\right).$$

3.3. **Entropy Estimation.** In this section, we prove our main theorem about entropy estimation, Theorem 2.3, restated below. We describe and analyze two upper bounds. The first is based on the empirical entropy estimator, and is described and analyzed in Section 3.3.1. The second is based on the method of best-polynomial approximation, and appears in Section 3.3.2. We prove the lower bound in Section 3.3.3.

Theorem 2.3. *Let $\lambda > 0$ be* any *small fixed constant. To be concrete, λ can be chosen to be any constant between* 0.01 *and* 1*. We have the following upper bounds on the sample complexity of entropy estimation:*

$$C(H, \alpha, \varepsilon) = O\left(\frac{k}{\alpha} + \frac{\log^2(\min\{k, n\})}{\alpha^2} + \frac{1}{\alpha\varepsilon}\log\left(\frac{1}{\alpha\varepsilon}\right)\right),$$

and

$$C(H, \alpha, \varepsilon) = O\left(\frac{k}{\lambda^2 \alpha \log k} + \frac{\log^2(\min\{k, n\})}{\alpha^2} + \left(\frac{1}{\alpha\varepsilon}\right)^{1+\lambda}\right).$$

Furthermore, [7]

$$C(H, \alpha, \varepsilon) = \Omega\left(\frac{k}{\alpha \log k} + \frac{\log^2(\min\{k, n\})}{\alpha^2} + \frac{\log k}{\alpha\varepsilon}\right).$$

3.3.1. *Upper Bound for Entropy Estimation: The Empirical Estimator.* Our first private entropy estimator is derived by adding Laplace noise into the empirical entropy estimator. Let $\hat{p}_n$ denote the empirical distribution from the samples X^n. Let $\Delta(H(\hat{p}_n))$ denote the sensitivity of the empirical entropy. Then, our privatized empirical entropy estimator adds a Laplace noise with parameter $\frac{\Delta(H(\hat{p}_n))}{\varepsilon}$, to $H(\hat{p}_n)$. By analyzing its sensitivity and bias of this estimator, we obtain the first upper bound in Theorem 2.3.

Let $\hat{p}_n$ be the empirical distribution, and let $H(\hat{p}_n)$ be the entropy of the empirical distribution. The theorem is based on the following three facts:

$$\Delta(H(\hat{p}_n)) = O\left(\frac{\log n}{n}\right), \tag{3.9}$$

$$|H(p) - \mathbb{E}\left[H(\hat{p}_n)\right]| = O\left(\frac{k}{n}\right), \tag{3.10}$$

$$\mathrm{Var}\left(H(\hat{p}_n)\right) = O\left(\frac{\log^2(\min\{k, n\})}{n}\right). \tag{3.11}$$

[7] A brief discussion on why the upper bounds are larger than the lower bound: (1) For the first upper bound, it is enough to show $k + 1/(\alpha\varepsilon)\log(1/(\alpha\varepsilon)) > \log k/(\alpha\varepsilon)$. This is equivalent to showing $k\alpha\varepsilon > \log(k\alpha\varepsilon)$, which is easy to see. (2) For the second upper bound, we only need to consider the case when $\log k/(\alpha\varepsilon) > k/(\alpha \log k)$, which implies $1/\varepsilon > k/(\log k)^2$. Then we have$(1/(\alpha\varepsilon))^{1+\lambda} \geq 1/(\alpha\varepsilon)(k/(\log k)^2)^{\lambda} = \Omega(\log k/(\alpha\varepsilon))$

With these three facts in hand, by Lemma 1.2, the sample complexity of the empirical estimator is any n for which $\Delta(H(\hat{p}_n)) \le \alpha\varepsilon$, $|H(p) - \mathbb{E}[H(\hat{p}_n)]| = O(\alpha)$ and $\mathrm{Var}(H(\hat{p}_n)) = O(\alpha^2)$. From the equations above, these are satisfied when $n = O\left(\frac{k}{\alpha} + \frac{\log^2(\min\{k,n\})}{\alpha^2}\right)$. We now prove the three facts.

Proof of (3.9). When one symbol changes, at most two N_x's change, each by at most one. Therefore,

$$\begin{aligned}
\Delta(H(\hat{p}_n)) &\le 2 \cdot \max_{j=1\ldots n-1} \left| \frac{j+1}{n} \log \frac{n}{j+1} - \frac{j}{n} \log \frac{n}{j} \right| \\
&= 2 \cdot \max_{j=1\ldots n-1} \left| \frac{j}{n} \log \frac{j}{j+1} + \frac{1}{n} \log \frac{n}{j+1} \right| \\
&\le 2 \cdot \max_{j=1\ldots n-1} \max \left\{ \left| \frac{j}{n} \log \frac{j}{j+1} \right|, \left| \frac{1}{n} \log \frac{n}{j+1} \right| \right\} \\
&\le 2 \cdot \max \left\{ \frac{1}{n}, \frac{\log n}{n} \right\}, \\
&= 2 \cdot \frac{\log n}{n}.
\end{aligned}$$

Proof of (3.10). The variance bound below is from Paninski (2003). By concavity of entropy,

$$\mathbb{E}[H(\hat{p}_n)] \le H(p).$$

Therefore,

$$\begin{aligned}
&\mathbb{E}[|H(p) - H(\hat{p}_n)|] = H(p) - \mathbb{E}[H(\hat{p}_n)] \\
&= \mathbb{E}\left[\sum_x (\hat{p}_n(x) \log \hat{p}_n(x) - p(x) \log p(x)) \right] \\
&= \mathbb{E}\left[\sum_x \hat{p}_n(x) \log \frac{\hat{p}_n(x)}{p(x)} \right] + \mathbb{E}\left[\sum_x (\hat{p}_n(x) - p(x)) \log p(x) \right] \\
&= \mathbb{E}[D(\hat{p}_n \| p)] \\
&\le \mathbb{E}\left[\mathrm{d}_{\chi^2}(\hat{p}_n \,||\, p) \right] \\
&= \mathbb{E}\left[\sum_x \frac{(\hat{p}_n(x) - p(x))^2}{p(x)} \right] \\
&\le \sum_x \frac{(p(x)/n)}{p(x)} \\
&= \frac{k}{n}.
\end{aligned}$$

Proof of (3.11). The variance bound of $\frac{\log^2 k}{n}$ is given precisely in Lemma 15 of Jiao et al. (2017). To obtain the bound of $\frac{\log^2 n}{n}$, we apply the bounded differences inequality in the form stated in Corollary 3.2 of Boucheron et al. (2013).

Lemma 3.5. *Let $f : \Omega^n \to \mathbb{R}$ be a function. Suppose further that*

$$\max_{z_1,\ldots,z_n,z_i'} \left| f(z_1,\ldots,z_n) - f(z_1,\ldots,z_{i-1},z_i',\ldots,z_n) \right| \le c_i.$$

Then for independent variables $Z_1,\ldots,Z_n$,

$$\mathrm{Var}(f(Z_1,\ldots,Z_n)) \le \frac{1}{4}\sum_{i=1}^{n} c_i^2.$$

Therefore, by Lemma 3.5 and (3.9)

$$\mathrm{Var}\left(H(\hat{p}_n)\right) \le n \cdot \left(\frac{4\log^2 n}{n^2}\right) = \frac{4\log^2 n}{n}.$$

3.3.2. *Upper Bound for Entropy Estimation: Best-Polynomial Approximation.* We now obtain the second upper bound in Theorem 2.3. Best polynomial approximation has been used to obtain sample-optimal entropy estimators in the non-private setting. We add a suitable Laplace noise to this estimate to privatize it.

