A Primer on Private Statistics

Gautam Kamath[*] Jonathan Ullman[†]

April 17, 2020

1 Introduction

Statistics and machine learning are now ubiquitous in data analysis. Given a dataset, one immediately wonders what it allows us to infer about the underlying population. However, modern datasets don't exist in a vacuum: they often contain sensitive information about the individuals they represent. Without proper care, statistical procedures will result in gross violations of privacy. Motivated by the shortcomings of ad hoc methods for data anonymization, Dwork, McSherry, Nissim, and Smith introduced the celebrated notion of differential privacy [DMNS06].

From its inception, some of the driving motivations for differential privacy were applications in statistics and the social sciences, notably disclosure limitation for the US Census. And yet, the lion's share of differential privacy research has taken place within the computer science community. As a result, the specific applications being studied are often not formulated using statistical terminology, or even as statistical problems. Perhaps most significantly, much of the early work in computer science (though definitely not all) focus on estimating some property *of a dataset* rather than estimating some property *of an underlying population.*

Although the earliest works exploring the interaction between differential privacy and classical statistics go back to at least 2009 [VS09, FRY10], the emphasis on differentially private statistical inference in the computer science literature is somewhat more recent. However, while earlier results on differential privacy did not always formulate problems in a statistical language, statistical inference was a key motivation for most of this work. As a result many of the techniques that were developed have direct applications in statistics, for example establishing minimax rates for estimation problems.

The purpose of this series of blog posts is to highlight some of those results in the computer science literature, and present them in a more statistical language. Specifically, we will discuss:

- Tight minimax lower bounds for privately estimating the mean of a multivariate distribution over $\mathbb{R}^d$, using the technique of *tracing attacks* developed in [BUV14, SU17a, DSS$^+$15, BSU17, SU17b, KLSU19].

[*] Cheriton School of Computer Science, University of Waterloo. g@csail.mit.edu. Supported by a University of Waterloo startup grant.

[†] Khoury College of Computer Sciences, Northeastern University. jullman@ccs.neu.edu. Supported by NSF grants CCF-1750640, CNS-1816028, and CNS-1916020.

- Upper bounds for estimating a distribution in Kolmogorov distance, using the ubiquitous *binary-tree mechanism* introduced in [DNPR10, CSS11].

In particular, we hope to encourage computer scientists working on differential privacy to pay more attention to the applications of their methods in statistics, and share with statisticians many of the powerful techniques that have been developed in the computer science literature.

1.1 Formulating Private Statistical Inference

Essentially every differentially private statistical estimation task can be phrased using the following setup. We are given a dataset $X = (X_1, \ldots, X_n)$ of size n, and we wish to design an algorithm $M \in \mathcal{M}$ where $\mathcal{M}$ is the class of mechanisms that are both:

1. differentially private, and
2. accurate, either in expectation or with high probability, according to some task-specific measure.

A few comments about this framework are in order. First, although the accuracy requirement is stochastic in nature (i.e., an algorithm might not be accurate depending on the randomness of the algorithm and the data generation process), the privacy requirement is worst-case in nature. That is, the algorithm must protect privacy for every dataset X, even those we believe are very unlikely.

Second, the accuracy requirement is stated rather vaguely. This is because the notion of accuracy of an algorithm is slightly more nuanced, depending on whether we are concerned with *empirical* or *population* statistics. A particular emphasis of these blog posts is to explore the difference (or, as we will see, the lack of a difference) between these two notions of accuracy. The former estimates a quantity of the observed dataset, while the latter estimates a quantity of an unobserved distribution which is assumed to have generated the dataset.

More precisely, the former can be phrased in terms of empirical loss, of the form:

$$\min_{M \in \mathcal{M}} \max_{X \in \mathcal{X}} \mathop{\mathbb{E}}_{M}(\ell(M(X), f(X))),$$

where $\mathcal{M}$ is some class of *randomized estimators* (e.g., differentially private estimators), $\mathcal{X}$ is some class of *datasets*, f is some quantity of interest, and ℓ is some *loss function*. That is, we're looking to find an estimator that has small expected loss on *any dataset* in some class.

In contrast, statistical minimax theory looks at statements about population loss, of the form:

$$\min_{M \in \mathcal{M}} \max_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X \sim P, M}(\ell(M(X), f(P))),$$

where $\mathcal{P}$ is some family of *distributions* over datasets (typically consisting of i.i.d. samples). That is, we're looking to find an estimator that has small expected loss on random data from *any distribution* in some class. In particular, note that the randomness in this objective additionally includes the data generating procedure $X \sim P$.

These two formulations are formally very different in several ways. First, the empirical formulation requires an estimator to have small loss on *worst-case* datasets, whereas the statistical formulation only requires the estimator to have small loss *on average* over datasets drawn from certain distributions. Second, the statistical formulation requires that we estimate the unknown quantity $f(P)$, and thus necessitates a solution to the non-private estimation problem. On the other hand, the empirical formulation only asks us to estimate the known quantity $f(X)$, and thus if there were no privacy constraint it would always be possible to compute $f(X)$ exactly. Third, typically in the statistical formulation, we require that the dataset is drawn i.i.d., which means that we are more constrained when proving lower bounds for estimation than we are in the empirical problem.

However, in practice,[1] these two formulations are more alike than they are different, and results about one formulation often imply results about the other formulation. On the algorithmic side, classical statistical results will often tell us that $\ell(f(X), f(P))$ is small, in which case algorithms that guarantee $\ell(M(X), f(X))$ is small also guarantee $\ell(M(X), f(P))$ is small.

Moreover, typical lower bound arguments for empirical quantities are often statistical in nature. These typically involving constructing some simple "hard distribution" over datasets such that no private algorithm can estimate well on average for this distribution, and thus these lower bound arguments also apply to estimating population statistics for some simple family of distributions.

We will proceed to give some examples of estimation problems that were originally studied by computer scientists with the empirical formulation in mind. These results either implicitly or explicitly provide solutions to the corresponding population versions of the same problems—our goal is to spell out and illustrate these connections.

2 Differential Privacy Background

Let $X = (X_1, X_2, \ldots, X_n) \in \mathcal{X}^n$ be a collection of n samples where each individual sample comes from the domain $\mathcal{X}$. We say that two samples $X, X' \in \mathcal{X}^*$ are *adjacent*, denoted $X \sim X'$, if they differ on at most one individual sample. Intuitively, a randomized algorithm M, which is often called a *mechanism* for historical reasons, is *differentially private* if the distribution of $M(X)$ and $M(X')$ are similar for every pair of adjacent samples X, X'.

