Proceedings of Machine Learning Research vol 65:1–30, 2017

Generalization for Adaptively-chosen Estimators via Stable Median

Vitaly Feldman *IBM Research – Almaden* VITALY@POST.HARVARD.EDU

Thomas Steinke *IBM Research – Almaden* MEDIAN@THOMAS-STEINKE.NET

Abstract

Datasets are often reused to perform multiple statistical analyses in an adaptive way, in which each analysis may depend on the outcomes of previous analyses on the same dataset. Standard statistical guarantees do not account for these dependencies and little is known about how to provably avoid overfitting and false discovery in the adaptive setting. We consider a natural formalization of this problem in which the goal is to design an algorithm that, given a limited number of i.i.d. samples from an unknown distribution, can answer adaptively-chosen queries about that distribution.

We present an algorithm that estimates the expectations of k arbitrary adaptively-chosen real-valued estimators using a number of samples that scales as $\sqrt{k}$. The answers given by our algorithm are essentially as accurate as if fresh samples were used to evaluate each estimator. In contrast, prior work yields error guarantees that scale with the worst-case sensitivity of each estimator. We also give a version of our algorithm that can be used to verify answers to such queries where the sample complexity depends logarithmically on the number of queries k (as in the reusable holdout technique).

Our algorithm is based on a simple approximate median algorithm that satisfies the strong stability guarantees of differential privacy. Our techniques provide a new approach for analyzing the generalization guarantees of differentially private algorithms.

1. Introduction

Modern data analysis is an iterative process in which multiple algorithms are run on the same data, often in an adaptive fashion: the algorithms used in each step depend on the results from previous steps. Human decisions may also introduce additional, implicit dependencies between the data and the algorithms being run. In contrast, theoretical analysis in machine learning and statistical inference focuses on fixed and pre-specified algorithms run on "fresh" data. The misuse of theoretical guarantees given by such analyses in real-world data analysis can easily lead to overfitting and false discovery.

While the issue has been widely recognized and lamented for decades, until recently the only known "safe" approach for dealing with general dependencies that can arise in the adaptive setting was "data splitting" — that is, dividing up the data so that untouched data is available when needed. While easy to analyze, data splitting is overly conservative for most settings and prohibitively expensive for analyses that include multiple adaptive steps. In a recent work, Dwork, Feldman, Hardt, Pitassi, Reingold, and Roth (2014) have proposed a new formalization of this problem. In their formulation, the analyst (or data analysis algorithm) does not have direct access to the dataset and instead can ask queries about the (unknown) data distribution. Each query corresponds to a procedure that the analyst wishes to execute on the data. The challenge is thus to design an algorithm that provides answers to queries that are close to answers that would have been obtained had the corresponding analyses been run on samples freshly drawn from the data distribution.

Previous Work

Dwork et al. (2014) consider *statistical queries* (Kearns, 1998). Each statistical query is specified by a function $\psi : \mathcal{X} \to [-1, 1]$ and corresponds to analyst wishing to compute $\mathbf{E}_{X \sim \mathcal{P}}[\psi(X)]$, where $\mathcal{P}$ is the data distribution (usually done by using the empirical mean of ψ on the dataset). A value $v \in \mathbb{R}$ is a valid response to such query if $|v - \mathbf{E}_{X\sim\mathcal{P}}[\psi(X)]| \le \tau$, where the parameter τ determines the desired accuracy. Dwork et al. demonstrated that, if a query answering algorithm $M : \mathcal{X}^n \to [-1, 1]$ is *differentially private*[1] as well as empirically accurate — that is, $\left|M(S) - \frac{1}{n}\sum_{i\in[n]}\psi(S_i)\right| \le \frac{\tau}{2}$ with high probability — then the answers produced by $M(S)$ will be valid with high probability if we take $S \sim \mathcal{P}^n$ to be n i.i.d. samples from the data distribution. Using algorithms from the differential privacy literature, they demonstrated several improvements over sample splitting. For example, they showed that simply perturbing the empirical mean by adding Gaussian noise allows answering k queries using a dataset whose size scales as $\sqrt{k}$. In contrast, either sample splitting or answering using empirical averages both require a dataset whose size scales linearly in k (Dwork et al., 2014). (Note that if the queries were non-adaptive, $\log k$ scaling would be sufficient.) They also showed a computationally inefficient algorithm that has optimal $\log k$ dependence (at the expense of an additional $\sqrt{\log|\mathcal{X}|}$ factor, where $\mathcal{X}$ is the set of possible data points, i.e. the potential support of $\mathcal{P}$.).

Bassily, Nissim, Smith, Steinke, Stemmer, and Ullman (2016) extended the results of Dwork et al. to *low-sensitivity queries*, where a query is specified by a function $\phi : \mathcal{X}^n \to \mathbb{R}$ such that $|\phi(s) - \phi(s')| \le 1/n$ if s and s' differ on a single element. Each such query corresponds to computing a real-valued estimator (or statistic) on the dataset. A value $v \in \mathbb{R}$ is a valid response to such query if $|v - \mathbf{E}_{S\sim\mathcal{P}^n}[\phi(S)]| \le \tau$. (Note that the empirical mean of a statistical query ψ is a low-sensitivity query, namely $\phi(s) = \frac{1}{n}\sum_{i\in[n]}\psi(s_i)$.) They also introduced a simpler and quantitatively sharper analysis showing that answering k statistical queries $\psi : \mathcal{X} \to [-1, 1]$ to accuracy τ is possible efficiently with $n = \tilde{O}(\sqrt{k}/\tau^2)$ samples and inefficiently with $n = \tilde{O}(\sqrt{\log|\mathcal{X}|}\log(k)/\tau^3)$ samples — in both cases improving the bounds of Dwork et al. by a $\sqrt{\tau}$ factor.