In the non-private setting the optimal sample complexity of estimating $H(p)$ over Δ_k is given by Theorem 1 of Wu and Yang (2016)

$$\Theta\left(\frac{k}{\alpha \log k} + \frac{\log^2(\min\{k,n\})}{\alpha^2}\right).$$

However, this estimator can have a large sensitivity. Acharya et al. (2017a) designed an estimator that has the same sample complexity but a smaller sensitivity. We restate Lemma 6 of Acharya et al. (2017a) here:

Lemma 3.6. *Let $\lambda > 0$ be a fixed small constant, which may be taken to be any value between 0.01 and 1. Then there is an entropy estimator with sample complexity*

$$\Theta\left(\frac{1}{\lambda^2}\cdot\frac{k}{\alpha \log k} + \frac{\log^2(\min\{k,n\})}{\alpha^2}\right),$$

and has sensitivity n^{λ}/n.

We can now invoke Lemma 1.2 on the estimator in this lemma to obtain the upper bound on private entropy estimation.

3.3.3. *Lower Bound for Entropy Estimation.* We now prove the lower bound for entropy estimation. Note that any lower bound on privately testing two distributions p, and q such that $H(p) - H(q) = \Theta(\alpha)$ is a lower bound on estimating entropy.

We analyze the following construction for Proposition 2 of Wu and Yang (2016). The two distributions p and q over $[k]$ are defined as:

$$p(1) = \frac{2}{3},\ p(i) = \frac{1-p(1)}{k-1}, \text{for } i = 2, \ldots, k, \tag{3.12}$$

$$q(1) = \frac{2-\eta}{3},\ q(i) = \frac{1-q(1)}{k-1}, \text{for } i = 2, \ldots, k. \tag{3.13}$$

Then, by the grouping property of entropy,

$$H(p) = h(2/3) + \frac{1}{3} \cdot \log(k-1), \text{ and } H(q) = h((2-\eta)/3) + \frac{1+\eta}{3} \cdot \log(k-1),$$

which gives

$$H(p) - H(q) = \Omega(\eta \log k).$$

For $\eta = \alpha / \log k$, the entropy difference becomes $\Theta(\alpha)$.

The total variation distance between p and q is $\eta/3$. By Lemma 3.1, there is a coupling over X^n and Y^n generated from p and q with expected Hamming distance at most $d_{\mathrm{TV}}(p,q) \cdot n$. This along with Lemma 1.3 gives a lower bound of $\Omega(\log k / \alpha\varepsilon)$ on the sample complexity.

4. Experiments

We evaluated our methods for entropy estimation and support coverage on both synthetic and real data. Overall, we found that privacy is quite cheap: private estimators achieve accuracy which is comparable or near-indistinguishable to non-private estimators in many settings. Our results on entropy estimation and support coverage appear in Sections 4.1 and 4.2, respectively. Code of our implementation is available at `https://github.com/HuanyuZhang/INSPECTRE` and archived as Acharya et al. (2020).

4.1. **Entropy.** We compare the performance of our entropy estimator with a number of alternatives, both private and non-private. Non-private algorithms considered include the plug-in estimator (`plug-in`), the Miller-Madow Estimator (`MM`) (Miller, 1955), the sample optimal polynomial approximation estimator (`poly`) of Wu and Yang (2016). We analyze the privatized versions of plug-in and `poly` in Sections 3.3.1 and 3.3.2, respectively. The implementation of the latter is based on code from the authors of Wu and Yang (2016)[8]. We compare performance on different distributions including uniform, a distribution with two steps, Zipf(1/2), a distribution with Dirichlet-1 prior, and a distribution with Dirichlet-1/2 prior, and over varying support sizes.

While `plug-in`and `MM` are parameter free, `poly` (and its private counterpart) have to choose the degree L of the polynomial to use, which manifests in the parameter λ in the statement of Theorem 2.3. Wu and Yang (2016) suggests the value of $L = 1.6 \log k$ in their experiments. However, since we add further noise, we choose a single L as follows: (i) Run privatized `poly` for different L values and distributions for $k = 2000$, $\varepsilon = 1$, (b) Choose the value of L that performs well across different distributions (See Figure 1). We choose $L = 1.2 \cdot \log k$ from this, and use it for all other experiments. To evaluate the

[8]See `https://github.com/Albuso0/entropy` for their code for entropy estimation.

sensitivity of poly, we computed the estimator's value at all possible input values, computed the sensitivity, (namely, $\Delta = \max_{d_{ham}(X^n,Y^n)\le 1} |\texttt{poly}(X^n) - \texttt{poly}(Y^n)|$), and added noise distributed as Lap $\left(0, \frac{\Delta}{\varepsilon}\right)$.

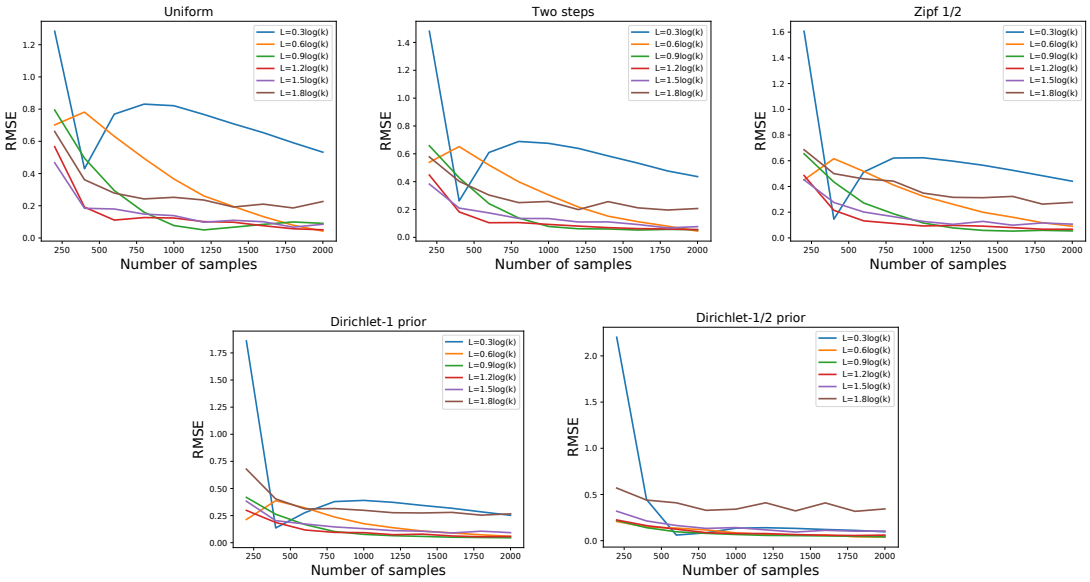


FIGURE 1. RMSE comparison between private Polynomial Approximation Estimators for entropy with various values for degree L, $k = 2000$, $\varepsilon = 1$. The degree L represents a bias-variance tradeoff: a larger degree decreases the bias but increases the sensitivity, necessitating the addition of Laplace noise with a larger variance.

The RMSE (Root Mean Square Error) of various estimators for $k = 1000$ and $\varepsilon = 1$ for various distributions are illustrated in Figure 2. The RMSE is averaged over 100 iterations in the plots.

We observe that the performance of our private-poly is near-indistinguishable from the non-private poly, particularly as the number of samples increases. It also performs significantly better than all other alternatives, including the non-private Miller-Madow and the plug-in estimator. The cost of privacy is minimal for several other settings of k and ε, for which results appear in Appendix A.