Definition 2.1 ([DMNS06]). *A mechanism $M \colon \mathcal{X}^n \to \mathcal{R}$ is* (ε, δ)-differentially private *if for every pair of adjacent datasets $X \sim X'$, and every (measurable) $R \subseteq R$*

$$\mathbb{P}(M(X) \in R) \le e^{\varepsilon} \cdot \mathbb{P}\big(M(X') \in R\big) + \delta.$$

We let $\mathcal{M}_{\varepsilon,\delta}$ denote the set of mechanisms that satisfy (ε, δ)-differential privacy.

Remark 2.2. *To simplify notation, and to maintain consistency with the literature, we adopt the convention of defining the mechanism only for a fixed sample size n. What this means in practice is that the mechanisms we describe treat the sample size n is* public information *that need not be kept private. While one could define a more general model where n is not fixed, it wouldn't add anything to this discussion other than additional complexity.*

Remark 2.3. *In these blog posts, we stick to the most general formulation of differential privacy, so-called* approximate differential privacy*, i.e. (ε, δ)-differential privacy for $\delta > 0$*

[1] More precisely, in the practice of doing theoretical research.

essentially because this is the notion that captures the widest variety of private mechanisms. Almost all of what follows would apply equally well, with minor technical modifications, to slightly stricter notions of concentrated differential privacy *[DR16, BS16, BDRS18],* Rényi differential privacy *[Mir17], or* Gaussian differential privacy *[DRS19]. While so-called* pure differential privacy*, i.e.* $(\varepsilon, 0)$*-differential privacy has also been studied extensively, this notion is artificially restrictive and excludes many differentially private mechanisms.*

A key property of differential privacy that helps when desinging efficient estimators is *closure under postprocessing*:

Lemma 2.4 (Post-Processing [DMNS06])**.** *If* $M\colon \mathcal{X}^n \to \mathcal{R}$ *is* (ε, δ)*-differentially private and* $M'\colon \mathcal{R} \to \mathcal{R}'$ *is any randomized algorithm, then* $M' \circ M$ *is* (ε, δ)*-differentially private.*

The estimators we present in this work will use only one tool for achieving differential privacy, the *Gaussian Mechanism.*

Lemma 2.5 (Gaussian Mechanism)**.** *Let* $f\colon \mathcal{X}^n \to \mathbb{R}^d$ *be a function and let*

$$\Delta_f = \sup_{X \sim X'} \|f(X) - f(X')\|_2$$

denote its ℓ_2-sensitivity*. The* Gaussian mechanism

$$M(X) = f(X) + \mathcal{N}\left(0, \frac{2\log(2/\delta)}{\varepsilon^2} \cdot \Delta_f^2 \cdot \mathbb{I}_{d \times d}\right)$$

satisfies (ε, δ)*-differential privacy.*

3 Mean Estimation in $\mathbb{R}^d$

Let's take a dive into the problem of *private mean estimation* for some family $\mathcal{P}$ of multivariate distributions over $\mathbb{R}^d$. This problem has been studied for various families $\mathcal{P}$ and various choices of loss function. Here we focus on perhaps the simplest variant of the problem, in which $\mathcal{P}$ contains distributions of bounded support $[\pm 1]^d$ and the loss is the ℓ_2^2 error. We emphasize, however, that the methods we discuss here are quite versatile and can be used to derive minimax bounds for other variants of the mean-estimation problem.

Note that, by a simple argument, the non-private minimax rate for this class is achieved by the empirical mean, and is

$$\max_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X_{1\cdots n} \sim P} \left(\|\overline{X} - \mu\|_2^2\right) = \frac{d}{n}. \tag{1}$$

The main goal of this section is to derive the minimax bound

$$\min_{M \in \mathcal{M}_{\varepsilon, \frac{1}{n}}} \max_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X_{1\cdots n} \sim P} \left(\|M(X_{1\cdots n}) - \mu\|_2^2\right) = \frac{d}{n} + \tilde{\Theta}\left(\frac{d^2}{\varepsilon^2 n^2}\right).^{2} \tag{2}$$

[2] $\tilde{\Theta}(f(n))$ is a slight abuse of notation – it refers to a function which is both $O(f(n)\log^{c_1} f(n))$ and $\Omega(f(n)\log^{c_2} f(n))$ for some constants c_1, c_2.

The proof of this lower bound is based on *robust tracing attacks*, also called *membership inference attacks*, which were developed in a chain of papers [BUV14, SU17a, DSS+15, BSU17, SU17b, KLSU19]. We remark that this lower bound is almost identical to the minimax bound for mean estimation proven in the much more recent work of Cai, Wang, and Zhang [CWZ19], but it lacks tight dependence on the parameter δ, which we discuss in the following remark.

Remark 3.1. *The choice of $\delta = 1/n$ in (2) may look strange at first. For the upper bound this choice is arbitrary—as we will see, we can upper bound the rate for any $\delta > 0$ at a cost of a factor of $O(\log(1/\delta))$. The lower bound applies only when $\delta \leq 1/n$. Note that the rate is qualitatively different when $\delta \gg 1/n$. However, we emphasize that (ε, δ)-differential privacy is not a meaningful privacy notion unless $\delta \ll 1/n$. In particular, the mechanism that randomly outputs δn elements of the sample satisfies $(0, \delta)$-differential privacy. However, when $\delta \gg 1/n$, this mechanism completely violates the privacy of $\gg 1$ person in the dataset. Moreover, taking the empirical mean of these δn samples gives rate $d/\delta n$, which would violate our lower bound when δ is large enough. On the other hand, we would expect the minimax rate to become slower when $\delta \ll 1/n$. This expectation is, in fact, correct, however the proof we present does not give the tight dependence on the parameter δ. See [SU17a] for a refinement that can obtain the right dependence on δ, and [CWZ19] for the details of how to apply this refinement in the i.i.d. setting.*

3.1 A Simple Upper Bound

Theorem 3.2. *For every $n \in \mathbb{N}$, and every $\varepsilon, \delta > 0$, there exists an (ε, δ)-differentially private private mechanism M such that*

$$\max_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X_{1\cdots n} \sim P} \left(\|M(X_{1\cdots n}) - \mu\|_2^2 \right) \leq \frac{d}{n} + \frac{2d^2 \log(2/\delta)}{\varepsilon^2 n^2}. \tag{3}$$

Proof. Define the mechanism

$$M(X_{1\cdots n}) = \overline{X} + \mathcal{N}\left(0, \frac{2d \log(2/\delta)}{\varepsilon^2 n^2} \cdot \mathbb{I}_{d \times d}\right). \tag{4}$$

This mechanism satisfies (ε, δ)-differential privacy by Lemma 2.5, noting that for any pair of adjacent samples $X_{1\cdots n}$ and $X'_{1\cdots n}$, $\|\overline{X} - \overline{X}'\|_2^2 \leq \frac{d}{n^2}$.

