Privacy-preserving Statistical Estimation with Optimal Convergence Rates*

[Extended Abstract]

Adam Smith†
Department of Computer Science and Engineering, Pennsylvania State University, University Park
asmith@psu.edu

ABSTRACT

Consider an analyst who wants to release aggregate statistics about a data set containing sensitive information. Using *differentially private* algorithms guarantees that the released statistics reveal very little about any particular record in the data set. In this paper we study the asymptotic properties of differentially private algorithms for statistical inference.

We show that for a large class of statistical estimators T and input distributions P, there is a differentially private estimator A_T with the same asymptotic distribution as T. That is, the random variables $A_T(X)$ and $T(X)$ converge in distribution when X consists of an i.i.d. sample from P of increasing size. This implies that $A_T(X)$ is essentially as good as the original statistic $T(X)$ for statistical inference, for sufficiently large samples. Our technique applies to (almost) any pair T, P such that T is asymptotically normal on i.i.d. samples from P—in particular, to parametric maximum likelihood estimators and estimators for logistic and linear regression under standard regularity conditions.

A consequence of our techniques is the existence of low-space streaming algorithms whose output converges to the same asymptotic distribution as a given estimator T (for the same class of estimators and input distributions as above).

Categories and Subject Descriptors

F.2.0 [**Theory of Computation**]: Analysis Of Algorithms And Problem Complexity—*General*; G.3 [**Mathematics of Computing**]: Probability And Statistics

General Terms

Algorithms,Theory

*A weaker version of the results of this paper appeared in an unpublished work by the same author [Smi08].

†Supported by NSF PECASE Award CCF-0747294 and NSF Award CCF-0729171

STOC'11, June 6–8, 2011, San Jose, California, USA.

Keywords

Differential Privacy, Statistical Inference, Asymptotic Distribution

1. INTRODUCTION

Private data analysis has recently emerged as a fascinating field at the interface of algorithms, statistical learning, complexity theory and geometry. One specific notion of privacy, *differential privacy* [DMNS06, Dwo06] has received significant attention, largely because it is conceptual simple and yet offers meaningful guarantees in the presence of arbitrary external information.

In this paper we consider the following basic problem: given a data set of sensitive information, what kinds of statistical inference and learning can we carry out without the results leaking sensitive information? For example, when can we guarantee that the published analysis of a clinical study's outcomes will not leak details of an individual patient's treatment?

We provide, in some sense, a "free lunch" for data analysts: we show that given a statistical estimator T that returns a relatively low-dimensional real vector, one can design a differentially private algorithm A_T with the following rough guarantee:

> If the input $X_1, X_2, ..., X_n$ is drawn i.i.d. (according to distribution that need not be known to A_T), and if $T(X_1, ..., X_n)$, appropriately rescaled, converges to a normal distribution as n grows, then $A_T(X_1, ..., X_n)$ converges to the *same distribution*.

Differential privacy is parameterized by a number $\varepsilon > 0$ (Definition 1). Our guarantee works even as ε tends to 0; we require only that $\frac{1}{\varepsilon}$ be bounded above by a small polynomial in the sample size n.

Our analysis requires additional conditions on the convergence of T; these are met by essentially all the "asymptotic normality" results from classical statistical theory. Estimators which fit our assumptions include the maximum likelihood estimator for "nice" parametric families of distributions; maximum-likelihood estimators for a host of regression problems, including linear regression and logistic regression; estimators for the parameters of low-dimensional mixture-models with a fixed number of components, and, in general, estimators for which T, viewed as a functional on the space of distributions, is Fréchet differentiable and has

a non-zero derivative at the distribution P from which the X_i are drawn.

In several cases, notably for parametric estimation and regression, the estimators to which our result applies are known to have minimum possible variance among a large class of estimators. In those cases, we give differentially private estimators whose variance is within a $1 + o(1)$ factor of optimal. In that sense, we provide a free lunch for private statistical inference: optimal accuracy (the same as the best nonprivate estimators) and strong guarantees of privacy.

Our algorithms also provide one-pass, space-efficient algorithms that compute an asymptotically similar statistic to a given statistic T. These algorithms use space $\tilde{O}(\sqrt{n})$ when T itself can be computed in quasilinear space. To the best of our knowledge, previously studied estimators require linear space to achieve the same guarantee. For a discussion of previous work on streaming algorithms for i.i.d. and randomly-ordered data, see Chien, Ligett and McGregor [CLM10].

Previous Work.

The general question of the design of differentially private algorithms has received much recent attention, too much to survey here. We focus on work directly relevant to our concerns. Blum et al. [BDMN05] considered a simple mechanism which expresses a learning or estimation algorithm in terms of a sequence of "statistical queries" (in the sense of Kearns [Kea98]) to the data, and adds noise the result of each query. In most cases, it is difficult to see how this reformulation and noise addition will affect the quality of the final answer. For the class of *exponential families* of probability distributions, the sufficient statistics for a family can be directly computed via statistical queries, and so the results of Blum et al. [BDMN05] (and follow-up work by Dwork et al. [DMNS06]) would apply. Unfortunately, many probability models, e.g. mixture models, do not have such simple sufficient statistics.

Kasiviswanathan et al. [KLN$^+$08] gave general inefficient algorithms with PAC learning guarantees for binary classification problems. They show, roughly, that the private and nonprivate sample complexities of PAC learning problems are polynomially related. The algorithms of [KLN$^+$08] run in exponential time, in general, though which makes them difficult to apply. In contrast, our algorithm A_T runs in time essentially proportional to the time needed to compute the original statistic T.

Most relevant to our work is the paper of Dwork and Lei [DL09], which drew inspiration from *robust statistics* to give differentially private estimators for scale, location and regression problems. Their estimators converge at a rate of $n^{-1/2+c}$ (where $c > 0$ is a small constant) to the underlying true parameters they were trying to recover. In contrast, our estimators converge at the optimal rate of $n^{-1/2}$ in distributions similar to the ones they analyzed; our convergence rate even recovers the correct leading constant. Our techniques have a similar flavor to those of [DL09]. In particular, we modify a robust estimator of location, the Winsorized mean, to get an *efficient*[1] differentially private estimator for estimating the mean of a Gaussian distribution. Our analysis sheds some light on the relationship between robustness and differential privacy. Another similarity between this work and [DL09] is the use of the subsample-and-aggregate technique of Nissim, Raskhodnikova and Smith [NRS07] as a basic building block.

Wasserman and Zhou [WZ10] consider differentially private versions of several nonparametric density estimators. In general, the convergence rates they obtain are suboptimal and the algorithms they describe take exponential time to run. However, the problem they tackle is essentially infinite dimensional; it is unclear if the techniques we discuss here can be applied in their setting.