4.2. **Support Coverage.** We investigate the cost of privacy for the problem of support coverage. We provide a comparison between the Smoothed Good-Toulmin estimator (SGT) of Orlitsky et al. (2016) and our algorithm, which is a privatized version of their statistic (see Section 3.1.1). Our implementation is based on code provided by the authors of Orlitsky et al. (2016). As shown in our theoretical results, the sensitivity of SGT is at most $2(1 + e^r(t-1))$, necessitating the addition of Laplace noise with parameter $2(1 + e^{r(t-1)})/\varepsilon$. Note that while the theory suggests we select the parameter $r = \log(1/\alpha)$, α is unknown. We instead set $r = \frac{1}{2t} \log_e \frac{n(t+1)^2}{t-1}$, as previously done in Orlitsky et al. (2016).

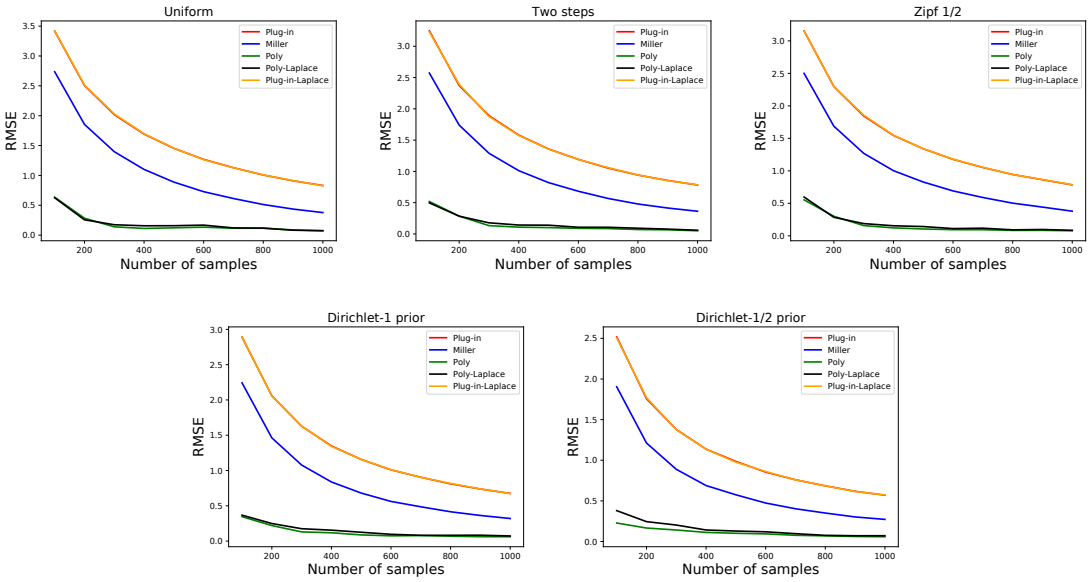


FIGURE 2. Comparison of various estimators for entropy, $k = 1000$, $\varepsilon = 1$.

4.2.1. *Evaluation on Synthetic Data.* In our synthetic experiments, we consider different distributions over different support sizes k. We generate $n = k/2$ samples, and then estimate the support coverage at $m = n \cdot t$. For large t, estimation is harder. Some results of our evaluation on synthetic data are displayed in Figure 3. We compare the performance of SGT and privatized versions of SGT with parameters $\varepsilon = 1, 2$, and 10. For this instance, we fixed the domain size $k = 20000$. We ran the methods described above with $n = k/2$ samples and estimated the support coverage at $m = nt$, for t ranging from 1 to 10. The performance of the estimators is measured in terms of RMSE over 1000 iterations.

We observe that, in this setting, the cost of privacy is relatively small for reasonable values of ε. This is as predicted by our theoretical results, where unless ε is extremely small (less than $1/k$) the non-private sample complexity dominates the privacy requirement. However, we found that for smaller support sizes (as shown in Section A.2), the cost of privacy can be significant. We provide an intuitive explanation for why no private estimator can perform well on such instances. To minimize the number of parameters, we instead argue about the related problem of support-size estimation. Suppose we are trying to distinguish between distributions which are uniform over supports of size 100 and 200. We note that, if we draw $n = 50$ samples, the "profile" of the samples (i.e., the histogram of the histogram) will be very similar for the two distributions. In particular, if one modifies only a few samples (say, five or six), one could convert one profile into the other. In other words, these two profiles are almost-neighboring datasets, but simultaneously correspond to very different support sizes. This pits the two goals of privacy and accuracy at odds with each other, thus resulting in a degradation in accuracy.

4.2.2. *Evaluation on Census Data and Hamlet.* We conclude with experiments for support coverage on two real-world datasets, the 2000 US Census data and the text of Shakespeare's play Hamlet, inspired by investigations in Orlitsky et al. (2016) and Valiant and Valiant

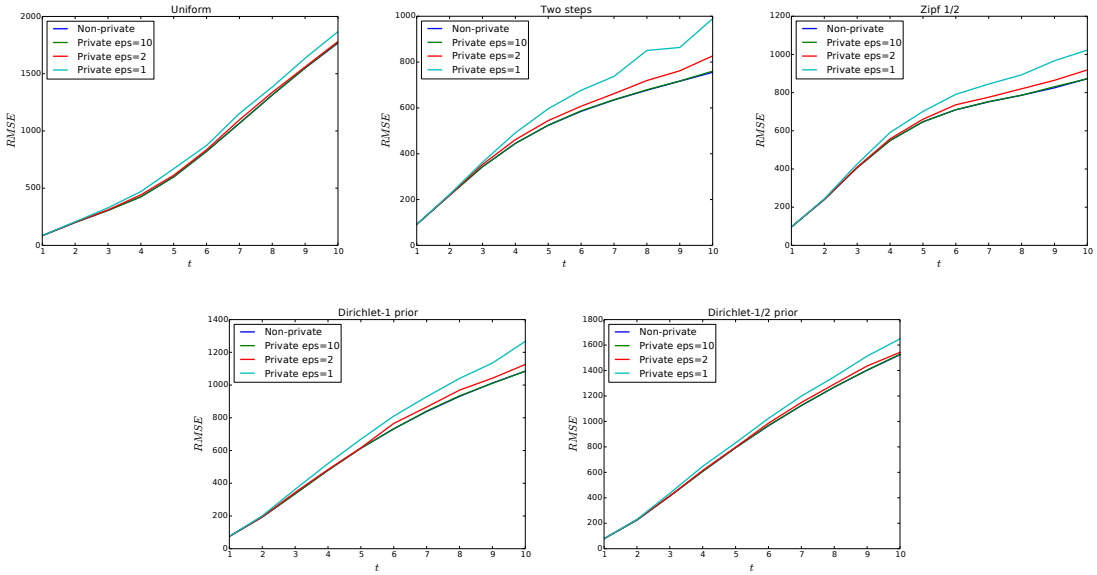


FIGURE 3. Comparison between our private support coverage estimator with non-private SGT when $k = 20000$

(2017). Our investigation on US Census data is also inspired by the fact that this is a setting where privacy is of practical importance, evidenced by the proposed adoption of differential privacy in the 2020 US Census (Dajani et al., 2017).

The Census dataset contains a list of last names that appear at least 100 times. Since the dataset is so oversampled, even a small fraction of the data is likely to contain almost all the names. As such, we make the task non-trivial by subsampling $m_{total} = 86080$ individuals from the data, obtaining 20412 distinct last names. We then sample n of the m_{total} individuals without replacement and attempt to estimate the total number of last names. Figure 4 displays the RMSE over 100 iterations of this process. We observe that even with an exceptionally stringent privacy budget of $\varepsilon = 0.5$, the performance is almost indistinguishable from the non-private SGT estimator.