Let $\sigma^2 = \frac{2d\log(2/\delta)}{\varepsilon^2 n^2}$. Note that since the Gaussian noise has mean 0 and is independent of $\overline{X} - \mu$, we have

$$\begin{aligned}
\mathbb{E}\left(\|M(X_{1\cdots n}) - \mu\|_2^2\right) &= \mathbb{E}\left(\|\overline{X} - \mu\|_2^2\right) + \mathbb{E}\left(\|M(X_{1\cdots n}) - \overline{X}\|_2^2\right) \\
&\leq \frac{d}{n} + \mathbb{E}\left(\|M(X_{1\cdots n}) - \overline{X}\|_2^2\right) \\
&= \frac{d}{n} + \mathbb{E}\left(\|\mathcal{N}(0, \sigma^2 \mathbb{I}_{d \times d})\|_2^2\right) \\
&= \frac{d}{n} + \sigma^2 d \\
&= \frac{d}{n} + \frac{2d^2 \log(2/\delta)}{\varepsilon^2 n^2}.
\end{aligned}$$

□

3.2 Minimax Lower Bounds via Tracing

Theorem 3.3. *For every $n, d \in \mathbb{N}$, $\varepsilon > 0$, and $\delta < 1/96n$, if $\mathcal{P}$ is the class of all product distributions on $\{\pm 1\}^d$, then for some constant $C > 0$,*

$$\min_{M \in \mathcal{M}_{\varepsilon,\delta}} \max_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X_{1\cdots n} \sim P, M} \left(\|M(X_{1\cdots n}) - \mu\|_2^2 \right) = \Omega\left(\min\left\{ \frac{d^2}{\varepsilon^2 n^2}, d \right\} \right).$$

Note that it is trivial to achieve error d for any distribution using the mechanism $M(X_{1\cdots n}) \equiv 0$, so the result says that the error must be $\Omega(d^2/\varepsilon^2 n^2)$ whenever this error is significantly smaller than the trivial error of d.

3.2.1 Tracing Attacks

Before giving the formal proof, we will try to give some intuition for the high-level proof strategy. The proof can be viewed as constructing a *tracing attack* [DSSU17] (sometimes called a *membership inference attack*) of the following form. There is an attacker who has the data of some individual Y chosen in one of the two ways: either Y is a random element of the sample X, or Y is an independent random sample from the population P. The attacker is given access to the true distribution P and the outcome of the mechanism $M(X)$, and wants to determine which of the two is the case. If the attacker can succeed, then M cannot be differentially private. To understand why this is the case, if Y is a member of the dataset, then the attacker should say Y is in the dataset, but if we consider the adjacent dataset X' where we replace Y with some independent sample from P, then the attacker will now say Y is independent of the dataset. Thus, $M(X)$ and $M(X')$ cannot be close in the sense required by differential privacy.

Thus, the proof works by constructing a test statistic $Z = Z(M(X), Y, P)$, that the attacker can use to distinguish the two possibilities for Y. In particular, we show that there is a distribution over populations P such that $\mathbb{E}(Z)$ is small when Y is independent of X, but for *every* sufficiently accurate mechanism M, $\mathbb{E}(Z)$ is large when Y is a random element of X.

3.2.2 Proof of Theorem 3.3

We start by constructing a "hard distribution" over the family of product distributions $\mathcal{P}$. Let $\mu = (\mu^1, \ldots, \mu^d) \in [-1, 1]^d$ consist of d independent draws from the uniform distribution on $[-1, 1]$ and let P_μ be the product distribution over $\{\pm 1\}^d$ with mean μ. Let $X_1, \ldots, X_n \sim P_\mu$ and $X = (X_1, \ldots, X_n)$.

Let $M \colon \{\pm 1\}^{n \times d} \to [\pm 1]^d$ be any (ε, δ)-differentially private mechanism and let

$$\alpha^2 = \mathop{\mathbb{E}}_{\mu, X, M} \left(\|M(X) - \mu\|_2^2 \right) \tag{5}$$

be its expected loss. We will prove the desired lower bound on α^2.

For every element i, we define the random variables

$$Z_i = Z_i(M(X), X_i, \mu) = \langle M(X) - \mu, X_i - \mu \rangle \tag{6}$$

$$Z_i' = Z_i'(M(X_{\sim i}), X_i, \mu) = \langle M(X_{\sim i}) - \mu, X_i - \mu \rangle, \tag{7}$$

where $X_{\sim i}$ denotes $(X_1, \dots, X_i', \dots, X_n)$ where X_i' is an independent sample from P_μ. Our goal will be to show that, privacy and accuracy imply both upper and lower bounds on $\mathbb{E}(\sum_i Z_i)$ that depend on α, and thereby obtain a bound on α^2.

The first claim says that, when X_i is *not* in the sample, then the likelihood random variable has mean 0 and variance controlled by the expected ℓ_2^2 error of the mechanism.

Claim 3.4. *For every i, $\mathbb{E}(Z_i') = 0$, $\mathrm{Var}(Z_i') \leq 4\alpha^2$, and $\|Z_i'\|_\infty \leq 4d$.*

Proof of Claim 3.4. Conditioned on any value of μ, $M(X_{\sim i})$ is independent from X_i. Moreover, $\mathbb{E}(X_i - \mu) = 0$, so we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mu,X,M}(\langle M(X_{\sim i}) - \mu, X_i - \mu\rangle) &= \mathop{\mathbb{E}}_{\mu}\left(\mathop{\mathbb{E}}_{X,M}(\langle M(X_{\sim i}) - \mu, X_i - \mu\rangle)\right) \\
&= \mathop{\mathbb{E}}_{\mu}\left(\left\langle \mathop{\mathbb{E}}_{X,M}(M(X_{\sim i}) - \mu), \mathop{\mathbb{E}}_{X,M}(X_i - \mu)\right\rangle\right) \\
&= \mathop{\mathbb{E}}_{\mu}\left(\left\langle \mathop{\mathbb{E}}_{X,M}(M(X_{\sim i}) - \mu), 0\right\rangle\right) \\
&= 0.
\end{aligned}$$

For the second part of the claim, since $(X_i - \mu)^2 \leq 4$, we have $\mathrm{Var}(Z_i') \leq 4 \cdot \mathbb{E}\big(\|M(X) - \mu\|_2^2\big) = 4\alpha^2$. The final part of the claim follows from the fact that every entry of $M(X_{\sim i}) - \mu$ and $X_i - \mu$ is bounded by 2 in absolute value, and Z_i' is a sum of d such entries, so its absolute value is always at most $4d$. □

The next claim says that, because M is differentially private, Z_i has similar expectation to Z_i', and thus its expectation is also small.