Hardt and Ullman (2014) and, subsequently with tighter parameters, Steinke and Ullman (2015) showed that answering k adaptively-chosen statistical queries (even to fixed accuracy, e.g. $\tau = 0.1$) requires a number of samples n that scales with $\sqrt{k}$ in the worst case. This lower bound holds under one of two assumptions: Either the data universe is large – that is, $|\mathcal{X}| \ge 2^k$ – or M is assumed to be computationally efficient so that it cannot break symmetric cryptography with $\log|\mathcal{X}|$-bit secret keys.

Still, in many applications the effects of adaptivity on generalization are much milder than in the adversarial setting considered in the lower bounds. For such settings, Dwork et al. (2015b) proposed the *reusable holdout* technique, which allows answering a large number of "verification" queries. In this technique, the analyst has unresricted access to

[1] Differential privacy is a strong notion of algorithmic stability developed in the context of privacy preserving data analysis (Dwork et al., 2006a). See Definition 2.1.

most of the dataset, but sets aside a subset of data as a "holdout" that is only accessed via queries. Given a query ψ and an estimate v, the goal of the algorithm is to verify that $|v - \mathbf{E}_{X\sim\mathcal{P}}[\psi(X)]| \lesssim \tau$. If the estimate is valid, then the algorithm only needs to respond "Yes". Otherwise, it outputs "No" and a valid estimate. They demonstrated that it is possible to verify k statistical queries while correcting at most ℓ of them using $n = \tilde{O}(\sqrt{\ell}\log(k)/\tau^2)$ samples.

Generalization guarantees in the adaptive setting can also be derived via other approaches such as bounding the approximate max-information or mutual information between the choice of estimator and the data (Dwork et al., 2015a; Rogers et al., 2016; Russo and Zou, 2016) or using other notions of stability that compose in the adaptive setting (Bassily et al., 2016; Bassily and Freund, 2016).

Our Results

In this work, we demonstrate a simple algorithm that can provide accurate answers to adaptively-chosen queries corresponding to arbitrary real-valued estimators. Specifically, let $\phi : \mathcal{X}^t \to \mathbb{R}$ be an arbitrary estimator, where the expectation $\mathbf{E}_{Z\sim\mathcal{P}^t}[\phi(Z)]$ is equal (in which case the estimator is referred to as unbiased) or sufficiently close to some parameter or value of interest. The number of samples t that ϕ uses corresponds to the desired estimation accuracy (naturally, larger values of t allow more accurate estimators). Our algorithm estimates the expectation $\mathbf{E}_{Z\sim\mathcal{P}^t}[\phi(Z)]$ to within (roughly) the standard deviation of $\phi(Z)$ — i.e. $\tau \approx \mathrm{sd}(\phi(\mathcal{P}^t)) \doteq \sqrt{\mathbf{Var}_{Z\sim\mathcal{P}^t}[\phi(Z)]}$ — or, more generally, to within the interquartile range of the distribution of ϕ on fresh data (i.e. the distribution of $\phi(Z)$ for $Z \sim \mathcal{P}^t$, which we denote by $\phi(\mathcal{P}^t)$). If $\phi(s) = \frac{1}{t}\sum_{i=1}^t \psi(s_i)$ for a function $\psi : \mathcal{X} \to [-1,1]$, then the error roughly scales as $\tau \approx \mathrm{sd}(\psi(\mathcal{P}))/\sqrt{t}$. This gives a natural interpretation of t as an accuracy parameter.

In contrast, given a comparable number of samples, the existing algorithms for statistical queries (Dwork et al., 2014; Bassily et al., 2016) give an estimate with accuracy roughly $\tau \approx \sqrt{1/t}$ regardless of the variance of the query. This is not just an artifact of existing analysis since, to ensure the necessary level of differential privacy, this algorithm adds noise whose standard deviation scales as $\sqrt{1/t}$. For a statistical query $\psi : \mathcal{X} \to [-1,1]$, the standard deviation of $\psi(\mathcal{P})$ is upper bounded by 1, but is often much smaller. For example, when estimating the accuracy of a binary classifier with (low) error p our algorithm will give an estimate with accuracy that scales as $\sqrt{p/t}$, rather than $\sqrt{1/t}$.

We now describe the guarantees of our algorithm more formally starting with the simple case of statistical queries. Given a statistical query $\psi : \mathcal{X} \to [-1,1]$ and t fresh random i.i.d. samples $Z \in \mathcal{X}^t$ drawn from $\mathcal{P}$, the empirical mean estimator $\phi(Z) \doteq \frac{1}{t}\sum_{i\in[t]}\psi(Z_i)$ is an unbiased estimator of $\mathbf{E}_{X\sim\mathcal{P}}[\psi(X)]$ with standard deviation equal to $\mathrm{sd}(\psi(\mathcal{P}))/\sqrt{t}$. Applied to such estimators, our algorithm answers adaptively-chosen statistical queries with accuracy that scales as $\mathrm{sd}(\psi(\mathcal{P}))/\sqrt{t}$.