Finally, Chaudhuri, Monteleoni and Sarwate [CMS11] and Rubinstein et al. [RBHT] consider differentially private algorithms tailored to algorithms that minimize a convex loss function. Their convergence guarantees are incomparable to ours, although it seems likely that their techniques outperform ours in the specific settings for which they were designed. In particular, the dependency on the dimension of the problem is probably much better with their techniques (our results require the dimension to be bounded above by a small polynomial in n, whereas their techniques may work for dimensions up to $\sqrt{n}$).

Basic Tools.

We design our differentially private algorithms by composing a number of existing tools from the literature, namely the Laplace mechanism of Dwork, McSherry, Nissim and Smith [DMNS06], the exponential mechanism of McSherry and Talwar [MT07] and the subsample-and-aggregate framework of Nissim, Raskhodnikova and Smith [NRS07]. See Section 2.1 for further discussion.

Preliminaries.

Given a random variable X, we denote its cumulative distribution function F_X. Similarly, given a probability measure P we use F_P for the corresonding c.d.f. We sometimes abuse notation and equate a random variable X with the corresponding probability distribution.

To compare probability distributions, we use two metrics. First, given two distributions P, Q defined on the same underlying space of measurable sets, consider the *statistical difference* (or *total variation distance*)

$$\mathsf{SD}(P, Q) = \sup_{\text{measurable } S} |P(S) - Q(S)| .$$

One often thinks of the sets S as possible tests for distinguishing P from Q. The definition states that no such set can distinguish between (draws from) P and Q with probability better than $\frac{1}{2} + \mathsf{SD}(P, Q)$.

Second, for probability distributions on $\mathbb{R}^d$, we will use the Kolmogorov-Smirnov distance, which can be thought of as a relaxed version of the statistical difference that considers only axis-parallel rectangles as possible distinguishers:

$$\mathsf{KS}(P, Q) = \sup_{\text{rectangles} R} |P(R) - Q(R)| .$$

For one-dimensional distributions, this is $\sup_{a,b\in\mathbb{R}} |P[a, b] - Q[a, b]|$. Note that the KS distance, as defined here, is closely related to the L_∞ norm on the cumulative distribution func-

[1] In statistical theory, an estimator is called *efficient* if it has the best possible convergence rate (among an appropriate class of estimators); that is, if it is efficient in its use of data. In this paper we have tried to carefully distinguish this meaning of "efficient" from the notion of computational efficiency, but we apologize in advance if some ambiguities remain.

tions of P and Q. Namely, $\|F_P - F_Q\|_\infty \leq \mathsf{KS}(P,Q) \leq 2^d \|F_P - F_Q\|_\infty$ where d is the dimension.

Recally that a sequence of distributions $Q_1, Q_2, ...$ on $\mathbb{R}^d$ *converges in distribution* to Q, denoted $Q_n \xrightarrow{\mathcal{D}} Q$ if $F_{Q_n}(x) \to F_Q(x)$ as $n \to \infty$, for all points x at which F_Q is continuous (this has several equivalent definitions in terms of expectations of continuous functions). If F_Q is continuous everywhere, then convergence in distribution is equivalent to the condition that $\mathsf{KS}(Q_n, Q) \to 0$ as $n \to \infty$. We occasionally abuse notation and write $P_n \xrightarrow{\mathcal{D}} Q_n$ if all the Q_n's have continuous c.d.f.'s and $\mathsf{KS}(P_n, Q_n)$ tends to 0 as n grows.

We extend the notions of statistical difference, KS distance and convergence in distribution to random variables in the natural way. For example, we say $X_n \xrightarrow{\mathcal{D}} Y_n$ if Y_n has a continuous c.d.f. for all n and the KS distance between the distributions of X_n and Y_n tends to 0.

Differential Privacy.

In the sequel, $\mathcal{D}$ will denote a generic domain in which data points lie. $\mathcal{D}^*$ is the set of all finite-length sequences of elements in $\mathcal{D}$. Because order will generally not matter, we will sometimes think of the elements in $\mathcal{D}^*$ as multisets.

We say two finite multisets (alternatively, sequences) $x, x' \subseteq \mathcal{D}$ are neighbors if they differ in a single element: $x \triangle x' = 1$. Differential privacy requires that inserting or removing one input tuple should change the distribution of inputs very little, as measured by parameters ε and δ.

Definition 1 ([DMNS06, DKM+06]). *Algorithm A is (ε, δ)-*differentially private *if* for all *neighboring data sets $x, x' \subseteq \mathcal{D}$ and for all events E,*

$$\Pr(A(x) \in E) \leq e^\varepsilon \Pr(A(x') \in E) + \delta\,.$$

We write ε-differential privacy as shorthand for $(\varepsilon, 0)$-differential privacy.

Most of the algorithms in this paper are ε-differentially private (that is, they satisfy the definition above with $\delta = 0$).

For simplicity, in most of the paper we will assume that the size n of the data set is publicly known. This assumption can be removed at the expense of some technical complication by using a differentially private approximation to n; see Section 2.4.1.

We consider statistics $T : \mathcal{D}^* \to \mathbb{R}^d$ that return a vector of d real numbers. Our main algorithm assumes that T in fact takes values only in a bounded cube $[-\frac{\Lambda}{2}, \frac{\Lambda}{2}]^d$, where $\Lambda > 0$ is known to the algorithm. This assumption can be removed at the price achieving only (ε, δ) differential privacy. See Section 2.4.1 for details.

Asymptotic Normality.

The central limit theorem provides the best known example of asymptotically normal statistics: any statistic which is a sum of $h(X_i)$ for some function h will converge to a normal distribution as long as $h(X_i)$ has finite expectation and variance. Many other statistics also exhibit this type of convergence, however. Typically, a careful inspection of the convergence theorems reveals additional properties: the standard deviation of these statistics usually scales as $n^{-1/2}$; the bias of these statistics, surprisingly, shrinks more quickly, usually at a rate of $1/n$; finally, additional moments of the asymptotic distribution can usually be bounded under mild assumptions. We dub statistics that exhibit these additional properties *generically* asymptotically normal.

Definition 2 (Generic Normality). *A statistic $T : \mathcal{D}^* \to \mathbb{R}$ is* generically asymptotically normal *at distribution P if there exists a "true value" $T(P)$ and constant $\sigma_P^2 > 0$ such that if $X_1, \ldots, X_n \sim P$ i.i.d., then*

1. *(Normality)* $\dfrac{T(X) - T(P)}{\sigma_P/\sqrt{n}} \xrightarrow{\mathcal{D}} N(0,1)$ *as $n \to \infty$,*
2. *(Linear Bias)* $E[T(X)] - T(P) = O(1/n)$*, and*
3. *(Bounded third moment)* $E(\frac{|T(X)-T(P)|}{\sigma_P/\sqrt{n}})^3 = O(1)$. $\diamond$

This definition generalizes to $d > 1$ by replacing σ_p with a symmetric positive definite matrix $\Sigma_P \in \mathbb{R}^{d \times d}$: we require that $Z(n) = \sqrt{n}\Sigma_P^{-1}(T(X) - T(P))$ converges to $N(0^d, I)$ and that the third moment condition holds for the projections along any given direction (that is, for any unit vector u we should have $E[|u^\top Z(n)|^3] = O(1)$).