Claim 3.5. $\mathbb{E}(\sum_{i=1}^n Z_i) \leq 4n\alpha\varepsilon + 8n\delta d$.

Proof. The proof is a direct calculation using the following inequality, whose proof is relatively simple using the definition of differential privacy:

$$\mathbb{E}(Z_i) \leq \mathbb{E}\big(Z_i'\big) + 2\varepsilon\sqrt{\mathrm{Var}(Z_i')} + 2\delta\|Z_i'\|_\infty. \tag{8}$$

Given the inequality and Claim 3.4, we have

$$\mathbb{E}(Z_i) \leq 0 + (2\varepsilon)(2\alpha) + (2\delta)(2d) = 4\varepsilon\alpha + 8\delta d.$$

The claim now follows by summing over all i. □

The final claim says that, because M is accurate, the expected sum of the random variables Z_i is large.

Claim 3.6. $\mathbb{E}(\sum_{i=1}^n Z_i) \geq \frac{d}{3} - \alpha^2$.

The proof relies on the following key lemma, whose proof we omit.

Lemma 3.7 (Fingerprinting Lemma [BSU17])**.** *If $\mu \in [\pm 1]$ is sampled uniformly, $X_1, \dots, X_n \in \{\pm 1\}^n$ are sampled independently with mean μ, and $f\colon \{\pm 1\}^n \to [\pm 1]$ is any function, then*

$$\mathop{\mathbb{E}}_{\mu,X}\left((f(X) - \mu) \cdot \sum_{i=1}^n (X_i - \mu)\right) \geq \frac{1}{3} - \mathop{\mathbb{E}}_{\mu,X}\big((f(X) - \mu)^2\big).$$

The lemma is somewhat technical, but for intuition, consider the case where $f(X) = \frac{1}{n}\sum_i X_i$ is the empirical mean. In this case we have

$$\mathop{\mathbb{E}}_{\mu,X}\left((f(X)-\mu)\cdot\sum_{i=1}^{n}(X_i-\mu)\right) = \mathop{\mathbb{E}}_{\mu}\left(\frac{1}{n}\sum_i \mathop{\mathbb{E}}_{X}\big((X_i-\mu)^2\big)\right) = \mathop{\mathbb{E}}_{\mu}(\mathrm{Var}(X_i)) = \frac{1}{3}.$$

The lemma says that, when μ is sampled this way, then any modification of f that reduces the correlation between $f(X)$ and $\sum_i X_i$ will increase the mean-squared-error of f proportionally.

Proof of Claim 3.6. We can apply the lemma to each coordinate of the estimate $M(X)$.

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n} Z_i\right) &= \mathbb{E}\left(\sum_{i=1}^{n}\langle M(X)-\mu, X_i-\mu\rangle\right) \\
&= \sum_{j=1}^{d}\mathbb{E}\left((M^j(X)-\mu^j)\cdot\sum_{i=1}^{n}(X_i^j-\mu^j)\right) \\
&\geq \sum_{j=1}^{d}\left(\frac{1}{3}-\mathbb{E}\big((M^j(X)-\mu^j)^2\big)\right) && \text{(Lemma 3.7)} \\
&= \frac{d}{3}-\mathbb{E}\big(\|M(X)-\mu\|_2^2\big) = \frac{d}{3}-\alpha^2.
\end{aligned}$$

□

Combining Claims 3.5 and 3.6 gives

$$\frac{d}{3}-\alpha^2 \leq 4n\alpha\varepsilon + 8n\delta d. \tag{9}$$

Now, if $\alpha^2 \geq \frac{d}{6}$ then we're done, so we'll assume that $\alpha^2 \leq \frac{d}{6}$. Further, by our assumption on the value of δ, $8n\delta d \leq \frac{d}{12}$. In this case we can rearrange terms and square both sides to obtain

$$\alpha^2 \geq \frac{1}{16\varepsilon^2 n^2}\left(\frac{d}{3}-\alpha^2-8n\delta d\right)^2 \geq \frac{1}{16\varepsilon^2 n^2}\left(\frac{d}{12}\right)^2 = \frac{d^2}{2304\varepsilon^2 n^2}. \tag{10}$$

Combining the two cases for α^2 gives $\alpha^2 \geq \min\{\frac{d}{6}, \frac{d^2}{2304\varepsilon^2 n^2}\}$, as desired.

4 CDF Estimation for Discrete, Univariate Distributions

Suppose we have a distribution P over the ordered, discrete domain $\{1,\ldots,D\}$ and let $\mathcal{P}$ be the family of all such distributions. The CDF of the distribution is the function $\Phi_P : \{1,\ldots,D\} \to [0,1]$ given by

$$\Phi_P(j) = \mathbb{P}(P \leq j). \tag{11}$$

A natural measure of distance between CDFs is the ℓ_∞ distance, as this is the sort of convergence guarantee that the empirical CDF satisfies. That is, in the non-private setting, the empirical CDF will achieve the minimax rate, which it known by [DKW56, Mas90] to be

$$\max_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{X_{1\cdots n}\sim P}(\|\Phi_X-\Phi_P\|_\infty) = O\left(\sqrt{\frac{1}{n}}\right). \tag{12}$$

4.1 Private CDF Estimation

Theorem 4.1. *For every $n \in \mathbb{N}$ and every $\varepsilon, \delta > 0$, there exists an (ε, δ)-differentially private mechanism M such that*

$$\max_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X_{1 \cdots n} \sim P}(\|M(X_{1 \cdots n}) - \Phi_P\|_\infty) = O\left(\sqrt{\frac{1}{n}} + \frac{\log^{3/2}(D) \log^{1/2}(1/\delta)}{\varepsilon n}\right). \tag{13}$$

Proof. Assume without loss of generality that $D = 2^d$ for an integer $d \geq 1$. Let $X_{1 \cdots n} \sim P$ be a sample. By the triangle inequality, we have

$$\begin{aligned}\mathop{\mathbb{E}}_{X_{1 \cdots n} \sim P}(\|M(X_{1 \cdots n}) - \Phi_P\|_\infty) &\leq \mathop{\mathbb{E}}_{X_{1 \cdots n} \sim P}(\|\Phi_X - \Phi_P\|_\infty + \|M(X_{1 \cdots n}) - \Phi_X\|_\infty) \\ &\leq O(\sqrt{1/n}) + \mathop{\mathbb{E}}_{X_{1 \cdots n} \sim P}(\|M(X_{1 \cdots n}) - \Phi_X\|_\infty),\end{aligned}$$

so we will focus on constructing M to approximate Φ_X.