Theorem 1.1 *For all $\zeta > 0$, $t \in \mathbb{N}$, $\beta > 0$, $k \in \mathbb{N}$, and $n \geq n_0 = O(t\sqrt{k\log(1/\beta)}\log(k/\beta\zeta))$, there exists an efficient algorithm M that, given n samples from an unknown distribution $\mathcal{P}$, provides answers $v_1, \ldots, v_k \in [-1,1]$ to an adaptively-chosen sequence of queries*

$\psi_1, \ldots, \psi_k : \mathcal{X} \to [-1,1]$ *and satisfies:*

$$\Pr_{\substack{S \sim \mathcal{P}^n \\ M(S)}} \left[\forall j \in [k] \quad \left| v_j - \mathop{\mathbf{E}}_{X \sim \mathcal{P}} [\psi_j(X)] \right| \le 2 \cdot \frac{\mathrm{sd}(\psi_j(\mathcal{P}))}{\sqrt{t}} + \zeta \right] \ge 1 - \beta.$$

Note that our guarantees also have an additional ζ *precision* term. The dependence of sample complexity on $1/\zeta$ is logarithmic and can be further improved using more involved algorithms. Thus the precision term ζ, like the failure probability β, can be made negligibly small at little cost. Previous work (Dwork et al., 2014; Bassily et al., 2016) gives an error bound of $\sqrt{1/t}$, which is at least as large (up to constant factors) as our bound $\mathrm{sd}(\psi_j(\mathcal{P}))/\sqrt{t} + \zeta$, when $\zeta \le 1/t$. For comparison, in the non-adaptive setting the same error guarantee can be obtained using $n = O(t \log(k/\beta))$ samples (with $\zeta = 1/t$).

Note that, for simplicity, in Theorem 1.1 the range of each query is normalized to $[-1,1]$. However this normalization affects only the precision term ζ. In particular, for queries whose range is in an interval of length at most b, the number of samples that our result requires to get precision ζ scales logarithmically in b/ζ. In contrast, the sample complexity of previous results scales quadratically in b. Further, a more refined statement discussed below allows us to handle queries with arbitrary range.

For general real-valued estimators of the form $\phi : \mathcal{X}^t \to [-1,1]$ our algorithm gives the following guarantees. (For simplicity we will assume that t is fixed for all the queries.)

Theorem 1.2 *For all $\zeta > 0$, $t \in \mathbb{N}$, $\beta > 0$, $k \in \mathbb{N}$, and $n \ge n_0 = O(t\sqrt{k \log(1/\beta)} \log(k/\beta\zeta))$, there exists an efficient algorithm M that, given n samples from an unknown distribution $\mathcal{P}$, provides answers $v_1, \ldots, v_k \in [-1,1]$ to an adaptively-chosen sequence of queries $\phi_1, \ldots, \phi_k : \mathcal{X}^t \to [-1,1]$ and satisfies:*

$$\Pr_{\substack{S \sim \mathcal{P}^n \\ M(S)}} \left[\forall j \in [k] \quad \left| v_j - \mathop{\mathbf{E}}_{Z \sim \mathcal{P}^t} [\phi_j(Z)] \right| \le 2 \cdot \mathrm{sd}(\phi_j(\mathcal{P}^t)) + \zeta \right] \ge 1 - \beta.$$

For general estimators, previous work gives accuracy guarantees in terms of the worst-case sensitivity. More formally, for $\phi : \mathcal{X}^n \to \mathbb{R}$, let $\Delta(\phi)$ denote the worst-case sensitivity of ϕ — that is, $\Delta(\phi) = \max_{z,z'} \phi(z) - \phi(z')$, where the maximum is over $z, z' \in \mathcal{X}^n$ that differ in a single element. The analysis of Bassily et al. (2016) shows that for k adaptively-chosen queries $\phi_1, \ldots, \phi_k$, each query ϕ_i, can be answered with accuracy $\sqrt{n \cdot \sqrt{k}} \cdot \Delta(\phi_i)$ (ignoring logarithmic factors and the dependence on the confidence β). This setting is somewhat more general than ours, since each query is applied to the entire dataset available to the algorithm, whereas in our setting each query is applied to fixed-size subsamples of the dataset. This means that in this setting the space of estimators that can be applied to data is richer than in ours and might allow better estimation of some underlying quantity of interest. At the same time, our techniques gives better accuracy guarantees for finding the expectation of each estimator.

To see the difference between our setting and that in (Bassily et al., 2016), consider estimation of the lowest expected loss of a function from a family $\mathcal{F}$, namely $L^* \doteq \min_{f \in \mathcal{F}} \mathbf{E}_{X \sim \mathcal{P}}[L(f,X)]$, where $L : \mathcal{F} \times \mathcal{X} \to [-R,R]$ is some loss function. Given a dataset s of size n, the standard ERM estimator is defined as $\phi_n(s) \doteq \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i \in [n]} L(f, s_i)$. Using uniform convergence results, one can often obtain that $|L^* - \mathbf{E}_{S \sim \mathcal{P}^n}[\phi_n(S)]| = O(d/\sqrt{n})$, for some d that measures the capacity of $\mathcal{F}$ and also depends on L. The sensitivity of the estimator ϕ_n is upper bounded by $2R/n$. Thus the algorithm in (Bassily et al., 2016) will give an estimate of $\mathbf{E}_{S\sim\mathcal{P}^n}[\phi_n(S)]$ within (roughly) $R\cdot\sqrt{\frac{\sqrt{k}}{n}}$ and an upper bound on the total error will scale as $(d+R\cdot k^{1/4})/\sqrt{n}$. In our setting, the estimator ϕ_t will be used, where t scales as $n/\sqrt{k}$. The bias of this estimator is $|L^* - \mathbf{E}_{S\sim\mathcal{P}^t}[\phi_t(S)]| = O(d/\sqrt{t}) = O(d\cdot k^{1/4}/\sqrt{n})$. At the same time we can estimate $\mathbf{E}_{S\sim\mathcal{P}^t}[\phi_t(S)]$ within (roughly) the standard deviation of ϕ_t. The standard deviation of ϕ_t is always upper bounded by $2R\cdot\sqrt{t} = R\cdot k^{1/4}/\sqrt{n}$, but is often much smaller. Hence, depending on the setting of the parameters and the distribution $\mathcal{P}$, our approach gives an error bound that is either higher by a factor of $k^{1/4}$ or lower by a factor of R/d (than the approach in (Bassily et al., 2016)). In other words, the two approaches provide incomparable guarantees for this problem.