The second and third condition are necessary because asymptotic normality alone doesn't imply that functionals of $T(X)$, such as its expectation and variance, also converge to the values one naturally expects. Asymptotic normality implies only that the c.d.f. of $T(X)$ is close, point-wise, to a Gaussian's. However, because the range can be very large, it could be that natural functionals of the distribution differ wildly from those of the Gaussian. For example, consider a distribution Q which is a mixture of $N(0,1)$ with weight $1-\rho$, and $N(1/\rho^2, 1)$ with weight ρ. The distance between Q and $N(0,1)$ is ρ, but the expectation of Q goes to ∞ as ρ goes to 0.

The thin tails condition implies that this sort of strange behavior does not occur: for "nicely behaved" functions f, one gets that $E(f(T(X))) \approx E(f(Z))$, where $Z \sim N(T(P), \frac{\sigma_P^2}{n})$.

For examples of asymptotically normal distributions, we refer to the textbook of Schervish [Sch96]. Some examples:

- Given a family of distributions parametrized by a finite number of real numbers, $\{P_\theta | \theta \in \Theta \subseteq \mathbb{R}^d\}$. Under mild regularity conditions, the *maximum likelihood estimator* is g.a.n. and has asymptotically optimal variance among all (asymptotically) unbiased estimators. See, for example, [Sch96, §7.3.5]. This includes as a special case mixture models with a fixed number of components which are described by a finite number of parameters but typically do not have sufficient statistics that are any less compact than the original data set.

- The estimators for common regression problems are also generically asymptotically normal under mild conditions (for example, the data matrix in linear and logistic regression problems should not have very small eigenvalues). In particular, we get estimators for linear and logistic regression coefficients with rate $O(n^{-1/2})$, improving the $n^{-1/2+\gamma}$ (for a small constant $\gamma > 0$) convergence rate of the estimators of Dwork and Lei [DL09].

- More generally, many statistics T can be viewed as functionals mapping the space of distributions (data

sets are just distributions with finite support and rational probability masses) to $\mathbb{R}^d$. With an appropriate topology on the space of distributions, one can define the differentiability of T at a particular distribution T. Differentiable statistics with non-zero derivative at P and satisfying mild moment conditions are generically asymptotically normal. See Fernholz [Fer83] for a compact exposition of the theory.

One striking example where our framework does *not* apply is the differentially private evaluation of *fit* of various models. Many statistics used to evaluate goodness of fit do not have asymptotically normal behavior.

1.1 Main Theorem

THEOREM 3 (MAIN THEOREM, UNBOUNDED RANGE). *Given a statistic $T : \mathcal{D}^n \to \mathbb{R}^d$, there exists a (ε, δ)-differentially private algorithm $A = A_{T,\varepsilon,\delta}$ such that if T is generically asymptotically normal at distribution P, the random variables $A(X)$ and $T(X)$ converge in distribution when X is an i.i.d. sample of size n from P, that is,*

$$\mathsf{KS}(A(X), T(X)) \to 0 \quad \text{as } n \to \infty\,.$$

Moreover, there is a constant $c > 0$ such that convergence continues to hold even if d, ε, Λ change with n as long as $d, \frac{1}{\varepsilon}$ and $\log(\frac{1}{\delta})$ are all at most n^c.

This theorem in fact follows from a slightly different result, which assumes a known bound on the range of T but achieves a stronger privacy guarantee. The reduction from one form to the other of the main theorem is given in Section 2.4.1.

THEOREM 4 (MAIN THEOREM, BOUNDED RANGE). *Given a statistic $T : \mathcal{D}^n \to [-\frac{\Lambda}{2}, \frac{\Lambda}{2}]^d$, there exists a $(\varepsilon, 0)$-differentially private algorithm $A = A_{T,\varepsilon,\Lambda}$ such that if T is generically asymptotically normal at distribution P, the random variables $A(X)$ and $T(X)$ converge in distribution when X is an i.i.d. sample from P, that is,*

$$\mathsf{KS}(A(X), T(X)) \to 0 \quad \text{as } n \to \infty\,.$$

Moreover, there is a constant $c > 0$ such that convergence continues to hold even if d, ε, δ change with n as long as $d, \frac{1}{\varepsilon}$ and $\log(\Lambda)$ are all at most n^c.

Because the KS distance bounds the difference between the cumulative distribution functions of two random variables, the theorem implies that $A(X)$ and $T(X)$ converge in distribution: $\|F_{A(X)} - F_{T(X)}\|_\infty \to 0$ as $n \to \infty$, where F_Y denotes the c.d.f. of random variable Y.

Note that we provide no bound on the rate at which the KS distance converges to 0 in the theorems above. Such a bound would require further assumptions about T (namely, about how quickly T itself converges to the normal distribution).

We focus here simply on establishing convergence. Note that the theorem is meaningful even with a very slow convergence rate. Once the KS distance is below a small constant, say 1/20, the median error of A in estimating $T(P)$ is at most 1.05 times greater than the median error of the nonprivate estimator T.

2. THE ESTIMATORS

In this section we explain the estimator we use and state the main claims about their behavior. Proofs and more detailed analysis are deferred to the next section.

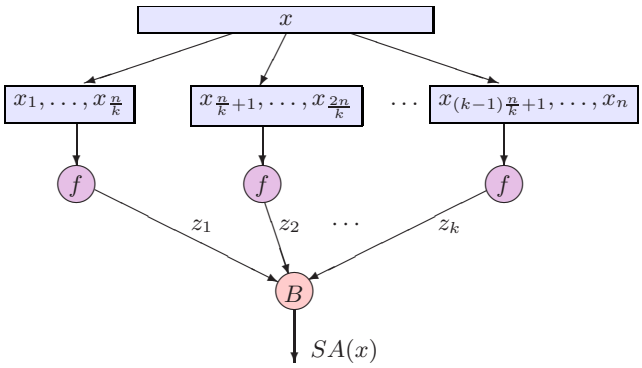

Figure 1: Subample-and-aggregate with a generic aggregation algorithm B. We analyze two specific candidates for B: the sample average and the noisy "widened Winsorized mean".

2.1 Subsample and Aggregate

Nissim et al. [NRS07] introduced the *subsample-and-aggregate* framework for "smoothing" a function f, defined on $\mathcal{D}^*$, to obtain a randomized algorithm which is differentially private. When f is sufficiently well-behaved, the resulting algorithm outputs a value close to that of f with high-probability.