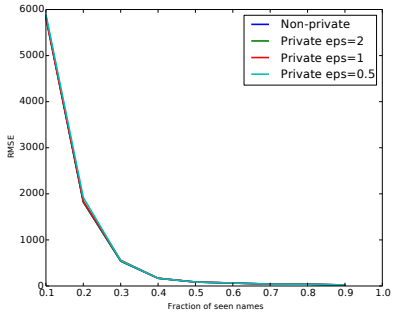


FIGURE 4. Comparison between our private support coverage estimator with the SGT on Census Data.

The Hamlet dataset has $m_{total} = 31,999$ words, of which 4804 are distinct. Since the distribution is not as oversampled as the Census data, we do not need to subsample the data. Besides this difference, the experimental setup is identical to that of the Census dataset. Once again, as we can see in Figure 5, we get near-indistinguishable performance between the non-private and private estimators, even for very small values of ε. Our experimental results demonstrate that privacy is realizable in practice, with particularly accurate performance on real-world datasets.

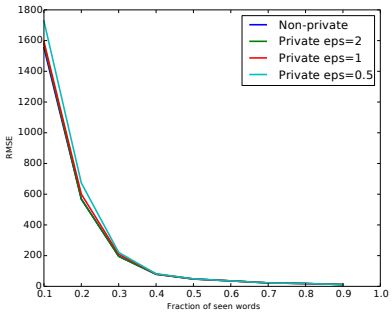


FIGURE 5. Comparison between our private support coverage estimator with the SGT on Hamlet.

References

Acharya, Jayadev, Clément L. Canonne, Cody Freitag, and Himanshu Tyagi (2019a) "Test without Trust: Optimal Locally Private Distribution Testing," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, AISTATS '19: JMLR, Inc. URL: https://arxiv.org/abs/1808.02174.

Acharya, Jayadev, Hirakendu Das, Alon Orlitsky, and Ananda Theertha Suresh (2017a) "A Unified Maximum Likelihood Approach for Estimating Symmetric Properties of Discrete Distributions," in *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pp. 11–21: JMLR, Inc. URL: http://proceedings.mlr.press/v70/acharya17a.html.

Acharya, Jayadev, Constantinos Daskalakis, and Gautam Kamath (2015) "Optimal Testing for Properties of Distributions," in *Advances in Neural Information Processing Systems 28*, NIPS '15, pp. 3577–3598, URL: http://papers.nips.cc/paper/5839-optimal-testing-for-properties-of-distributions.pdf.

Acharya, Jayadev, Gautam Kamath, Ziteng Sun, and Huanyu Zhang (2018a) "INSPECTRE: Privately Estimating the Unseen," in *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pp. 30–39: JMLR, Inc. URL: http://proceedings.mlr.press/v80/acharya18a.html.

——— (2020) "Code and data for INSPECTRE: Privately Estimating the Unseen," Software v202004-jpc, Zenodo.

Acharya, Jayadev, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi (2017b) "Estimating Rényi Entropy of Discrete Distributions," *IEEE Transactions on Information Theory*, Vol. 63, No. 1, pp. 38–56, DOI: 10.1109/TIT.2016.2620435.

Acharya, Jayadev, Ziteng Sun, and Huanyu Zhang (2018b) "Differentially Private Testing of Identity and Closeness of Discrete Distributions," in *Advances in Neural Information Processing Systems 31*, NeurIPS '18, pp. 6879–6891, URL: http://papers.nips.cc/paper/7920-differentially-private-testing-of-identity-and-closeness-of-discrete-distributions.pdf.

——— (2019b) Personal communication.

——— (2019c) "Hadamard Response: Estimating Distributions Privately, Efficiently, and with Little Communication," in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, AISTATS '19: JMLR, Inc. URL: https://arxiv.org/abs/1802.04705.

Adam, Nabil R. and John C. Worthmann (1989) "Security-Control Methods for Statistical Databases: A Comparative Study," *ACM Computing Surveys (CSUR)*, Vol. 21, No. 4, pp. 515–556, DOI: 10.1145/76894.76895.

Agrawal, Dakshi and Charu C. Aggarwal (2001) "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: ACM, DOI: 10.1145/375551.375602.

Aliakbarpour, Maryam, Ilias Diakonikolas, and Ronitt Rubinfeld (2018) "Differentially Private Identity and Closeness Testing of Discrete Distributions," in *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pp. 169–178: JMLR, Inc. URL: http://proceedings.mlr.press/v80/aliakbarpour18a/aliakbarpour18a.pdf.

Awan, Jordan and Aleksandra Slavkovic (2018) "Differentially Private Uniformly Most Powerful Tests for Binomial Data," in *Advances in Neural Information Processing Systems 31*, NeurIPS '18, pp. 4212–4222, URL: https://papers.nips.cc/paper/7675-differentially-private-uniformly-most-powerful-tests-for-binomial-data.pdf.

Balcan, Maria-Florina, Travis Dick, Yingyu Liang, Wenlong Mou, and Hongyang Zhang (2017) "Differentially Private Clustering in High-Dimensional Euclidean Spaces," in *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pp. 322–331: JMLR, Inc. URL: http://proceedings.mlr.press/v70/balcan17a.html.

Berry, Michael J., David K Warland, and Markus Meister (1997) "The Structure and Precision of Retinal Spike Trains," *Proceedings of the National Academy of Sciences*, Vol. 94, No. 10, pp. 5411–5416, DOI: 10.1073/pnas.94.10.5411.

Boucheron, Stephane, Gabor Lugosi, and Pierre Massart (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence*: Oxford University Press, DOI: 10.1093/acprof:oso/9780199535255.001.0001.

Brenner, Hai and Kobbi Nissim (2014) "Impossibility of Differentially Private Universally Optimal Mechanisms," *SIAM Journal on Computing*, Vol. 43, No. 5, pp. 1513–1540, DOI: 10.1109/FOCS.2010.13.

Cai, Bryan, Constantinos Daskalakis, and Gautam Kamath (2017) "Priv'IT: Private and Sample Efficient Identity Testing," in *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pp. 635–644: JMLR, Inc. URL: http://proceedings.mlr.press/v70/cai17a.html.

Campbell, Zachary, Andrew Bray, Anna Ritz, and Adam Groce (2018) "Differentially Private ANOVA Testing," in *Proceedings of the 2018 International Conference on Data Intelligence and Security*, ICDIS '18, pp. 281–285, Washington, DC, USA: IEEE Computer Society, DOI: 10.1109/icdis.2018.00052.

Canonne, Clément L., Gautam Kamath, Audra McMillan, Adam Smith, and Jonathan Ullman (2019) "The Structure of Optimal Private Tests for Simple Hypotheses," in *Proceedings of the 50th Annual ACM Symposium on the Theory of Computing*, STOC '19, New York, NY, USA: ACM, URL: https://arxiv.org/abs/1811.11148.

Chaudhuri, Kamalika, Anand D. Sarwate, and Kaushik Sinha (2013) "A Near-Optimal Algorithm for Differentially-Private Principal Components," *Journal of Machine Learning Research*, Vol. 14, No. Sep, pp. 2905–2943, URL: http://dl.acm.org/citation.cfm?id=2567709.2567754.

Colwell, Robert K., Anne Chao, Nicholas J. Gotelli, Shang-Yi Lin, Chang Xuan Mao, Robin L. Chazdon, and John T. Longino (2012) "Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages," *Journal of Plant Ecology*, Vol. 5, No. 1, pp. 3–21, DOI: 10.1093/jpe/rtr044.