Both Theorem 1.2 and Theorem 1.1 are corollaries of the following more general result. Define

$$\mathrm{qi}_{\mathcal{D}}(a,b) \doteq \{v\in\mathbb{R} \ : \ \Pr_{Y\sim\mathcal{D}}[Y\le v] > a \ \wedge \ \Pr_{Y\sim\mathcal{D}}[Y<v] < b\}$$

to be the (a,b)-*quantile interval* of the distribution $\mathcal{D}$. We refer to the $(1/4, 3/4)$-quantile interval as the *interquartile interval.*

Theorem 1.3 *For all $T\subset\mathbb{R}$ with $|T|<\infty$, $t\in\mathbb{N}$, $\beta>0$, $k\in\mathbb{N}$, and*

$$n \ge n_0 = O(t\sqrt{k\log(1/\beta)}\log(|T|k/\beta)),$$

there exists an efficient algorithm M that, given n samples from an unknown distribution $\mathcal{P}$, provides answers $v_1,\dots,v_k\in T$ to an adaptively-chosen sequence of queries $\phi_1,\dots,\phi_k : \mathcal{X}^t\to T$ and satisfies:

$$\Pr_{\substack{S\sim\mathcal{P}^n\\ M(S)}}\left[\forall j\in[k] \quad v_j\in\mathrm{qi}_{\phi_j(\mathcal{P}^t)}\left(\frac{1}{4},\frac{3}{4}\right)\right] \ge 1-\beta.$$

We make two remarks about the guarantee of the theorem:

Accuracy in terms of interquartile interval: The accuracy guarantee of Theorem 1.3 is that each returned answer lies in the $(1/4, 3/4)$-quantile interval of the distribution of the query function on fresh data (i.e. the distribution $\phi(\mathcal{P}^t)$). The length of this interval is referred to as the *interquartile range*. This guarantee may appear strange at first sight, but it is actually a strengthening of Theorem 1.2: by Chebyshev's inequality, the interquartile interval of any distribution $\mathcal{D}$ lies within two standard deviations of the mean:

$$\mathrm{qi}_{\mathcal{D}}\left(\frac{1}{4},\frac{3}{4}\right) \subseteq \left[\mathop{\mathbf{E}}_{Y\sim\mathcal{D}}[Y] - 2\cdot\mathrm{sd}(\mathcal{D}), \mathop{\mathbf{E}}_{Y\sim\mathcal{D}}[Y] + 2\cdot\mathrm{sd}(\mathcal{D})\right].$$

However, the interquartile interval may be significantly smaller if $\mathcal{D}$ is heavy-tailed. If, for example, the distribution $\mathcal{D}$ has infinite variance, then our guarantee is still useful, whereas bounds in terms of standard deviation will be meaningless. This formulation does not even assume that the quantity of interest is the expectation of $\phi(\mathcal{P}^t)$ (or even that this expectation exists). In fact, ϕ could be a biased estimator of some other parameter of interest.

Intuitively, we can interpret this accuracy guarantee as follows. If we know that a sample from $\phi(\mathcal{P}^t)$ is an acceptable answer with probability at least $3/4$ and the set of acceptable answers forms an interval, then with high probability the answer returned by our algorithm is acceptable. The constants $1/4$ and $3/4$ are, of course, arbitrary. More generally we can demand $v_j \in \mathrm{qi}_{\phi_j(\mathcal{P}^t)}(a, b)$ for any $0 \le a < b \le 1$. However, this reduction increases the sample complexity n_0 by a factor of $1/(b-a)^2$.

Finite range T: Theorem 1.3 assumes that the queries have a finite range. This is necessary for our algorithm, as the required number of samples grows with the size of T, albeit slowly.[2] To obtain Theorem 1.2 from Theorem 1.3, we simply set T to be the discretization of $[-1, 1]$ with granularity ζ and round the output of $\phi : \mathcal{X}^t \to [-1, 1]$ to the nearest point in T; this introduces the additive error ζ. However, allowing T to be arbitrary provides further flexibility. For example, T could be a grid on a logarithmic scale, yielding a multiplicative, rather than additive, accuracy guarantee. Furthermore, in some settings the range of the query is naturally finite and the scale-free guarantee of Theorem 1.3 comes at no additional cost. We also note that, in general, our result allows a different T_j to be chosen (adaptively) for each query ϕ_j as long as the size of each T_j is upper-bounded by some fixed value.