For any $\ell = 0, \ldots, d-1$ and $j = 1, \ldots, 2^{d-\ell}$, consider the statistics

$$f_{\ell,j}(X_{1 \cdots n}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{(j-1)2^\ell + 1 \leq X_i \leq j2^\ell\}. \tag{14}$$

Let $f : \{1, \ldots, D\}^n \to [0,1]^{2D-2}$ be the function whose output consists of all $2D - 2$ such counts. To decipher this notation, for a given ℓ, the counts $f_{\ell,\cdot}$ form a histogram of $X_{1 \cdots n}$ using consecutive bins of width 2^ℓ, and we consider the $\log(D)$ histograms of geometrically increasing width $1, 2, 4, \ldots, D$.

First, we claim that the function f has low sensitivity—for adjacent samples X and X',

$$\|f(X) - f(X')\|_2^2 \leq \frac{2\log(D)}{n^2}. \tag{15}$$

Thus, we can use the Gaussian mechanism:

$$M'(X_{1 \cdots n}) = f(X_{1 \cdots n}) + \mathcal{N}\left(0, \frac{2\log(D)\log(1/\delta)}{\varepsilon^2 n^2} \cdot \mathbb{I}_{2D \times 2D}\right). \tag{16}$$

As we will argue, there exists a matrix $A \in \mathbb{R}^{2D \times 2D}$ such that $\Phi_X = A \cdot f(X_{1 \cdots n})$. We will let $M(X_{1 \cdots n}) = A \cdot M'(X_{1 \cdots n})$. Since differential privacy is closed under post-processing, M inherits the privacy of M'.

We will now show how to construct the matrix A and analyze the error of M. For any $j = 1, \ldots, D$, we can form the interval $\{1, \ldots, j\}$ as the union of at most $\log D$ disjoint intervals of the form we've computed, and therefore we can obtain $\Phi_X(j)$ as the sum of at most $\log D$ of the entries of $f(X)$. For example, if $j = 5$ then we can write

$$\{1, \ldots, 7\} = \{1, \ldots, 4\} \cup \{5, 6\} \cup \{7\} \tag{17}$$

and

$$\Phi_X(5) = f_{2,1} + f_{1,3} + f_{0,7}. \tag{18}$$

See Figure 1 for a visual representation of the decomposition. Thus we can construct the matrix A using this information.

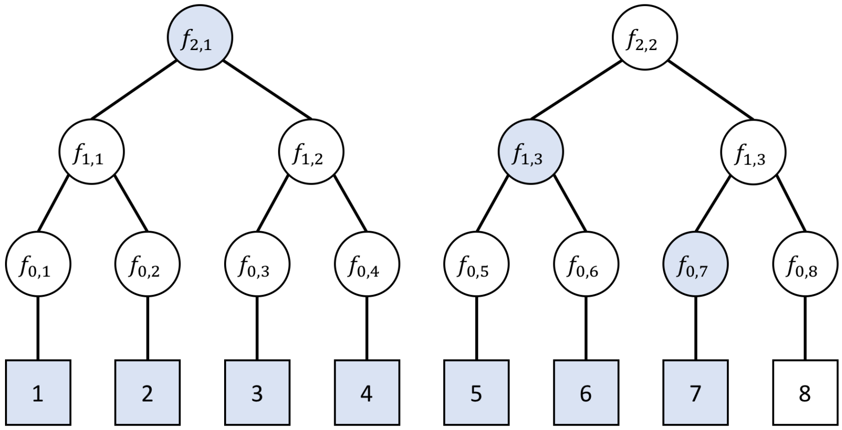

Figure 1: A diagram showing the hierarchical decomposition of the domain $\{1, \dots, 8\}$ using 14 intervals. The highlighted squares represent the interval $\{1, \dots, 7\}$ and the highlighted circles show the decomposition of this interval into a union of 3 intervals in the tree.

Note that each entry of $Af(X)$ is the sum of at most $\log(D)$ entries of $f(X)$. Thus, if we use the output of $M'(X_{1\cdots n})$ in place of $f(X_{1\cdots n})$, for every j we obtain

$$\Phi_X(j) + \mathcal{N}(0, \sigma^2) \quad \text{for} \quad \sigma^2 = \frac{2\log^2(D)\log(1/\delta)}{\varepsilon^2 n^2}. \tag{19}$$

Applying standard bounds on the expected supremum of a Gaussian process, we have

$$\mathbb{E}(\|M(X_{1\cdots n}) - \Phi_X\|_\infty) = O(\sigma\sqrt{\log D}) = O\left(\frac{\log^{3/2}(D)\log^{1/2}(1/\delta)}{\varepsilon n}\right). \tag{20}$$

□

4.2 Why restrict the domain?

A drawback of the estimator we constructed is that it only applies to distributions of finite support $\{1, 2, \dots, D\}$, albeit with a relatively mild dependence on the support size. If privacy isn't a concern, then no such restriction is necessary, as the bound (12) applies equally well to any distribution over $\mathbb{R}$. Can we construct a differentially private estimator for distributions with infinite support?

Perhaps surprisingly, the answer to this question is no! Any differentially private estimator for the CDF of the distribution has to have a rate that depends on the support size, and cannot give non-trivial rates for distributions with infinite support.

Theorem 4.2 ([BNSV15])**.** *If $\mathcal{P}$ consists of all distributions on $\{1, \dots, D\}$, then*

$$\min_{M \in \mathcal{M}_{1,\frac{1}{n}}} \max_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X_{1\cdots n}\sim P}(\|M(X_{1\cdots n}) - \Phi_P\|_\infty) = \Omega\left(\frac{\log^* D}{n}\right).^{3} \tag{21}$$

[3] The notation $\log^* D$ refers to the iterated logarithm.

We emphasize that this theorem shouldn't meet with too much alarm, as $\log^* D$ grows remarkably slowly with D. There are differentially private CDF estimators that achieve very mild dependence on D [BNS13, BNSV15], including one nearly matching the lower bound in Theorem 4.2. Moreover, if we want to estimate a distribution over $\mathbb{R}$, and are willing to make some mild regularity conditions on the distribution, then we can approximate it by a distribution with finite support and only increase the rate slightly. However, what Theorem 4.2 shows is that there is no "one-size-fits-all" solution to private CDF estimation that achieves similar guarantees to the empirical CDF. That is, the right algorithm has to be tailored somewhat to the application and the assumptions we can make about the distribution.