Couch, Simon, Zeki Kazan, Kaiyan Shi, Andrew Bray, and Adam Groce (2019) "Differentially Private Nonparametric Hypothesis Testing," *arXiv preprint arXiv:1903.09364*, URL: https://arxiv.org/abs/1903.09364.

Cover, Thomas M. and Joy A. Thomas (2006) *Elements of Information Theory*: Wiley, DOI: 10.1002/047174882X.

Dajani, Aref N., Amy D. Lauger, Phyllis E. Singer, Daniel Kifer, Jerome P. Reiter, Ashwin Machanavajjhala, Simson L. Garfinkel, Scot A. Dahl, Matthew Graham, Vishesh Karwa, Hang Kim, Philip Lelerc, Ian M. Schmutte, William N. Sexton, Lars Vilhuber, and John M. Abowd (2017) "The Modernization of Statistical Disclosure Limitation at the U.S. Census

Bureau," URL: https://www2.census.gov/cac/sac/meetings/2017-09/statistical-disclosure-limitation.pdf.

Dalenius, Tore (1977) "Towards a Methodology for Statistical Disclosure Control," *Statistisk Tidskrift*, Vol. 15, pp. 429–444, URL: https://www.semanticscholar.org/paper/Towards-a-methodology-for-statistical-disclosure-Dalenius/c40cf1310102d09d9818853fda0b5a9efe0a21c2.

Diakonikolas, Ilias, Moritz Hardt, and Ludwig Schmidt (2015) "Differentially Private Learning of Structured Discrete Distributions," in *Advances in Neural Information Processing Systems 28*, NIPS '15, pp. 2566–2574, URL: http://papers.nips.cc/paper/5713-differentially-private-learning-of-structured-discrete-distributions.pdf.

Differential Privacy Team, Apple (2017) "Learning with Privacy at Scale," December, URL: https://machinelearning.apple.com/docs/learning-with-privacy-at-scale/appledifferentialprivacysystem.pdf.

Ding, Bolin, Janardhan Kulkarni, and Sergey Yekhanin (2017) "Collecting telemetry data privately," in *Advances in Neural Information Processing Systems*, pp. 3571–3580, URL: https://papers.nips.cc/paper/6948-collecting-telemetry-data-privately.pdf.

Dinur, Irit and Kobbi Nissim (2003) "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: ACM, DOI: 10.1145/773153.773173.

Duchi, John C., Michael I. Jordan, and Martin J. Wainwright (2018) "Minimax Optimal Procedures for Locally Private Estimation," *Journal of the Americal Statistical Association*, Vol. 113, No. 521, pp. 182–201, DOI: 10.1080/01621459.2017.1389735.

Dwork, Cynthia (2008) "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: Springer, URL: https://doi.org/10.1007/978-3-540-79228-4_1.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith (2006) "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: Springer, URL: https://doi.org/10.1007/978-3-540-32732-5_32.

Dwork, Cynthia and Aaron Roth (2014) "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends® in Machine Learning*, Vol. 9, No. 3–4, pp. 211–407, DOI: 10.1561/0400000042.

Erlingsson, Úlfar, Vasyl Pihur, and Aleksandra Korolova (2014) "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response," in *Proceedings of the 2014 ACM Conference on Computer and Communications Security*, CCS '14, pp. 1054–1067, New York, NY, USA: ACM, DOI: 10.1145/2660267.2660348.

Fukuchi, Kazuto and Jun Sakuma (2017) "Minimax Optimal Estimators for Additive Scalar Functionals of Discrete Distributions," in *Proceedings of the 2017 IEEE International Symposium on Information Theory*, ISIT '17, pp. 2103–2107, Washington, DC, USA: IEEE Computer Society, DOI: 10.1109/ISIT.2017.8006900.

Gaboardi, Marco, Hyun-Woo Lim, Ryan M. Rogers, and Salil P. Vadhan (2016) "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. URL: http://proceedings.mlr.press/v48/rogers16.html.

Gaboardi, Marco and Ryan Rogers (2018) "Local Private Hypothesis Testing: Chi-Square Tests," in *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pp. 1626–1635: JMLR, Inc. URL: http://proceedings.mlr.press/v80/gaboardi18a/gaboardi18a.pdf.

Good, I.J. and G.H. Toulmin (1956) "The Number of New Species, and the Increase in Population Coverage, when a Sample is Increased," *Biometrika*, Vol. 43, No. 1-2, pp. 45–63, DOI: 10.2307/2333577.

Hardt, Moritz and Eric Price (2014) "The Noisy Power Method: A Meta Algorithm with Applications," in *Advances in Neural Information Processing Systems 27*, NIPS '14, pp. 2861–2869, URL: https://papers.nips.cc/paper/5326-the-noisy-power-method-a-meta-algorithm-with-applications.

Homer, Nils, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig (2008) "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures using High-Density SNP Genotyping Microarrays," *PLoS Genetics*, Vol. 4, No. 8, pp. 1–9, DOI: 10.1371/journal.pgen.1000167.

Jiao, Jiantao, Kartik Venkat, Yanjun Han, and Tsachy Weissman (2017) "Minimax estimation of functionals of discrete distributions," *IEEE Transactions on Information Theory*, Vol. 61, No. 5, pp. 2835–2885, DOI: 10.1109/TIT.2015.2412945.

Johnson, Aaron and Vitaly Shmatikov (2013) "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: ACM, DOI: 10.1145/2487575.2487687.

Kakizaki, Kazuya, Jun Sakuma, and Kazuto Fukuchi (2017) "Differentially Private Chi-squared Test by Unit Circle Mechanism," in *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pp. 1761–1770: JMLR, Inc. URL: http://proceedings.mlr.press/v70/kakizaki17a.html.

Kamath, Gautam, Jerry Li, Vikrant Singhal, and Jonathan Ullman (2018) "Privately Learning High-Dimensional Distributions," *arXiv preprint arXiv:1805.00216*, URL: https://arxiv.org/abs/1805.00216.

Kapralov, Michael and Kunal Talwar (2013) "On Differentially Private Low Rank Approximation," in *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '13, pp. 1395–1414, Philadelphia, PA, USA: SIAM, URL: https://dl.acm.org/citation.cfm?id=2627918.

Karwa, Vishesh and Salil Vadhan (2018) "Finite Sample Differentially Private Confidence Intervals," in *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pp. 44:1–44:9, Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, URL: https://privacytools.seas.harvard.edu/files/privacytools/files/itcs_version.pdf.

Keinan, Alon and Andrew G. Clark (2012) "Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants," *Science*, Vol. 336, No. 6082, pp. 740–743, DOI: 10.1126/science.1217283.

Kifer, Daniel and Ryan M. Rogers (2017) "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. URL: http://proceedings.mlr.press/v54/rogers17a/rogers17a.pdf.

Miller, George A. (1955) "Note on the Bias of Information Estimates," *Information Theory in Psychology: Problems and Methods*, Vol. 2, pp. 95–100, URL: https://www.scienceopen.com/document?vid=357d299f-62fa-4bda-8dd2-e4d5b5abde5d.

Nelson, Matthew R., Daniel Wegmann, Margaret G. Ehm, Darren Kessner, Pamela St. Jean, Claudio Verzilli, Judong Shen, Zhengzheng Tang, Silviu-Alin Bacanu, Dana Fraser, Liling Warren, Jennifer Aponte, Matthew Zawistowski, Xiao Liu, Hao Zhang, Yong Zhang, Jun Li, Yun Li, Li Li, Peter Woollard, Simon Topp, Matthew D. Hall, Keith Nangle, Jun Wang, Gonçalo Abecasis, Lon R. Cardon, Sebastian Zöllner, John C. Whittaker, Stephanie L. Chissoe, John Novembre, and Vincent Mooser (2012) "An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People," *Science*, Vol. 337, No. 6090, pp. 100–104, DOI: 10.1126/science.1217876.