Another advantage of this formulation is that it removes the dependence on the entire range of ϕ: If we know that the interquartile interval of $\phi(\mathcal{P}^t)$ lies in some interval $[a, b]$, we can truncate the output range of ϕ to $[a, b]$ (and then discretize if necessary). This operation does not affect the interquartile interval of $\phi(\mathcal{P}^t)$ and hence does not affect the guarantees of our algorithm. In particular, this means that in Theorem 1.2 we do not need to assume that the range of each ϕ is bounded in $[-1, 1]$; it is sufficient to assume that the interquartile interval of $\phi(\mathcal{P}^t)$ lies in $[-1, 1]$ to obtain the same guarantee. For example, if we know beforehand that the mean of $\phi(\mathcal{P}^t)$ lies in $[-1, 1]$ and its standard deviation is at most 1 then we can truncate the range of ϕ to $[-3, 3]$.

Verification queries: We next consider queries that ask for "verification" of a given estimate of the expectation of a given estimator — each query is specified by a function $\phi : \mathcal{X}^t \to \mathbb{R}$ and a value (or "guess") $v \in \mathbb{R}$. The task of our algorithm is to check whether or not $v \in \mathrm{qi}_{\phi(\mathcal{P}^t)}(\rho, 1-\rho)$ for some ρ chosen in advance. Such queries are used in the reusable holdout setting (Dwork et al., 2015b) and in the `EffectiveRounds` algorithm that uses fresh subset of samples when a verification query fails (Dwork et al., 2014). We give an algorithm for answering adaptively-chosen verification queries with the following guarantees.

Theorem 1.4 *For all $\alpha, \beta, \rho \in (0, 1/4)$ with $\alpha < \rho$ and $t, \ell, k, n \in \mathbb{N}$ with*

$$n \ge n_0 = O(t\sqrt{\ell \log(1/\alpha\beta)} \log(k/\beta)\rho/\alpha^2),$$

there exists an efficient algorithm M that, given n samples from an unknown distribution $\mathcal{P}$, provides answers to an adaptively-chosen sequence of queries $(\phi_1, v_1), \ldots, (\phi_k, v_k)$ (where $\phi_j : \mathcal{X}^t \to \mathbb{R}$ and $v_j \in \mathbb{R}$ for all $j \in [k]$) and satisfies the following with probability at least $1 - \beta$: for all $j \in [k]$

[2] Using a more involved algorithm from Bun et al. (2015), it is possible to improve the dependence on the size of T from $O(\log |T|)$ to $2^{O(\log^* |T|)}$, where $\log^*$ denotes the iterated logarithm — an extremely slow-growing function.

- *If* $v_j \in \mathrm{qi}_{\phi_j(\mathcal{P}^t)}(\rho, 1-\rho)$*, then the algorithm outputs "Yes".*
- *If* $v_j \notin \mathrm{qi}_{\phi_j(\mathcal{P}^t)}(\rho-\alpha, 1-\rho+\alpha)$ *the algorithm outputs "No".*[3]

However, once the algorithm outputs "No" in response to ℓ *queries, it stops providing answers.*

To answer the ℓ queries that do not pass the verification step we can use our algorithm for answering queries in Theorem 1.3 (with k there set to ℓ here).

Answering many queries: Finally, we show an (inefficient) algorithm that requires a dataset whose size scales as $\log k$ at the expense of an additional $\sqrt{t \log |\mathcal{X}|}$ factor.

Theorem 1.5 *For all* $T \subset \mathbb{R}$ *with* $|T| < \infty$, $t \in \mathbb{N}$, $\beta > 0$, $k \in \mathbb{N}$, *and*

$$n \geq n_0 = O(t^{3/2} \cdot \sqrt{\log |\mathcal{X}| \log(1/\beta)} \cdot \log(k \log |T|/\beta)),$$

there exists an algorithm M *that, given* n *samples from an unknown distribution* $\mathcal{P}$ *supported on* $\mathcal{X}$*, provides answers* $v_1, \ldots, v_k \in T$ *to an adaptively-chosen sequence of queries* $\phi_1, \ldots, \phi_k : \mathcal{X}^t \to T$ *and satisfies*

$$\Pr_{\substack{S \sim \mathcal{P}^n \\ M(S)}} \left[\forall j \in [k] \quad v_j \in \mathrm{qi}_{\phi_j(\mathcal{P}^t)}\left(\frac{1}{4}, \frac{3}{4}\right) \right] \geq 1 - \beta.$$

We remark that when applied to low-sensitivity queries with $t = 1/\tau^2$, this algorithm improves dependence on $|\mathcal{X}|$ and τ from $n = \tilde{O}(\log(|\mathcal{X}|)/\tau^4)$ in Bassily et al. (2016) to $n = \tilde{O}(\sqrt{\log |\mathcal{X}|}/\tau^3)$ (although, as pointed out above, the setting in that work is not always comparable to ours).

Techniques

Like the prior work (Dwork et al., 2014; Bassily et al., 2016), we rely on properties of differential privacy (Dwork et al., 2006a). Differential privacy is a stability property of an algorithm, namely it requires that replacing any element in the input dataset results in a small change in the output distribution of the algorithm. As a result, a function output by a differentially private algorithm on a given dataset generalizes to the underlying distribution (Dwork et al., 2014; Bassily et al., 2016). Specifically, if a differentially private algorithm is run on a dataset drawn i.i.d from any distribution and the algorithm outputs a low-sensitivity function, then the empirical value of that function on the input dataset is close to the expectation of that function on a fresh dataset drawn from the same distribution.