5 More Private Statistics

Of course, the story doesn't end here! There's a whole wide world of differentially private statistics beyond what we've mentioned already. We proceed to survey just a few other directions of study in private statistics.

5.1 Parameter and Distribution Estimation

A number of the early works in differential privacy give methods for differentially private statistical estimation for i.i.d. data. The earliest works [DN03, DN04, BDMN05, DMNS06], which introduced the Gaussian mechanism, among other foundational results, can be thought of as methods for estimating the mean of a distribution over the hypercube $\{0,1\}^d$ in the ℓ_∞ norm. Tight lower bounds for this problem follow from the tracing attacks introduced in [BUV14, SU17a, DSS$^+$15, BSU17, SU17b]. A very recent work of Acharya, Sun, and Zhang [ASZ20] adapts classical tools for proving estimation and testing lower bounds (lemmata of Assouad, Fano, and Le Cam) to the differentially private setting. Steinke and Ullman [SU17b] give tight minimax lower bounds for the weaker guarantee of selecting the largest coordinates of the mean, which were refined by Cai, Wang, and Zhang [CWZ19] to give lower bounds for sparse mean-estimation problems.

Nissim, Raskhodnikova, and Smith introduced the highly general sample-and-aggregate paradigm, which they apply to several learning problems (e.g., learning mixtures of Gaussians) [NRS07]. Later, Smith [Smi11] showed that this paradigm can be used to transform any estimator for any asymptotically normal, univariate statistic over a bounded data domain into a differentially private one with the same asymptotic convergence rate.

Subsequent work has focused on both relaxing the assumptions in [Smi11], particularly boundedness, and on giving finite-sample guarantees. Karwa and Vadhan investigated the problem of Gaussian mean estimation, proving the first near-optimal bounds for this setting [KV18]. In particular, exploiting concentration properties of Gaussian data allows us to achieve non-trivial results even with unbounded data, which is impossible in general. Following this, Kamath, Li, Singhal, and Ullman moved to the multivariate setting, investigating the estimation of Gaussians and binary product distributions in total variation distance [KLSU19]. In certain cases (i.e., Gaussians with identity covariance), this is equivalent to mean estimation in ℓ_2-distance, though not always. For example, for binary product distribution, one must estimate the mean in a type of χ^2-distance instead. The perspective of distribution estimation rather than parameter estimation can be valuable. Bun, Kamath, Steinke, and Wu [BKSW19] develop a primitive for private hypothesis selection, which they apply to learn any coverable class of distributions under pure differential privacy. Through the lens of distribution estimation, their work implies an upper bound for mean estimation of binary product distributions that bypasses lower bounds for the same problem in the empirical setting. In addition to work on mean estimation in the sub-Gaussian setting, such as the results discussed earlier, mean estimation has also been studied under weaker moment conditions [BS19, KSU20]. Beyond these settings, there has also been study of estimation of discrete multinomials, including estimation in Kolmogorov distance [BNSV15] and in total variation distance for structured distributions [DHS15], and parameter estimation of Markov Random Fields [ZKKW20].

A different approach to constructing differentially private estimators is based on robust statistics. This approah begins with the influential work of Dwork and Lei [DL09], which introduced the propose-test-release framework, and applied to estimating robust statistics such as the median and interquartile range. While the definitions in robust statistics and differential privacy are semantically similar, formal connections between the two remain relatively scant, which suggests a productive area for future study.

5.2 Hypothesis Testing

An influential work of Homer et al. [HSR+08] demonstrated the vulnerability of classical statistics in a genomic setting, showing that certain χ^2-statistics on many different variables could allow an attacker to determine the presence of an individual in a genome-wide association study (GWAS). Motivated by these concerns, an early line of work from the statistics community focused on addressing these issues [VS09, USF13, YFSU14].

More recently, work on private hypothesis testing can be divided roughly into two lines. The first focuses on the minimax sample complexity, in a line initiated by Cai, Daskalakis, and Kamath [CDK17], who give an algorithm for privately testing goodness-of-fit (more precisely, a statistician might refer to this problem as one-sample testing of multinomial data). A number of subsequent works have essentially settled the complexity of this problem [ASZ18, ADR18], giving tight upper and lower bounds. Other papers in this line study related problems, including the two-sample version of the problem, independence testing, and goodness-of-fit testing for multivariate product distributions [ASZ18, ADR18, ADKR19, CKM+19b]. A related paper studies the minimax sample complexity of property *estimation*, rather than testing of discrete distributions, including support size and entropy [AKSZ18]. Other recent works in this vein focus on testing of simple hypotheses [CKM+18, CKM+19a]. In particular [CKM+19a] proves an analogue of the Neyman-Pearson Lemma for differentially private testing of simple hypotheses. A paper of Awan and Slavkovic [AS18] gives a universally optimal test when the domain size is two, however Brenner and Nissim [BN14] shows that such universally optimal tests cannot exist when the domain has more than two elements. A related problem in this space is private change-point detection [CKM+18, CKM+19a, CKLZ19] – in this setting, we are given a time series of datapoints which are sampled from a distribution, which at some point, changes to a different distribution. The goal is to (privately) determine when this point occurs.

Complementary to minimax hypothesis testing, a line of work [WLK15, GLRV16, KR17, KSF17, CBRG18, SGHG+19, CKS+19] designs differentially private versions of popular test statistics for testing goodness-of-fit, closeness, and independence, as well as private ANOVA, focusing on the performance at small sample sizes. Work by Wang et al. [WKLK18] focuses on generating statistical approximating distributions for differentially private statistics, which they apply to hypothesis testing problems.

5.3 Differential Privacy on Graphs

There is a significant amount of work on differentially private analysis of graphs. We remark that these algorithms can satisfy either edge or node differential privacy. The former (easier) guarantee defines a neighboring graph to be one obtained by adding or removing a single edge, while in the latter (harder) setting, a neighboring graph is one that can be obtained by modifying the set of edges connected to a single node. The main challenge in this area is that most graph statistics can have high sensitivity in the worst-case.

The initial works in this area focused on the empirical setting, and goals range from counting subgraphs [KRSY11, BBDS13, KNRS13, CZ13, RS16] to outputting a privatized graph which approximates the original [GRU12, BBDS12, Upa13, AU19, EKKL20]. In contrast to the setting discussed in most of this series, it seems that there are larger qualitative differences between the study of empirical and population statistics due to the fact that many graph statistics have high worst-case sensitivity, but may have smaller sensitivity on typical graphs from many natural models.