The idea is to randomly partition the input $x \in \mathcal{D}^n$ into k blocks of size roughly n/k each. The function f is applied to each block to obtain k estimates $z_1, ..., z_k$. Finally, these estimates are aggregated using a differentially private function B:

$$SA_{f,B,k}(x) = B(f(x_1, ..., x_{\frac{n}{k}}), ..., f(x_{(k-1)\frac{n}{k}+1}, ..., x_n))\,,$$

as depicted in Figure 1. Nissim et al. observed that this construction is always differentially private, regardless of the structure of f:

LEMMA 5 (PRIVACY OF SA [NRS07]). *Let f be an arbitrary (possibly randomized) function with inputs in $\mathcal{D}_f^*$ and taking values in a domain $\mathcal{D}_B$. If B is ε-differentially private for inputs in $\mathcal{D}_B^*$, then $SA_{f,B,k}$ is ε-differentially private for inputs in $\mathcal{D}_f^*$.*

The intuition behind the framework is the following: if f gives reasonably consistent answers on data sets of size n/k, then the aggregation function can return a value close to the expected value of f on a random subset of x of size n/k. Roughly: SA_f should obtain the same accuracy as the best nonprivate algorithm could obtain on data sets of a smaller size, n/k.

What we show in this paper is that a careful application of the framework gives answers that are as good — up to a $1 + o(1)$ factor — as the best nonprivate algorithm can obtain on datasets of *the same size*.

2.2 A Nonprivate Estimator: Aggregation via Averaging

If we replace the aggregation algorithm B in the subsample-and-aggregate algorithm with a simple average, that is, if we output $\bar{z} = \frac{1}{k}\sum_{i=1}^k z_i$, then we get an estimator that is not differentially private (in particular, it may be deterministic) but which is asymptotically identical to $T(X)$ when T is asymptotically normal. Because we assume our

data is drawn i.i.d., the partitioning into blocks need not be random. We use the natural, consecutive partition for simplicity. We denote the estimator

$$\mathsf{Ave}_{k,T}(x_1, ..., x_n) = \bar{z} = \frac{1}{k}\sum_{i=1}^{k} T(x_{(i-1)\frac{n}{k}+1}, ..., x_{i\frac{n}{k}}) .$$

LEMMA 6 (CONVERGENCE OF AVERAGING). *Suppose that T is generically asymptotically normal at P. If $k = o(\sqrt{n})$, d is bounded above by a sufficiently small polynomial in n and $X_1, ..., X_n$ are drawn i.i.d. from P, then*

$$\sqrt{n} \cdot \mathsf{Ave}_{k,T}(X) \xrightarrow{\mathcal{D}} N(T(P), \sigma_P^2) \text{ as } n \to \infty .$$

Proof Idea: Separating Bias and Variance.

The basic intuition for the analysis is that the averaging step reduces the variance drastically, but does not reduce bias. As long as the bias of the individual estimators is sufficiently low, however, we can still get convergence to the right distribution.

First, since the sample X is i.i.d. from P, the Z_i terms are themselves i.i.d. and (by asymptotic normality) each Z_i is close in distribution to $N(T(P), \frac{k\sigma_P^2}{n})$.

Crudely, one would expect then that the average of the Z_i's would be also close to Gaussian with the same mean $T(P)$ and bias reduced by a factor of k down to $\frac{\sigma_P^2}{n}$.

This doesn't quite work, but it comes close. The variance calculation is essentially correct. One can show that the variance of each Z_i is close to $\frac{k\sigma_P^2}{n}$. Since the Z_i's are independent, the variance of the average $\bar{Z}$ is exactly $\frac{1}{k}$ times the variance of each of the terms, which gives

$$Var(\bar{Z}) \approx \frac{1}{k} \cdot \frac{k\sigma_P^2}{n} = \frac{\sigma_P^2}{n} .$$

On the other hand, the expectation of $\bar{Z}$ is exactly the expectation of the Z_i's, without division by k. The linear bias condition of generic asymptotic normality ensures that $E(Z_i) = T(P) \pm O(k/n)$. When k is asymptotically smaller than $\sqrt{n}$, we get

$$E(\bar{Z}) = T(P) \pm o(n^{-1/2}) .$$

This is not exactly the right expectation, but the $o(n^{-1/2})$ bias term is swamped by the variance and so we get $\bar{Z} \approx N(T(P), \frac{\sigma_P^2}{n})$.

To make this argument formal, we use the bound on the third moment of the Z_i's to convert between expectation/variance calculations and convergence in distribution. The details are explained in Section 3.1.

Relation to the Bootstrap and Jackknife.

The "subsample-and-average" technique above belongs to a family of resampling techniques (of which popular members are the bootstrap and jackknife estimators) used in statistics to estimate the sampling distribution of a statistic. The textbook of Politis, Romano and Wolf [PRW99] describes the general theory of these techniques (see Chapter 2, for example, for the treatment of the i.i.d. setting). The specific variant we use here, in which each observation gets used in exactly one subsample, is not standard. More importantly, to our knowledge our analysis of the bias of the estimator, and the extrapolation to asymptotic convergence, is novel.

2.3 Corollary: Streaming Estimators

Note that $\mathsf{Ave}_{k,T}$ can be evaluated in a single pass over the data set using memory at most n/k plus whatever memory requirements are needed to evaluate T on inputs of size n/k: one need only remember the values x_j for the current block and a running sum of the z_i's computed so far. Setting k to be slightly smaller than $\sqrt{n}$, we obtain:

THEOREM 7 (STREAMING STATISTICAL ESTIMATORS). *Let $T : \mathcal{D}^* \to \mathbb{R}^d$ be a statistic that is computable in linear space and polynomial time. $\tilde{T}(X) = \mathsf{Ave}_{\frac{\sqrt{n}}{\log n}, T}(x)$ can be computed in one pass with $\tilde{O}(\sqrt{n})$ space. If T is g.a.n. at P and X is an i.i.d. sample from P of size n, then $\tilde{T}(X) \xrightarrow{\mathcal{D}} T(X)$ as $n \to \infty$.*

To get some idea of the meaning of this result, consider the naive low-space estimator which estimates $T(P)$ by computing T on a single subsample of size $\sqrt{n}$ from the data set. The standard deviation of this estimator will be on the order of $n^{-1/4}$. In contrast, Theorem 7 gives an estimator with standard deviation $O(n^{-1/2})$.

This result has a quite different flavor from existing work on low-space statistical estimators, since it aims for optimal error and allows for a polynomial amount of space. As mentioned in the introduction, previous work [GM07, CK09, CLM10] focused on regimes with higher error and lower (generally polylogarithmic in n) space. An interesting direction, which we leave to future work, is understanding whether our analysis and techniques could also lead to improvements in lower space regimes.