The second crucial property of differential privacy is that it composes adaptively: running several differentially private algorithms on the same dataset still satisfies differential privacy (with somewhat worse parameters) even if each algorithm depends on the output of all the previous algorithms. This property makes it possible to answer adaptively-chosen queries with differential privacy and a number of algorithms have been developed for answering different types of queries. The generalization property of differential privacy then

[3] If neither of these two cases applies, the algorithm may output either "Yes" or "No."

implies that such algorithms can be used to provide answers to adaptively-chosen queries while ensuring generalization (Dwork et al., 2014).

For each query, the algorithm of Theorem 1.3 first splits its input, consisting of n samples, into $m = n/t$ disjoint subsamples of size t and computes the estimator $\phi : \mathcal{X}^t \to \mathbb{R}$ on each. It then outputs an approximate *median* of the resulting values in a differentially private manner. Here an approximate median is any value that falling between the $(1-\alpha)/2$-quantile and the $(1+\alpha)/2$-quantile of the m computed values (for some approximation parameter α). In addition to making the resulting estimator more stable, this step also amplifies the probability of success of the estimator. For comparison, the previous algorithms compute the estimator once on the whole sample and then output the value in a differentially private manner.

We show that differential privacy ensures that an approximate empirical median with high probability falls within the true interquartile interval of the estimator on the input distribution. Here we rely on the known strong connection between differential privacy and generalization (Dwork et al., 2014; Bassily et al., 2016). However, our application relies on stability to replacement of any one of the m subsamples (consisting of t points) used to evaluate the estimator, whereas previous analyses used the stability under replacement of any one out of the n data points. The use of this stronger condition is crucial for bypassing the limitations of previous techniques while achieving improved accuracy guarantees.

A number of differentially-private algorithms for approximating the empirical median of values in a dataset have been studied in the literature. One common approach to this problem is the use of local sensitivity (Nissim et al., 2007) (see also Dwork and Lei (2009)). This approach focuses on additive approximation guarantees and requires stronger assumptions on the data distribution to obtain explicit bounds on the approximation error.

In this paper we rely on a data-dependent notion of approximation to the median in which the goal is to output any value v between the $(1-\alpha)/2$-quantile and the $(1+\alpha)/2$-quantile of the empirical distribution for some approximation parameter α. It is easy to see that this version is essentially equivalent to the *interior point* problem in which the goal is to output a value between the smallest and largest values in the dataset. Bun et al. (2015) recently showed that the optimal sample complexity of privately finding an interior point in a range of values T lies between $m = 2^{(1+o(1))\log^* |T|}$ and $m = \Omega(\log^* |T|)$, where $\log^*$ is iterated logarithm or inverse tower function satisfying $\log^*(2^x) = 1 + \log^*(x)$.

The algorithm in (Bun et al., 2015) is relatively complex and, therefore, here we will use a simple and efficient algorithm that is based on the exponential mechanism (McSherry and Talwar, 2007), and is similar to an algorithm by Smith (2011) to estimate quantiles of a distribution. This algorithm has sample complexity $m = O(\log |T|)$ (for constant $\alpha > 0$). In addition, for our algorithm that answers many queries we will use another simple algorithm based on approximate binary search, which can yield sample complexity $m = O(\sqrt{\log |T|})$; it has the advantage that it reduces the problem to a sequence of statistical queries.[4]

Now the second part of our proof: we show that an approximate empirical median is also in the interquartile interval of the distribution. i.e. we show that any value in the *empirical* (3/8,5/8)-quantile interval falls in the *distribution's* (1/4, 3/4)-quantile interval. A value

[4] We are not aware of a formal reference but it is known that the exponential mechanism and binary search can be used to privately find an approximate median in the sense that we use in this work (*e.g.* (Raskhodnikova and Smith, 2010; Nissim and Sheffet, 2014)).

$v \in T$ is an approximate empirical median if the empirical cumulative distribution function at v is close to $1/2$ — that is, $\mathrm{cdf}_S(v) \doteq \mathbf{Pr}_{Y \sim S}[Y \leq v] \approx 1/2$, where $S \sim \mathcal{D}^n$ is the random samples and $Y \sim S$ denotes picking a sample from S randomly and uniformly. Note that $\mathrm{cdf}_S(v)$ is the empirical mean of a statistical query over S, whereas the true cumulative distribution function $\mathrm{cdf}_{\mathcal{D}}(v) \doteq \mathbf{Pr}_{Y \sim \mathcal{D}}[Y \leq v]$ is the true mean of the same statistical query. Hence we can apply the known strong connection between differential privacy and generalization for statistical queries (Dwork et al., 2014; Bassily et al., 2016) to obtain that $\mathrm{cdf}_{\mathcal{D}}(v) \approx \mathrm{cdf}_S(v)$. This ensures $\mathrm{cdf}_{\mathcal{D}}(v) \approx 1/2$ and hence v falls in $\mathrm{qi}_{\mathcal{D}}(1/4, 3/4)$ (with high probability).

For estimators that produce an accurate response with high probability (such as any well-concentrated estimator) we give a different, substantially simpler way to prove high probability bounds on the accuracy of the whole adaptive procedure. This allows us to bypass the known proofs of high-probability bounds that rely on relatively involved arguments (Dwork et al., 2014; Bassily et al., 2016; Rogers et al., 2016).