In the population statistics setting, recent work has focused on parameter estimation of the underlying random graph model. So far this work has given estimators for the β-model [KS16] and graphons [BCS15, BCSZ18]. Graphons are a generalization of the stochastic block model, which is, in turn, a generalization of the Erdős-Rényi model. Interestingly, the methods of Lipschitz-extensions introduced in the empirical setting by [BBDS13, KNRS13] are the main tool used in the statistical setting as well. While the first works on private graphon estimation were not computationally efficient, a recent focus has been on obviating these issues for certain important cases, such as the Erdős-Rényi setting [SU19].

Acknowledgments

Thanks to Adam Smith, who helped kick off this project, and Clément Canonne, Aaron Roth, and Thomas Steinke for helpful comments.

References

[ADKR19] Maryam Aliakbarpour, Ilias Diakonikolas, Daniel M. Kane, and Ronitt Rubinfeld. Private testing of distributions via sample permutations. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 10877–10888. Curran Associates, Inc., 2019.

[ADR18] Maryam Aliakbarpour, Ilias Diakonikolas, and Ronitt Rubinfeld. Differentially private identity and closeness testing of discrete distributions. In *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pages 169–178. JMLR, Inc., 2018.

[AKSZ18] Jayadev Acharya, Gautam Kamath, Ziteng Sun, and Huanyu Zhang. Inspectre: Privately estimating the unseen. In *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pages 30–39. JMLR, Inc., 2018.

[AS18] Jordan Awan and Aleksandra Slavković. Differentially private uniformly most powerful tests for binomial data. In *Advances in Neural Information Processing Systems 31*, NeurIPS '18, pages 4208–4218. Curran Associates, Inc., 2018.

[ASZ18] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private testing of identity and closeness of discrete distributions. In *Advances in Neural Information Processing Systems 31*, NeurIPS '18, pages 6878–6891. Curran Associates, Inc., 2018.

[ASZ20] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private Assouad, Fano, and Le Cam. *arXiv preprint arXiv:2004.06830*, 2020.

[AU19] Raman Arora and Jalaj Upadhyay. On differentially private graph sparsification and applications. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 13378–13389. Curran Associates, Inc., 2019.

[BBDS12] Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. The Johnson-Lindenstrauss transform itself preserves differential privacy. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science*, FOCS '12, pages 410–419, Washington, DC, USA, 2012. IEEE Computer Society.

[BBDS13] Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. Differentially private data analysis of social networks via restricted sensitivity. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, pages 87–96, New York, NY, USA, 2013. ACM.

[BCS15] Christian Borgs, Jennifer Chayes, and Adam Smith. Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 1369–1377. Curran Associates, Inc., 2015.

[BCSZ18] Christian Borgs, Jennifer Chayes, Adam Smith, and Ilias Zadik. Revealing network structure, confidentially: Improved rates for node-private graphon estimation. In *Proceedings of the 59th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '18, pages 533–543, Washington, DC, USA, 2018. IEEE Computer Society.

[BDMN05] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: The SuLQ framework. In *Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '05, pages 128–138, New York, NY, USA, 2005. ACM.

[BDRS18] Mark Bun, Cynthia Dwork, Guy N. Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual ACM Symposium on the Theory of Computing*, STOC '18, pages 74–86, New York, NY, USA, 2018. ACM.

[BKSW19] Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 156–167. Curran Associates, Inc., 2019.

[BN14] Hai Brenner and Kobbi Nissim. Impossibility of differentially private universally optimal mechanisms. *SIAM Journal on Computing*, 43(5):1513–1540, 2014.

[BNS13] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, RANDOM-APPROX '13, pages 363–378. Springer, 2013.

[BNSV15] Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 634–649, Washington, DC, USA, 2015. IEEE Computer Society.

[BS16] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Proceedings of the 14th Conference on Theory of Cryptography*, TCC '16-B, pages 635–658, Berlin, Heidelberg, 2016. Springer.

[BS19] Mark Bun and Thomas Steinke. Average-case averages: Private algorithms for smooth sensitivity and mean estimation. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 181–191. Curran Associates, Inc., 2019.

[BSU17] Mark Bun, Thomas Steinke, and Jonathan Ullman. Make up your mind: The price of online queries in differential privacy. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, pages 1306–1325, Philadelphia, PA, USA, 2017. SIAM.

[BUV14] Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the 46th Annual ACM Symposium on the Theory of Computing*, STOC '14, pages 1–10, New York, NY, USA, 2014. ACM.

[CBRG18] Zachary Campbell, Andrew Bray, Anna Ritz, and Adam Groce. Differentially private ANOVA testing. In *Proceedings of the 2018 International Conference on Data Intelligence and Security*, ICDIS '18, pages 281–285, Washington, DC, USA, 2018. IEEE Computer Society.

[CDK17] Bryan Cai, Constantinos Daskalakis, and Gautam Kamath. Priv'it: Private and sample efficient identity testing. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 635–644. JMLR, Inc., 2017.

[CKLZ19] Rachel Cummings, Sara Krehbiel, Yuliia Lut, and Wanrong Zhang. Privately detecting changes in unknown distributions. *arXiv preprint arXiv:1910.01327*, 2019.

[CKM+18] Rachel Cummings, Sara Krehbiel, Yajun Mei, Rui Tuo, and Wanrong Zhang. Differentially private change-point detection. In *Advances in Neural Information Processing Systems 31*, NeurIPS '18. Curran Associates, Inc., 2018.

[CKM+19a] Clément L. Canonne, Gautam Kamath, Audra McMillan, Adam Smith, and Jonathan Ullman. The structure of optimal private tests for simple hypotheses. In *Proceedings of the 51st Annual ACM Symposium on the Theory of Computing*, STOC '19, New York, NY, USA, 2019. ACM.

[CKM+19b] Clément L. Canonne, Gautam Kamath, Audra McMillan, Jonathan Ullman, and Lydia Zakynthinou. Private identity testing for high-dimensional distributions. *arXiv preprint arXiv:1905.11947*, 2019.

[CKS+19] Simon Couch, Zeki Kazan, Kaiyan Shi, Andrew Bray, and Adam Groce. Differentially private nonparametric hypothesis testing. In *Proceedings of the 2019 ACM Conference on Computer and Communications Security*, CCS '19, New York, NY, USA, 2019. ACM.

[CSS11] T-H Hubert Chan, Elaine Shi, and Dawn Song. Private and continual release of statistics. *ACM Transactions on Information and System Security (TISSEC)*, 14(3):26, 2011.

[CWZ19] T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *arXiv preprint arXiv:1902.04495*, 2019.

[CZ13] Shixi Chen and Shuigeng Zhou. Recursive mechanism: Towards node differential privacy and unrestricted joins. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 653–664, New York, NY, USA, 2013. ACM.