2.4 Making the Estimator Private: Widened Winsorized Mean

We can make the estimator above differentially private by replacing the empirical average with a differentially private aggregate $W(z_1, ..., z_n)$ which closely approximates the average as long as its input z is an i.i.d. sample from a distribution close to the normal distribution with bounded third moment.

A Winsorized mean rounds outliers to in a data set to lie within a fixed interval, usually defined in terms of some quantiles of the data set. An α-Winsorized mean rounds the αk smallest values up to $Z_{(\alpha k)}$ and rounds the αk largest values down to $Z_{((1-\alpha)k)}$. For example, when $\alpha = \frac{1}{4}$ one "squeezes" the data set to lie in its own interquartile range.

In order to get both differential privacy and statistical efficiency, we adapt this idea in two key ways:

First, we replace the true quartiles with differentially private estimates. There are several ways to estimate quartiles; we found an "exponential-mechanism"-based method due to McSherry and Talwar [MT07] as well as personal communication) to be most amenable to our analysis (the "smoothed sensitivity" approach of Nissim et al. [NRS07] and the "propose-test-release" approach of Dwork and Lei [DL09] also work here but make for messier theorem statements). For completeness, PrivateQuantile is described in Algorithm 2.

Second, we *widen* the Winsorization interval: we scale the interval up by a factor rad (a parameter we set to be roughly $\sqrt[3]{k}$). This scaling allows us to capture enough values in the

tail to get optimal variance but confines the data to a small enough interval that we can add a relatively small amount of noise and still get differential privacy.

The resulting algorithm W is described in Algorithm 1. Note that the privacy guarantee does not make any distributional assumptions about the input.

2.4.1 Simplifying Assumptions

In the remainder of the analysis, we make two simplifying assumptions.

First, we assume that the size of the database, n, is a publicly known value. In general, though, it may not be desirable to release n; we can instead first estimate n using the Laplace mechanism [DMNS06] and then use that estimate in place of n in the remaining algorithms. The data set can be made to have length $n' \neq n$ either by inserting a small number of default values or removing a small number of values. This will not change the asymptotics of the estimators analyzed here.

Second, we assume that the estimator T always takes values in a known range $[-\frac{\Lambda}{2}, \frac{\Lambda}{2}]^d$. One can circumvent this in at least two ways. The first is to monotonically rescale each of the real axes to take only values in $(-1/2, 1/2)$ (say via the map $x \mapsto \frac{1}{\pi}\arctan(x)$). However, this has the net effect of drastically increasing the error for parameter values that are far away from the origin. Another approach is to use the differentially private estimators of scale and location due to Dwork and Lei [DL09] inside W to obtain a crude interval that contains a $(1 - o(1))$ fraction of the data set and which is not too much larger than σ_P. Because the final error depends only logarithmically on the radius Λ of the interval, even a significant overestimate of the range of the data will not affect the final answer significantly. We defer the details to the full version of the paper.

Analysis of WWM.

LEMMA 8. *The algorithm W is ε-differentially private.*

PROOF. The algorithm is a composition of three differentially-private mechanisms: two calls to PrivateQuantile (with parameter $\varepsilon/4$ each) and one application of the Laplace mechanism (with parameter $\varepsilon/2$). Overall, the algorithm is ε-differentially private by the triangle inequality. □

We now turn to the more involved analysis of the efficiency $W(Z)$ as an estimator of the mean.

LEMMA 9. *Let W be the noisy widened Winsorized mean described in Algorithm 1. Let $\hat{\mu}(Z)$ be the initial estimate of $\bar{Z}$ that is computed by the algorithm, and let Y be the Laplace noise that is added to $\hat{\mu}(Z)$. There exists a constant $\beta > 0$ such that for every random variable Z such that $\mathsf{KS}(Z, N(\mu, \sigma^2)) \leq \frac{1}{20}$ and $E(\frac{|Z-\mu|}{\sigma})^3 \leq C$, if $Z_1, ..., Z_k$ are i.i.d. copies of Z, then the following holds:*

There exists an event E of probability at least $1 - Ck^{-\beta} - \Lambda e^{-\Omega(\varepsilon k)}$ such that conditioned on E:

1. *$\hat{\mu}(Z) = \bar{Z}$, and*
2. *$Var(Y) \leq Var(\bar{Z}) \cdot k^{-\beta}$.*

The event E in the preceding lemma is roughly that the quantile estimates obtained in the first part of W are accurate. In that case, the interval $[\ell, u]$ into which observations are projected is likely to contain all the points in Z; this latter event implies $\hat{\mu}(Z) = \bar{Z}$.

Since $\hat{\mu}(Z) = \bar{Z}$ are actually equal with high probability, their distributions are close in statistical difference. We immediately get that $\mathsf{KS}(\hat{\mu}, \bar{Z})$ is small since statistical difference upper bounds KS distance. Second, since under the same conditions, we know that $\bar{Z}$ is close to normal, adding noise Y with variance much smaller than that of Z will not change the distribution significantly. We obtain the following key corollary:

COROLLARY 10 (ACCURACY OF WWM). *Let W be the noisy widened Winsorized mean described in Algorithm 1. There exists a constant $\beta > 0$ such that for every random variable Z such that $\mathsf{KS}(Z, N(\mu, \sigma^2)) \leq \frac{1}{20}$ and $E(\frac{|Z-\mu|}{\sigma})^3 \leq C$, if $Z_1, ..., Z_k$ are i.i.d. copies of Z, then*

$$\mathsf{KS}(W(Z), \bar{Z}) \leq Ck^{-\beta} + \Lambda e^{-\Omega(\varepsilon k)}.$$

In particular, if k goes to infinity and $C, \frac{1}{\varepsilon}$ and $\log(\Lambda)$ are bounded above by sufficiently small polynomials in k, then $\mathsf{KS}(Z, N(\mu, \sigma^2))$ goes to 0.

2.5 Putting the Pieces Together

We can combine the guarantees on W together with the fact, shown above, that $\bar{Z} = \mathsf{Ave}_{k,T}(X)$ is asymptotically equivalent to $T(X)$ (when T is generically asymptotically normal at the distribution of X). The overall estimator A_T is described in Algorithm 3.

To get the final result, we must set k carefully in A_T to balance two factors: if k is too large, then the bias of the estimators on the individual subsamples becomes too big and dominates the variance of the estimator. If k is too small, however, the noise added to ensure the privacy of W becomes too large, and again the overall error becomes too large. We set $k = n^{1/2-\eta}$ where $\eta > 0$ is a small constant. This allows us to prove the bounded-range version of the main result.