Nemenman, Ilya, William Bialek, and Rob de Ruyter van Steveninck (2004) "Entropy and Information in Neural Spike Trains: Progress on the Sampling Problem," *Physical Review E*, Vol. 69, No. 5, pp. 056111:1–056111:6, DOI: 10.1103/PhysRevE.69.056111.

Nowozin, Sebastian (2012) "Improved Information Gain Estimates for Decision Tree Induction," in *Proceedings of the 29th International Conference on Machine Learning*, ICML '12, pp. 571–578: JMLR, Inc. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.371.359.

Obremski, Maciej and Maciej Skorski (2017) "Rényi Entropy Estimation Revisited," in *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques.*, APPROX '17, pp. 20:1–20:15, Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, DOI: 10.4230/LIPIcs.APPROX-RANDOM.2017.20.

Orlitsky, Alon, Ananda Theerta Suresh, and Yihong Wu (2016) "Optimal Prediction of the Number of Unseen Species," *Proceedings of the National Academy of Sciences*, Vol. 113, No. 47, pp. 13283–13288, DOI: 10.1073/pnas.1607774113.

Paninski, Liam (2003) "Estimation of Entropy and Mutual Information," *Neural Computation*, Vol. 15, No. 6, pp. 1191–1253, DOI: 10.1162/089976603321780272.

Raghunathan, Aditi, Greg Valiant, and James Zou (2017) "Estimating the Unseen from Multiple Populations," in *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pp. 2855–2863: JMLR, Inc. URL: http://proceedings.mlr.press/v70/raghunathan17a.html.

Raskhodnikova, Sofya, Dana Ron, Amir Shpilka, and Adam Smith (2009) "Strong Lower Bounds for Approximating Distribution Support Size and the Distinct Elements Problem," *SIAM Journal on Computing*, Vol. 39, No. 3, pp. 813–842, DOI: 10.1137/070701649.

Rogers, Ryan Michael (2017) *Leveraging Privacy in Data Analysis* Ph.D. dissertation, University of Pennsylvania, URL: https://repository.upenn.edu/edissertations/2554/.

Sheffet, Or (2017) "Differentially Private Ordinary Least Squares," in *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pp. 3105–3114: JMLR, Inc. URL: http://proceedings.mlr.press/v70/sheffet17a.html.

——— (2018) "Locally Private Hypothesis Testing," in *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pp. 4605–4614: JMLR, Inc. URL: http://proceedings.mlr.press/v80/sheffet18a/sheffet18a.pdf.

Simmons, Sean, Cenk Sahinalp, and Bonnie Berger (2016) "Enabling Privacy-preserving GWASs in Heterogeneous Human Populations," *Cell Systems*, Vol. 3, No. 1, pp. 54–61, DOI: 10.1016/j.cels.2016.04.013.

Smith, Adam (2011) "Privacy-Preserving Statistical Estimation with Optimal Convergence Rates," in *Proceedings of the 43rd Annual ACM Symposium on the Theory of Computing*,

STOC '11, pp. 813–822, New York, NY, USA: ACM, DOI: 10.1145/1993636.1993743.

Swanberg, Marika, Ira Globus-Harris, Iris Griffith, Anna Ritz, Adam Groce, and Andrew Bray (2019) "Improved Differentially Private Analysis of Variance," *Proceedings on Privacy Enhancing Technologies*, Vol. 2019, No. 3, URL: https://arxiv.org/abs/1903.00534.

Tennessen, Jacob A., Abigail W. Bigham, Timothy D. O'Connor, Wenqing Fu, Eimear E. Kenny, Simon Gravel, Sean McGee, Ron Do, Xiaoming Liu, Goo Jun, Hyun Min Kang, Daniel Jordan, Suzanne M. Leal, Stacey Gabriel, Mark J. Rieder, Goncalo Abecasis, David Altshuler, Deborah A. Nickerson, Eric Boerwinkle, Shamil Sunyaev, Carlos D. Bustamante, Michael J. Bamshad, Joshua M. Akey, Broad GO, Seattle GO, and on behalf of the NHLBI Exome Sequencing Project (2012) "Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes," *Science*, Vol. 337, No. 6090, pp. 64–69, DOI: 10.1126/science.1219240.

Uhler, Caroline, Aleksandra Slavković, and Stephen E. Fienberg (2013) "Privacy-preserving Data Sharing for Genome-wide Association Studies," *The Journal of Privacy and Confidentiality*, Vol. 5, No. 1, pp. 137–166, DOI: 10.1016/j.jbi.2014.01.008.

Valiant, Gregory and Paul Valiant (2013) "Estimating the Unseen: Improved Estimators for Entropy and Other Properties," in *Advances in Neural Information Processing Systems 26*, NIPS '13, pp. 2157–2165, URL: https://papers.nips.cc/paper/5170-estimating-the-unseen-improved-estimators-for-entropy-and-other-properties.

_____ (2016) "Instance Optimal Learning of Discrete Distributions," in *Proceedings of the 48th Annual ACM Symposium on the Theory of Computing*, STOC '16, pp. 142–155, New York, NY, USA: ACM, DOI: 10.1145/2897518.2897641.

_____ (2017) "Estimating the Unseen: Improved Estimators for Entropy and Other Properties," *Journal of the ACM*, Vol. 64, No. 6, pp. 37:1–37:41, DOI: 10.1145/3125643.

Vu, Duy and Aleksandra Slavković (2009) "Differential Privacy for Clinical Trial Data: Preliminary Evaluations," in *2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, pp. 138–143: IEEE, DOI: 10.1109/ICDMW.2009.52.

Wang, Yining, Yu-Xiang Wang, and Aarti Singh (2015b) "Differentially Private Subspace Clustering," in *Advances in Neural Information Processing Systems 28*, NIPS '15, pp. 1000–1008, URL: https://dl.acm.org/citation.cfm?id=2969351.

Wang, Yue, Jaewoo Lee, and Daniel Kifer (2015a) "Revisiting Differentially Private Hypothesis Tests for Categorical Data," *arXiv preprint arXiv:1511.03376*, URL: https://arxiv.org/abs/1511.03376.

Wu, Yihong and Pengkun Yang (2016) "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Transactions on Information Theory*, Vol. 62, No. 6, pp. 3702–3720, DOI: 10.1109/TIT.2016.2548468.

_____ (2019) "Chebyshev Polynomials, Moment Matching, and Optimal Estimation of the Unseen," *The Annals of Statistics*, Vol. 13, No. 2, pp. 768–774, URL: https://projecteuclid.org/euclid.aos/1547197241.

Yu, Fei, Stephen E. Fienberg, Aleksandra B. Slavković, and Caroline Uhler (2014) "Scalable Privacy-Preserving Data Sharing Methodology for Genome-Wide Association Studies," *Journal of Biomedical Informatics*, Vol. 50, pp. 133–141, DOI: 10.1016/j.jbi.2014.01.008.

Zou, James, Gregory Valiant, Paul Valiant, Konrad Karczewski, Siu On Chan, Kaitlin Samocha, Monkol Lek, Shamil Sunyaev, Mark Daly, and Daniel G. MacArthur (2016) "Quantifying Unobserved Protein-Coding Variants in Human Populations Provides a Roadmap for Large-Scale Sequencing Projects," *Nature Communications*, Vol. 7, DOI: 10.1038/ncomms13293.