Our algorithm for answering verification queries (Theorem 1.4), is obtained by reducing the verification step to verification of two statistical queries for which the domain of the function is $\mathcal{X}^t$ and the expectation is estimated relative to $\mathcal{P}^t$. To further improve the sample complexity, we observe that it is possible to calibrate the algorithm for verifying statistical queries from (Dwork et al., 2015b) to introduce less noise when ρ is small. This improvement relies on sharper analysis of the known generalization properties of differential privacy that has already found some additional uses (Steinke and Ullman, 2017; Nissim and Stemmer, 2017) (see Section 4.1 for details).

Our algorithm for answering many queries (Theorem 1.5) is also obtained by a reduction to statistical queries over $\mathcal{X}^t$. In this case we use statistical queries to find a value in the interquartile interval of the estimator via a binary search.

Finally, we remark that the differentially private algorithms that we use to answer queries might also be of interest for applications in private data analysis. These algorithms demonstrate that meaningful privacy and error guarantees can be achieved without any assumptions on the sensitivity of estimators. From this point of view, our approach is an instance of subsample-and-aggregate technique of Nissim et al. (2007), where the approximate median algorithm is used for aggregation. We note that Smith (2011) used a somewhat related approach that also takes advantage of the concentration of the estimator around its mean during the aggregation step. His algorithm first clips the tails of the estimator's distribution via a differentially private quantile estimation and then uses simple averaging with Laplace noise addition as the aggregation step. His analysis is specialized to estimators that are approximately normally distributed and hence the results are not directly comparable with our more general setting.

2. Preliminaries

For $k \in \mathbb{N}$, we denote $[k] = \{1, 2, \ldots, k\}$ and we use $a_{[k]}$ as a shorthand for the k-tuple $(a_1, a_2, \ldots, a_k)$. For a condition E we use $\mathbb{1}(E)$ to denote the indicator function of E. Thus $\mathbf{E}\left[\mathbb{1}(E)\right] = \mathbf{Pr}\left[E\right]$.

For a randomized algorithm M we use $Y \sim M$ to denote that Y is random variable obtained by running M. For a distribution $\mathcal{D}$ we use $Y \sim \mathcal{D}$ to denote that Y is a random variable distributed according to $\mathcal{D}$. For $0 \le \alpha < \beta \le 1$ and a distribution $\mathcal{D}$, we denote the (α, β)-quantile interval of $\mathcal{D}$ by

$$\mathrm{qi}_{\mathcal{D}}(\alpha, \beta) \doteq \left\{ v \in \mathbb{R} \ : \ \underset{Y \sim \mathcal{D}}{\mathbf{Pr}}[Y \le v] > \alpha \ \wedge \ \underset{Y \sim \mathcal{D}}{\mathbf{Pr}}[Y < v] < \beta \right\}.$$

We refer to $\mathrm{qi}_{\mathcal{D}}(1/4, 3/4)$ as the interquartile interval of $\mathcal{D}$ and the length of this interval is the interquartile range. For $s \in \mathbb{R}^n$ we denote by $\mathrm{qi}_s(\alpha, \beta)$ the empirical version of the quantity above, that is the (α, β)-quantile interval obtained by taking $\mathcal{D}$ to be the uniform distribution over the elements of s. In general, we view datasets as being equivalent to a distribution, namely the uniform distribution on elements of that dataset.

For our algorithms, a query is specified by a function $\phi : \mathcal{X}^t \to \mathbb{R}$. For notational simplicity we will often set $\mathcal{Z} = \mathcal{X}^t$ and partition a dataset $s \in \mathcal{X}^n$ into m points $s_1, \ldots, s_m \in \mathcal{Z}$. Therefore throughout our discussion we will have $n = mt$. For $s \in \mathcal{Z}^m$, we define $\phi(s) = (\phi(s_1), \ldots, \phi(s_m))$ to be the transformed dataset. Similarly, for a distribution $\mathcal{P}$ on $\mathcal{X}$, we define $\mathcal{D} = \mathcal{P}^t$ to be the corresponding distribution on $\mathcal{Z}$ and $\phi(\mathcal{D})$ to be the distribution obtained by applying ϕ to a random sample from $\mathcal{D}$. The expectations of these distributions are denoted $s[\phi] = \frac{1}{m}\sum_{i \in [m]} \phi(s_i)$ and $\mathcal{D}[\phi] = \mathbf{E}_{Z \sim \mathcal{D}}[\phi(Z)]$.

2.1. Adaptivity

A central topic in this paper is the interaction between two algorithms A — the analyst (who might even be adversarial) — and M — our query-answering algorithm. Let $\mathcal{Q}$ be the space of all possible queries and $\mathcal{A}$ be the set of all possible answers. In our applications $\mathcal{Q}$ will be a set of functions $\phi : \mathcal{Z} \to \mathbb{R}$, possibly with some additional parameters and $\mathcal{A}$ will be (a subset of) $\mathbb{R}$. We now set up the notation for this interaction.

```
Input $s \in \mathcal{Z}^m$ is given to $M$.
For $j = 1, 2, \ldots, k$:
  $A$ computes a query $q_j \in \mathcal{Q}$ and passes it to $M$
  $M$ produces answer $a_j \in \mathcal{A}$ and passes it to $A$
The output is the transcript $(q_1, q_2, \ldots, q_k, a_1, a_2, \ldots, a_k) \in \mathcal{Q}^k \times \mathcal{A}^k$.
```

Figure 1: $A \overset{\rightarrow}{\underset{\leftarrow}{}} M : \mathcal{Z}^m \to \mathcal{Q}^k \times \mathcal{A}^k$

Given interactive algorithms A and M, we define $A \overset{\rightarrow}{\underset{\leftarrow}{}} M(s)$ to be function which produces a random transcript of the interaction between A and M, where s is the input given to M. Formally, Figure 1 specifies how $A \overset{\rightarrow}{\underset{\leftarrow}{}} M : \mathcal{X}^n \to \mathcal{Q}^k \times \mathcal{A}^k$ is defined.