PROOF OF THEOREM 4. We have set $k = n^{1/2-\eta}$ so that the analysis of the convergence of simple averaging applies. We get that $\mathsf{KS}(\bar{Z}, N(T(P), \frac{\sigma_P}{n}))$ tends to 0 (by Lemma 6) and that $\mathsf{KS}(W(Z), \bar{Z})$ tends to 0 (by Lemma 9). Since the the KS distance satisfies the triangle inequality, we have that $\mathsf{KS}(W(Z), N(T(P), \frac{\sigma_P}{n}) \leq \mathsf{KS}(N(T(P), \frac{\sigma_P}{n}), \bar{Z}) + \mathsf{KS}(\bar{Z}, W(Z))$ tends to 0 as n grows to ∞. The same argument works in higher dimensions by a straightforward hybrid argument, although the distance bounds degrade by a factor of d. □

As mentioned earlier, the unbounded range version of the result follows by first obtaining a crude estimate of Λ using the technique of Dwork and Lei [DL09].

3. DETAILED ANALYSIS

3.1 Analysis of Simple Averaging

PROOF OF LEMMA 6. Consider first what happens when $d = 1$. Let $\rho_n = \mathsf{KS}(T(X), N(T(P), \sigma_P^2/n))$. By the normality assumption, $\rho_n \to 0$.

By the linear bias assumption, $E(\bar{Z}) = E(Z_1) = T(P) + O(k/n)$.

Algorithm 1: Widened Winsorized Mean W

Input: $Z = (Z_1, \ldots, Z_k) \in \mathbb{R}^k$, parameter $\varepsilon > 0$, bounding parameter $\Lambda > 0$
Output: Estimate $\hat{\mu} = W(Z)$
Set the parameter $\mathsf{rad} = k^{\frac{1}{3}+\eta}$, where $\eta = 1/10$.

`/* First, estimate the range of` Z `with` $[\ell, u]$. `*/`
`/* Use the exponential mechanism to estimate quantiles. */`
$\hat{a} \leftarrow \text{PrivateQuantile}(Z, \frac{1}{4}, \frac{\varepsilon}{4}, \Lambda)$;
$\hat{b} \leftarrow \text{PrivateQuantile}(Z, \frac{3}{4}, \frac{\varepsilon}{4}, \Lambda)$;
$\mu_{crude} \leftarrow \frac{\hat{a}+\hat{b}}{2}$;
$\mathsf{iqr}_{crude} \leftarrow |\hat{b} - \hat{a}|$;
$u \leftarrow \mu_{crude} + 4 \cdot \mathsf{rad} \cdot \mathsf{iqr}_{crude}$;
$\ell \leftarrow \mu_{crude} - 4 \cdot \mathsf{rad} \cdot \mathsf{iqr}_{crude}$;

`/* Now compute Winsorized mean for range` $[\ell, u]$. `*/`
Define $\Pi_{[\ell,u]}(x) \leftarrow \begin{cases} \ell & \text{if } x \le \ell \\ x & \text{if } \ell < x < u \\ u & \text{if } x > u \end{cases}$;
Let $\hat{\mu} \leftarrow \dfrac{1}{k}\sum_{i=1}^{k} \Pi_{[\ell,u]}(Z_i)$;
Sample $Y \sim \mathsf{Lap}(\frac{|u-\ell|}{2\varepsilon k})$ where $\mathsf{Lap}(\lambda)$ is a Laplace distribution with scale parameter λ;
Output $W(Z) = \hat{\mu} + Y$

Algorithm 2: PrivateQuantile(Z, α, ε)

Input: List of real numbers $Z = (Z_1, ..., Z_k)$, quantile $\alpha \in (0, 1)$, privacy parameter $\varepsilon > 0$, bounding parameter $\Lambda > 0$
Output: A real number $x \in [0, \Lambda]$ drawn according to the distribution with density proportional to $\exp(-\frac{\varepsilon}{2}|\alpha k - \#\{i : Z_i \le x\}|)$.
Sort Z_i in ascending order;
Replace $Z_i < 0$ with 0 and $Z_i > \Lambda$ with Λ;
Define $Z_0 = 0$ and $Z_{k+1} = \Lambda$;
For $i = 0, ..., k$, set $y_i = (Z_{i+1} - Z_i)\exp(-\varepsilon|i - \alpha k|)$.
Sample an integer $i \in \{0, ..., k\}$ with probability $y_i/(\sum_{i=0}^{k} y_i)$;
Output a uniform draw from $Z_{i+1} - Z_i$.

Computing the variance is slightly more delicate since, even though the Z_i's are close to $N(T(P), \frac{k\sigma_P^2}{n})$, the variance of Z_i may differ from $\frac{k\sigma_P^2}{n}$. Nevertheless, the bound on the third moment of the Z_i's (which is ensured by generic asymptotic normality assumption) allows us to conclude that the variance is indeed close to what we expect. Lemma 14 in Appendix A implies that

$$Var(\bar{Z}) = \frac{1}{k}Var(Z_1) \le \frac{1}{k}(\frac{\sigma_P^2}{n/k})(1 + O(\rho_{n/k}^{1/3})) = \frac{\sigma_P^2}{n}(1 + o(1)) .$$

Now $\bar{Z}$ is a sum of i.i.d. random variables. The central limit theorem states that such sums converge to a normal distribution. The difficulty here is that we are taking two limits at the same time: the distribution of the Z_i's also changes with k. Fortunately, the Berry-Esseen theorem gives a uniform bound on the convergence rate in terms of the third moment of the Z_i's (which are bounded by the conditions of generic asymptotic normality). The following statement is paraphrased from Feller [Fel71, §16.5]:

THEOREM 11 (BERRY-ESSÉEN). *Let $A_1, ..., A_k$ be i.i.d. realizations of a random variable A with mean 0, variance 1 and third absolute moment $c = E[|A|^3] < \infty$ and let G be a $N(0, 1)$ Gaussian random variable. Then $\mathsf{KS}(\sqrt{n} \cdot \bar{A}, G) \le 9c/\sqrt{n}$.*

Applying the Berry-Esséen theorem to $\frac{Z_i - E(Z)}{Var(Z)}$, we get that $Y_n = \frac{\bar{Z} - E(Z_1)}{Var(Z_1)/\sqrt{n/k}} = \frac{\bar{Z} - E(Z_1)}{(1+o(1))\sigma_P/\sqrt{n}} \xrightarrow{\mathcal{D}} N(0, 1)$.

Dropping the $1+o(1)$ term in the denominator, we get that $Y'_n = \frac{\bar{Z} - E(Z_1)}{\sigma_P/\sqrt{n}}$ converges to $N(0, 1)$ (since $Y'_n = (1+o(1))Y_n$, we get that for any real value y, $\Pr(Y_n \le y) = \Pr(Y'_n \le y(1+o(1)))$; this value converges to $\Pr(Y'_n \le y)$ since Y_n converges to a continuous distribution). Finally, we can replace $E(Z_1)$ with $T(P) \pm O(k/n)$:

$$Y'_n = \frac{\bar{Z} - T(P) \pm O(k/n)}{\sigma_P/\sqrt{n}} = \frac{\bar{Z} - T(P)}{\sigma_P/\sqrt{n}}(1 + O(\frac{k}{\sqrt{n}}))$$

Since $k = o(\sqrt{n})$, the factor $(1 + O(\frac{k}{\sqrt{n}}))$ is $(1 + o(1))$. Hence, $\frac{\bar{Z} - T(P)}{\sigma_P/\sqrt{n}} \xrightarrow{\mathcal{D}} N(0, 1)$, as desired.