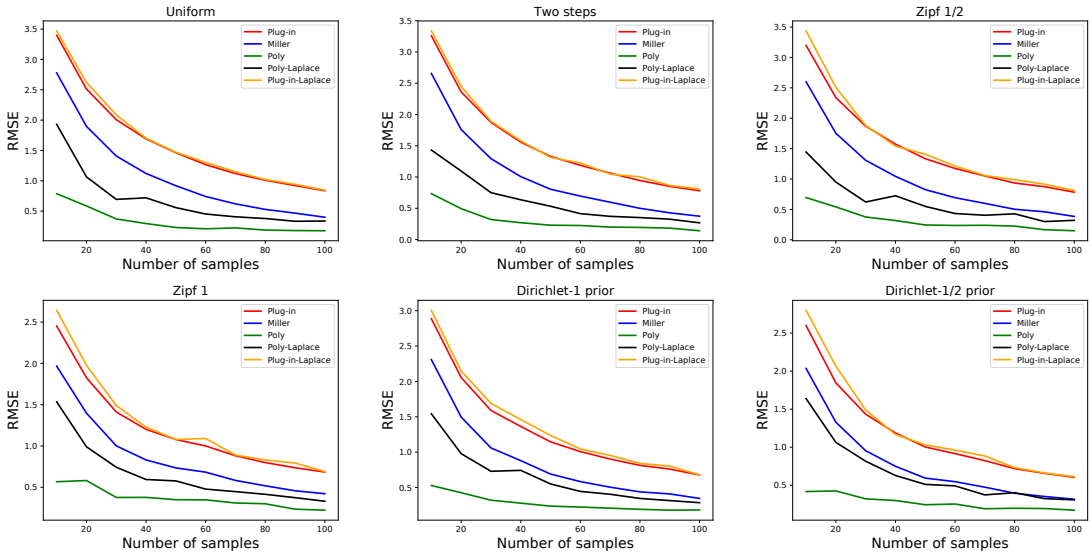


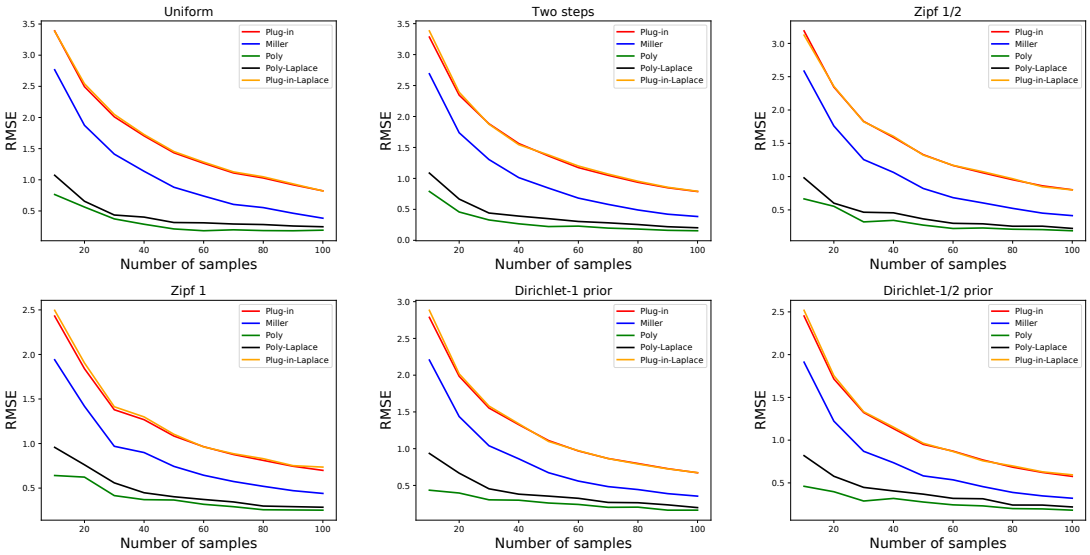
FIGURE 6. Comparison of various estimators for the entropy, $k = 100$, $\varepsilon = 1$.

APPENDIX A. ADDITIONAL EXPERIMENTAL RESULTS

This section contains additional plots of our synthetic experimental results. Section A.1 contains experiments on entropy estimation, while Section A.2 contains experiments on estimation of support coverage.

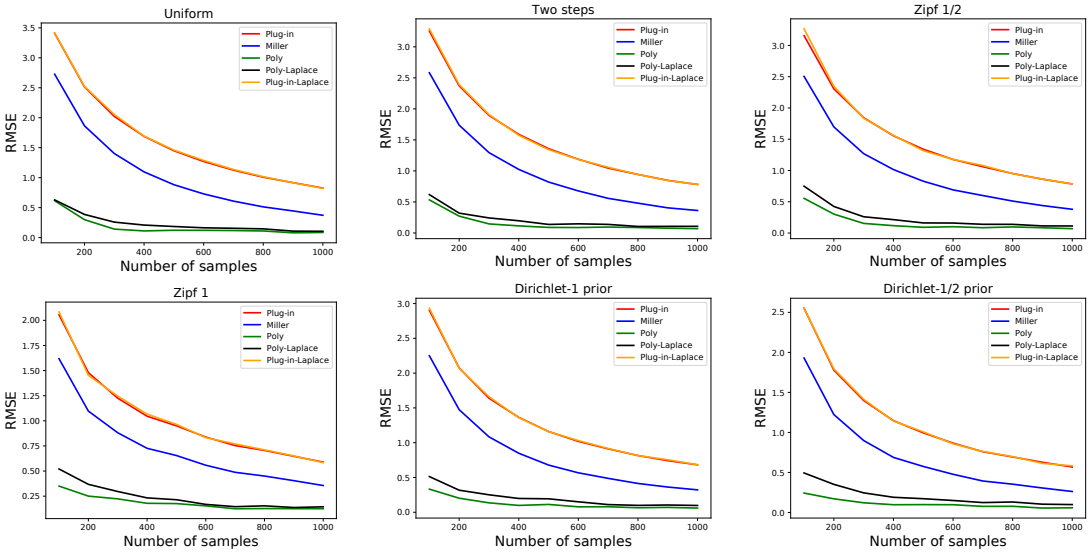
A.1. **Entropy Estimation.** We present four more plots of our synthetic experimental results for entropy estimation. Figures 6 and 7 are on a smaller support of $k = 100$, with $\varepsilon = 1$ and 2, respectively. Figures 8 and 9 are on a support of $k = 1000$, with $\varepsilon = 0.5$ and 2.

A.2. **Support Coverage.** We present three additional plots of our synthetic experimental results for support coverage estimation. In particular, Figures 10, 11, and 12 show support coverage for $k = 1000$, 5000, 100000.



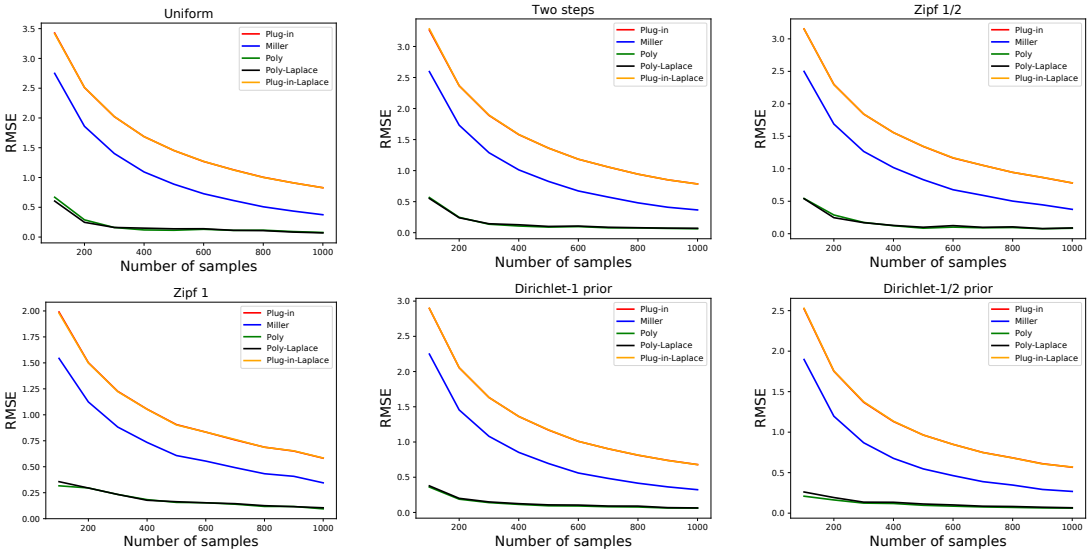
(a) (b) (c)

FIGURE 7. Comparison of various estimators for the entropy, $k = 100$, $\varepsilon = 2$.