[DHS15] Ilias Diakonikolas, Moritz Hardt, and Ludwig Schmidt. Differentially private learning of structured discrete distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 2566–2574. Curran Associates, Inc., 2015.

[DKW56] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 09 1956.

[DL09] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st Annual ACM Symposium on the Theory of Computing*, STOC '09, pages 371–380, New York, NY, USA, 2009. ACM.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer.

[DN03] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, pages 202–210, New York, NY, USA, 2003. ACM.

[DN04] Cynthia Dwork and Kobbi Nissim. Privacy-preserving datamining on vertically partitioned databases. In *Annual International Cryptology Conference*, pages 528–544. Springer, 2004.

[DNPR10] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. Differential privacy under continual observation. In *Symposium on Theory of Computing (STOC)*, pages 715–724. ACM, 2010.

[DR16] Cynthia Dwork and Guy N. Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[DRS19] Jinshuo Dong, Aaron Roth, and Weijie J. Su. Gaussian differential privacy. *arXiv preprint arXiv:1905.02383*, 2019.

[DSS$^+$15] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 650–669, Washington, DC, USA, 2015. IEEE Computer Society.

[DSSU17] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application*, 4:61–84, 2017.

[EKKL20] Marek Eliáš, Michael Kapralov, Janardhan Kulkarni, and Yin Tat Lee. Differentially private release of synthetic graphs. In *Proceedings of the 31st Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '20, pages 560–578, Philadelphia, PA, USA, 2020. SIAM.

[FRY10] Stephen E. Fienberg, Alessandro Rinaldo, and Xiaolin Yang. Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *Proceedings of the International Conference on Privacy in Statistical Databases*, PSD '10, Corfu, Greece, 2010. Springer.

[GLRV16] Marco Gaboardi, Hyun-Woo Lim, Ryan M. Rogers, and Salil P. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *Proceedings of the 33rd International Conference on Machine Learning*, ICML '16, pages 1395–1403. JMLR, Inc., 2016.

[GRU12] Anupam Gupta, Aaron Roth, and Jonathan Ullman. Iterative constructions and private data release. In *Proceedings of the 9th Conference on Theory of Cryptography*, TCC '12, pages 339–356, Berlin, Heidelberg, 2012. Springer.

[HSR$^+$08] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8):1–9, 2008.

[KLSU19] Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Proceedings of the 32nd Annual Conference on Learning Theory*, COLT '19, pages 1853–1902, 2019.

[KNRS13] Shiva Prasad Kasiviswanathan, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Analyzing graphs with node differential privacy. In *Proceedings of the 10th Conference on Theory of Cryptography*, TCC '13, pages 457–476, Berlin, Heidelberg, 2013. Springer.

[KR17] Daniel Kifer and Ryan M. Rogers. A new class of private chi-square tests. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, AISTATS '17, pages 991–1000. JMLR, Inc., 2017.

[KRSY11] Vishesh Karwa, Sofya Raskhodnikova, Adam Smith, and Grigory Yaroslavtsev. Private analysis of graph structure. *Proceedings of the VLDB Endowment*, 4(11):1146–1157, 2011.

[KS16] Vishesh Karwa and Aleksandra Slavković. Inference using noisy degrees: Differentially private β-model and synthetic graphs. *The Annals of Statistics*, 44(1):87–112, 2016.

[KSF17] Kazuya Kakizaki, Jun Sakuma, and Kazuto Fukuchi. Differentially private chi-squared test by unit circle mechanism. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 1761–1770. JMLR, Inc., 2017.

[KSU20] Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. *arXiv preprint arXiv:2002.09464*, 2020.

[KV18] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pages 44:1–44:9, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[Mas90] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 07 1990.

[Mir17] Ilya Mironov. Rényi differential privacy. In *Proceedings of the 30th IEEE Computer Security Foundations Symposium*, CSF '17, pages 263–275, Washington, DC, USA, 2017. IEEE Computer Society.

[NRS07] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the 39th Annual ACM Symposium on the Theory of Computing*, STOC '07, pages 75–84, New York, NY, USA, 2007. ACM.

[RS16] Sofya Raskhodnikova and Adam D. Smith. Lipschitz extensions for node-private graph statistics and the generalized exponential mechanism. In *Proceedings of* *the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 495–504, Washington, DC, USA, 2016. IEEE Computer Society.

[SGHG$^+$19] Marika Swanberg, Ira Globus-Harris, Iris Griffith, Anna Ritz, Adam Groce, and Andrew Bray. Improved differentially private analysis of variance. *Proceedings on Privacy Enhancing Technologies*, 2019(3), 2019.

[Smi11] Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the 43rd Annual ACM Symposium on the Theory of Computing*, STOC '11, pages 813–822, New York, NY, USA, 2011. ACM.

[SU17a] Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 7(2), 2017.

[SU17b] Thomas Steinke and Jonathan Ullman. Tight lower bounds for differentially private selection. In *58th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '17, pages 552–563, Berkeley, CA, 2017.

[SU19] Adam Sealfon and Jonathan Ullman. Efficiently estimating Erdos-Renyi graphs with node differential privacy. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 3765–3775. Curran Associates, Inc., 2019.

[Upa13] Jalaj Upadhyay. Random projections, graph sparsification, and differential privacy. In *Proceedings of the 19th Annual International Conference on the Theory and Application of Cryptology and Information Security*, ASIACRYPT '13, pages 276–295, Berlin, Heidelberg, 2013. Springer.

[USF13] Caroline Uhler, Aleksandra Slavković, and Stephen E. Fienberg. Privacy-preserving data sharing for genome-wide association studies. *The Journal of Privacy and Confidentiality*, 5(1):137–166, 2013.

[VS09] Duy Vu and Aleksandra Slavković. Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, pages 138–143. IEEE, 2009.

[WKLK18] Yue Wang, Daniel Kifer, Jaewoo Lee, and Vishesh Karwa. Statistical approximating distributions under differential privacy. *The Journal of Privacy and Confidentiality*, 8(1):1–33, 2018.

[WLK15] Yue Wang, Jaewoo Lee, and Daniel Kifer. Revisiting differentially private hypothesis tests for categorical data. *arXiv preprint arXiv:1511.03376*, 2015.

[YFSU14] Fei Yu, Stephen E. Fienberg, Aleksandra B. Slavković, and Caroline Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50:133–141, 2014.

[ZKKW20] Huanyu Zhang, Gautam Kamath, Janardhan Kulkarni, and Zhiwei Steven Wu. Privately learning Markov random fields. *arXiv preprint arXiv:2002.09463*, 2020.