The proof extends directly to higher d using higher-dimensional analogues of the central limit theorems and Berry-Esséen bounds. □

Algorithm 3: A_T: Subsample-and-aggregate using Widened Winsorized mean

Input: $X = (X_1, \ldots, X_n)$ in domain $\mathcal{D}$. Description of $T : \mathcal{D}^n \to \mathbb{R}^d$
Output: Estimate $A(X)$ to $T(X)$
Set $k = n^{\frac{1}{2}-\eta}$ where $\eta = 1/10$.
Randomly divide X into k blocks $X^{(1)}, \ldots, X^{(k)}$ of size n/k each;
Compute $Z_i = T(X^{(i)})$ for each block $i = 1, 2, \ldots, k$.
for *each dimension* $j = 1, \ldots, d$ **do**
  `/* Run noisy Widened Winsorized Mean in dimension` j `*/`
  $Z|_j \leftarrow$ projection of $Z = (Z_1, \ldots, Z_k)$ in dimension j;
  $A_j \leftarrow W(Z|_j\,, \frac{\varepsilon}{d})$;
Output $(A_1, \ldots, A_d)$.

3.2 Analysis of Widened Winsorized Mean

The analysis of W hinges on getting good estimates of the quartiles. The following lemma shows the estimates are accurate enough (that is, the interquartile range is correct within a constant factor) with high probability:

LEMMA 12. *Let $Z = (Z_1, ..., Z_k)$ be i.i.d. draws from distribution Q. If Q is within KS distance $\frac{1}{20}$ of $N(\mu, \sigma)$, then, with probability $1 - \frac{\Lambda}{\sigma}e^{-\Omega(\varepsilon k)}$ over the choice of Z, both $\mu - \hat{a}$ and $\hat{b} - \mu$ are in the interval $[\frac{1}{8}\sigma, 2\sigma]$.*

This lemma is proved in the next section. For now, we need only the following corollary:

COROLLARY 13. *If we define ℓ, u as $\mu_{crude} \pm 4(\hat{b} - \hat{a})\mathsf{rad}$, then the probability over the choice of Z that* both *$[\ell, u]$ contains $\mu \pm \sigma\mathsf{rad}$ and $|u - \ell| < 16\sigma\mathsf{rad}$ is $1 - \frac{\Lambda}{\sigma}e^{-\Omega(\varepsilon k)}$.*

We can now prove the main lemma on the accuracy of the noisy widened Winsorized mean.

PROOF OF LEMMA 9. Let E_1 be the event described in Corollary 13. Conditioned on E_1, the standard deviation of Y is $O(\frac{\sigma\mathsf{rad}}{\varepsilon k})$. Since we set $\mathsf{rad} = k^{1/3} + \eta$, the standard deviation of Y is $O(\sigma k^{-\frac{2}{3}+\eta})$, which is much smaller than $\sqrt{Var(\bar{Z})} = \frac{\sigma}{\sqrt{k}}(1 + o(1))$ (see the analysis of the simple averaging estimator for the bound on the variance of $\bar{Z}$).

Let E_2 be the event that all the points in Z lie in the interval $\mu \pm \sigma\mathsf{rad}$. The bound on the third moment of Z implies that

$$\Pr(\tfrac{|Z_1-\mu|}{\sigma} \le \mathsf{rad}) = \textstyle\int_{|w|>\mathsf{rad}} dQ(\mu + \sigma w)$$
$$\le \mathsf{rad}^{-3} \int_{|w|>\mathsf{rad}} |w|^3 dQ(\mu + \sigma w) = C \cdot \mathsf{rad}^{-3}\,.$$

hence the probability of E_2 is at most $kC/\mathsf{rad}^3 = Ck^{-3\eta}$.

To complete the proof of the lemma, let $E = E_1 \cap E_2$. The probability of E is close to 1 since each of E_1 and E_2 occurs with probability close to 1. Moreover, E_2 and E_1 together imply that all the points in Z lie in $[\ell, u]$, and so $\hat{\mu}(Z) = \bar{Z}$. The event E_1 alone implies that the Laplace noise added to the estimate is not too large, as desired. □

3.3 Analysis of PrivateQuantile

PROOF OF LEMMA 12. The constants 1/8 and 2 are chosen to be less than $F^{-1}(11/20) \approx 0.126$ and more than $F^{-1}(19/20) \approx 1.65$, respectively, where F is the c.d.f. of $N(0, 1)$.

We analyze here the case of the upper estiamte $\hat{b}$; the case of the lower estimate $\hat{a}$ is symmetric.

First, we can assume that $\mu = 0$ and $\sigma = 1$ by rescaling the interval $[0, \Lambda]$ to $[-\frac{\mu}{\sigma}, \frac{\mu+\Lambda}{\sigma}]$.

Second, we can assume that the empirical c.d.f. of $Z_1, ..., Z_k$, denoted $\hat{F}_z$, follows F_Z closely. By the Dvoretsky-Kifer-Wolfowitz theorem, the probability that $||\hat{F}_Z - F_Z||_\infty > \beta$ is exponentially small in $\beta^2 k$. We'll assume the c.d.f.'s differ by at most $\beta = 1/1000$; this occurs with probability $1 - \exp(\Omega(k))$. Thus, the overall distance between $\hat{F}_Z$ and F is at most $\rho + \beta$.

Our argument follows the by-now standard outline of analyses of the exponential mechanism [MT07].

Divide the real line into segments delimited by $F^{-1}(11/20), F^{-1}(14/20), F^{-1}(16/20)$, and $F^{-1}(19/20)$. Let BAD be the set of points *outside* of $[F^{-1}(11/20), F^{-1}(19/20)]$ and let $GOOD$ be the set of points *inside* $[F^{-1}(14/20), F^{-1}(16/20)]$.