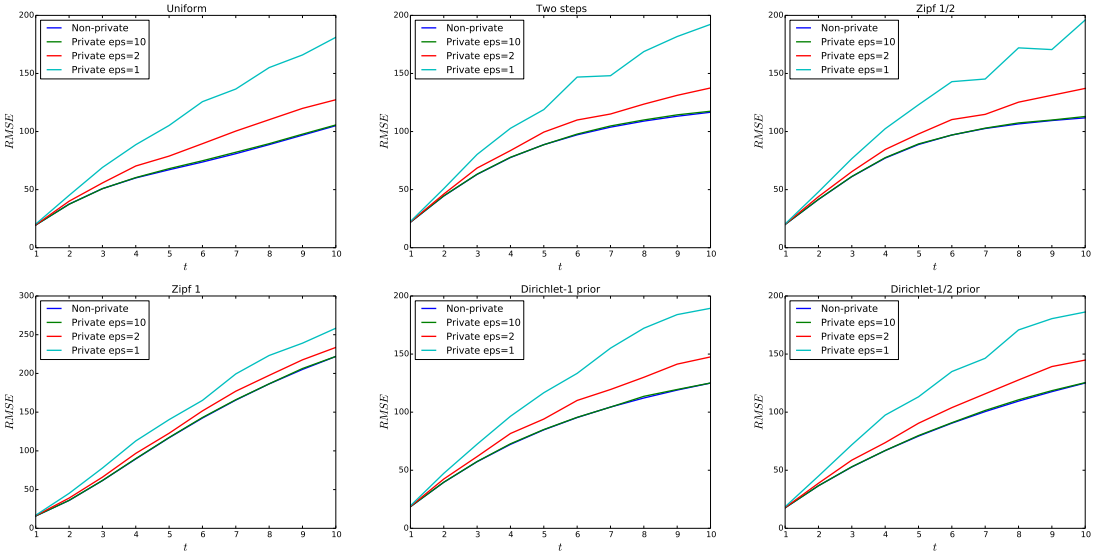
(a) (b) (c)

FIGURE 8. Comparison of various estimators for the entropy, $k = 1000$, $\varepsilon = 0.5$.



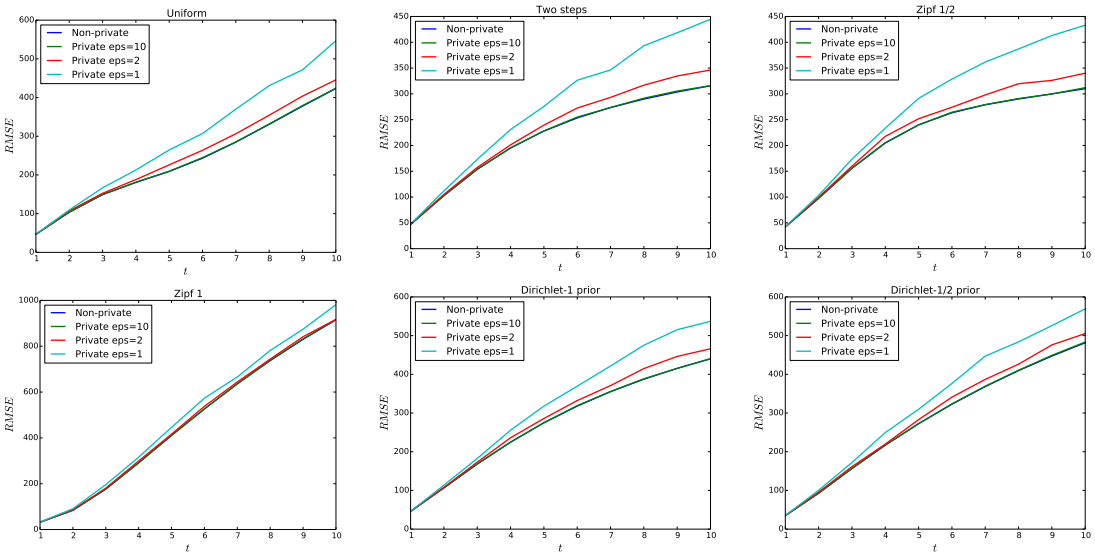
(a) (b) (c)

FIGURE 9. Comparison of various estimators for the entropy, $k = 1000$, $\varepsilon = 2$.



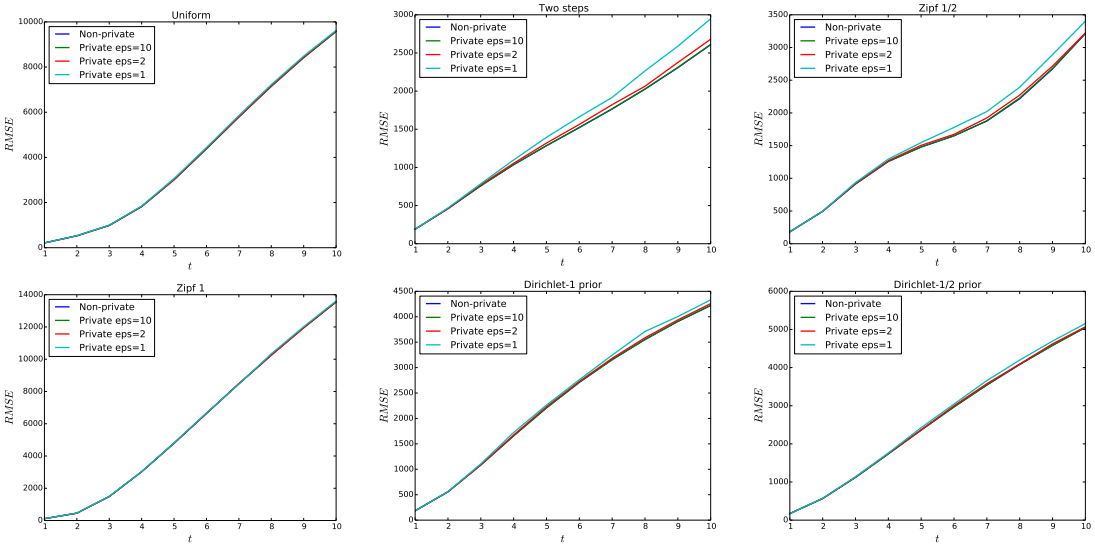
(a) (b) (c)

FIGURE 10. Comparison between the private estimator with the non-private SGT when $k = 1000$.



(a) (b) (c)

FIGURE 11. Comparison between the private estimator with the non-private SGT when $k = 5000$.



(a) (b) (c)

FIGURE 12. Comparison between the private estimator with the non-private SGT when $k = 100000$.