Consider the distribution sampled from by PrivateQuantile. Note that points in BAD have density in the distribution at most $\exp(-\frac{\varepsilon}{2}(\frac{4k}{20} - \rho - \beta))/M$ where M is a normalizing constant. Points in $GOOD$ have mass at least $\exp(-\frac{\varepsilon}{2}(\frac{1k}{20} + \rho + \beta))/M$. Moreover, the measure of BAD is at most Λ/σ; the measure of $GOOD$ is $F^{-1}(16/20) - F^{-1}(14/20)$. Thus, the probability of a point in BAD being sampled by PrivateQuantile is at most

$$\frac{(\Lambda/\sigma)\exp(-\frac{\varepsilon}{2}(\frac{4k}{20} - \rho - \beta))}{(F^{-1}(16/20) - F^{-1}(14/20))\exp(-\frac{\varepsilon}{2}(\frac{1k}{20} + \rho + \beta))}$$
$$= O(\frac{\Lambda}{\sigma}) \cdot \exp(-\frac{\varepsilon}{2}(\frac{3k}{20} - 2\rho - 2\beta))\,.$$

Recalling that $\rho < 1/20$ and $\beta < 1/1000$, we get the desired bound. □

Acknowledgements

This work benefited from conversations with many colleagues from both statistics and computer science over the last three years. In particular, I am grateful for helpful discussions with Cynthia Dwork, Sofya Raskhodnikova, Steve Fienberg, Sesa Slavković, Bing Li, Jing Lei, Andrew McGregor and Larry Wasserman.

4. REFERENCES

[BDMN05] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: The SuLQ framework. In *PODS*, pages 128–138. ACM, 2005.

[CK09] Kevin L. Chang and Ravi Kannan. Pass-efficient algorithms for learning mixtures of uniform distributions. *SIAM J. Comput.*, 39(3):783–812, 2009.

[CLM10] Steve Chien, Katrina Ligett, and Andrew McGregor. Space-efficient estimation of robust statistics and distribution testing. In *Innovations in Computer Since (ICS)*, pages 251–265, 2010.

[CMS11] K. Chaudhuri, C. Monteleoni, and A.D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 2011.

[DKM+06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, LNCS, pages 486–503. Springer, 2006.

[DL09] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *STOC '09: Proceedings of the 41st annual ACM symposium on Theory of computing*, pages 371–380, 2009.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284. Springer, 2006.

[Dwo06] Cynthia Dwork. Differential privacy. In *ICALP*, LNCS, pages 1–12, 2006.

[Fel71] W. Feller. *An Introduction to Probability Theory and Its Applications*. Wiley, 3rd edition, 1971.

[Fer83] Luisa Turrin Fernholz. *von Mises Calculus for Statistical Functionals*. Springer-Verlag, 1983.

[GM07] Sudipto Guha and Andrew McGregor. Space-efficient sampling. In *AISTATS*, pages 169—176, 2007.

[Kea98] Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998. Preliminary version in *proceedings of STOC'93*.

[KLN+08] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? In *FOCS*, pages 531–540. IEEE Computer Society, 2008.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, pages 94–103. IEEE, 2007.

[NRS07] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, pages 75–84. ACM, 2007.

[PRW99] D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer-Verlag, 1999.

[RBHT] Benjamin I. P. Rubinstein, Peter L. Bartlett, Ling Huang, and Nina Taft. Learning in a large function space: Privacy-preserving mechanisms for svm learning. arXiv:0911.5708v1 [cs.LG].

[Sch96] Mark Schervish. *Theory of Statistics*. Springer, 1996.

[Smi08] Adam Smith. Efficient, differentially private point estimators. *CoRR*, abs/0809.4794, 2008.

[WZ10] Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, Mar 2010. arXiv:0811.2501v1 [math.ST].

APPENDIX

A. VARIANCE BOUNDS

LEMMA 14. *Let Z be drawn from a distribution within KS distance ρ of $N(\mu,\sigma)$ and such that $E(\frac{|Z-\mu|}{\sigma})^s \le C$ for some $s > 2$ and $C \ge 2\sqrt{2\pi}$. Then $Var(Z) = \sigma^2(1 \pm 6C^{\frac{2}{s}}\rho^{\frac{s-2}{s}})$. In particular, when $s = 3$ and C is constant, we get $Var(Z) = \sigma^2(1 \pm O(\rho^{1/3}))$.*

PROOF. Let Q be the distribution of Z. Without loss of generality, we can assume $\mu = 0$ and $sigma = 1$; the general result follows by rescaling. $Var(Z) \le E_Q(Z)^2 = \int_{z=-\infty}^{\infty}(z)^2 dQ(z)$. We can break this interval into two pieces and write it as $\int_{w=0}^{\infty} w^2 dQ(-w) + \int_{w=0}^{\infty} w^2 dQ(w)$. We will show that each of these is pieces very close to $\sigma^2/2$, which is the value of the analogous integral for a $N(0,1)$ standard Gaussian random variable. Consider the component $w > 0$; the corresponding argument for $w < 0$ is symmetric. Let N be the distribution of the Gaussian:

$$\int_{w=0}^{\infty} w^2 dQ(w) - \sigma^2 = \int_{w=0}^{\infty} w^2 dQ(w) - \int_{w=0}^{\infty} w^2 dN(w)$$

Given a bound $T > 0$, which we will set later, we can break both integrals into two pieces, one for $w < T$ and the other for $w \ge T$. We bound the contributions of "small" w by rewriting the integral in terms of the c.d.f., using Lemma 15 (stated below).

$$\int_{w=0}^{T} w^2 dQ(w) - \int_{w=0}^{T} w^2 dN(w) = \int_{t=0}^{T} 2tQ(Z \ge t)dt - \int_{t=0}^{T} 2tN(Z \ge t)dt$$

By the assumption that Q is close to N in KS distance, we have $|Q(Z \ge t) - N(Z \ge t)| \le \rho$, and so the expression above is within $\pm \int_{t=0}^{T} 2t\rho dt = \rho T^2$.

The contribution of "large" w (that is, $w > T$) is bounded using the moment condition:

$$\int_{w=T}^{\infty} w^2 dQ(w) \le T^{s-2} \int_{w=T}^{\infty} w^s dQ(w) = CT^{s-2}\,.$$

Finally, the same bound applies to the analogous integral for the normal, as long as C at least as large as the third (absolute) moment of the normal, which is $2\sqrt{2\pi}$.

Putting the pieces together, and multiplying by two to account for the contribution of $w < 0$, we get

$$|Var(Z) - \sigma^2| \le 2(\rho T^2 + 2CT^{s-2})\,.$$

Setting $T = (\frac{\rho}{C})^{-1/s}$, we get that $|Var(Z) - \sigma^2| \le 6C^{\frac{2}{s}}\rho^{\frac{s-2}{s}}$, as desired. □

LEMMA 15 (WRITING EXPECTATIONS USING CDFS). *Let $f : \mathbb{R} \to \mathbb{R}$ be differentiable on $[0,\infty]$ such that $f(0) = 0$ and let Z be a nonnegative random variable. Then*

$$E_Q(f(Z)) = \int_{t=0}^{\infty} f'(t)Q(Z \ge t)dt\,,$$

as long as the integral on the right converges absolutely (that is, as long as $\int_{t=0}^{\infty} |f'(t)|Q(Z \ge t)dt < \infty$).

Note that if Q has bounded support and f is bounded on the support of Q, then the integral on the right-hand side will always converge absolutely.