The transcript function $A \overset{\rightarrow}{\underset{\leftarrow}{}} M$ provides a "non-interactive view" of the output of an interactive process. Our goal is thus to construct M such that, for all A, the output of $A \overset{\rightarrow}{\underset{\leftarrow}{}} M$ has the desired accuracy and stability properties.

2.2. Differential Privacy

We begin with the standard definition of differential privacy (Dwork et al., 2006a,b).

Definition 2.1 (Differential Privacy) *An algorithm $M : \mathcal{Z}^m \to \mathcal{Y}$ is (ε, δ)-differentially private if, for all datasets $s, s' \in \mathcal{Z}^m$ that differ on a single element,*

$$\forall E \subseteq \mathcal{Y} \qquad \mathbf{Pr}\left[M(s) \in E\right] \le e^{\varepsilon}\, \mathbf{Pr}\left[M(s') \in E\right] + \delta.$$

Note that, throughout, we consider an algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ and let $\mathcal{Z} = \mathcal{X}^t$ with $n = mt$ so that $M : \mathcal{Z}^m \to \mathcal{Y}$. Then we define differential privacy with respect to the latter view (that is, with respect to changing a whole tuple of t elements in the original view of M). This is a stronger condition.

However, Definition 2.1 only covers non-interactive algorithms. Thus we extend it to interactive algorithms:

Definition 2.2 (Interactive Differential Privacy) *An interactive algorithm M is (ε, δ)-differentially private if, for all interactive algorithms A, the (non-interactive) algorithm $A \overset{\rightarrow}{\leftarrow} M : \mathcal{Z}^m \to \mathcal{Y}$ is (ε, δ)-differentially private.*

We now record the basic properties of differential privacy. See the textbook of Dwork and Roth (2014) for proofs and discussion of these results.

Theorem 2.3 (Postprocessing) *Let $M : \mathcal{Z}^m \to \mathcal{Y}$ be (ε, δ)-differentially private. Let $F : \mathcal{Y} \to \mathcal{Y}'$ be an arbitrary randomized algorithm. Define $M' : \mathcal{Z}^m \to \mathcal{Y}'$ by $M'(s) = F(M(s))$. Then M' is also (ε, δ)-differentially private.*

Postprocessing is important as it allows us to perform further computation on the output of a differentially private algorithm without breaking the privacy guarantee.

We next state the key adaptive composition property of differential privacy, which bounds how rapidly differential privacy degrades under repeated use of the same dataset (Dwork et al., 2010) (with sharper constants from (Bun and Steinke, 2016)).

Theorem 2.4 (Adaptive Composition (Dwork et al., 2010; Bun and Steinke, 2016)) *Fix $k \in \mathbb{N}$ and $\varepsilon_1, \ldots, \varepsilon_k, \delta_1, \ldots, \delta_k > 0$. Let $M_1, \ldots, M_k : \mathcal{Z}^m \times \mathcal{Y} \to \mathcal{Y}$ be randomized algorithms. Suppose that for all $j \in [k]$ and all fixed $y \in \mathcal{Y}$, the randomized algorithm $x \mapsto M_j(x, y)$ is $(\varepsilon_j, \delta_j)$-differentially private. Define $\hat{M}_1, \ldots, \hat{M}_k : \mathcal{Z}^m \to \mathcal{Y}$ inductively by $\hat{M}_1(x) = M_1(x, y_0)$ where $y_0 \in \mathcal{Y}$ is fixed and $\hat{M}_{j+1}(x) = M_{j+1}(x, \hat{M}_j(x))$ for $j \in [k-1]$. Then $\hat{M}_k$ is $(\hat{\varepsilon}, \hat{\delta})$-differentially private for*

$$\hat{\varepsilon} = \frac{1}{2} \sum_{j \in [k]} \varepsilon_j^2 + \sqrt{2 \log(1/\delta') \sum_{j \in [k]} \varepsilon_j^2} \qquad \textit{and} \qquad \hat{\delta} = \delta' + \sum_{j \in [k]} \delta_j,$$

where $\delta' \in (0, 1)$ is arbitrary.

The analyst that asks queries can be seen as a postprocessing step on the output of a differentially private algorithm that answers the queries. Thus, by combining the adaptive composition and postprocessing properties of differential privacy we obtain that in order to ensure that an interactive algorithm is differentially private it is sufficient to ensure that each of the individual queries is answered with differential privacy.

Theorem 2.5 *Fix $k \in \mathbb{N}$ and $\varepsilon, \delta > 0$. Let $M : \mathcal{Z}^m \times \mathcal{Q} \to \mathcal{A}$ be an algorithm, such that $M(s, q)$ provides an answer to the query q using the dataset s and M is (ε, δ)-differentially private for every fixed $s \in \mathcal{Z}^m$.*

Define an interactive algorithm $M^{\otimes k}$ that takes as input $s \in \mathcal{Z}^m$ and answers k adaptively-chosen queries $q_1, \ldots, q_k \in \mathcal{Q}$ where, for each $j \in [k]$, $M^{\otimes k}$ produces an answer by independently running $M(s, q_j)$. Then $M^{\otimes k}$ is $\left(\frac{1}{2}k\varepsilon^2 + \varepsilon\sqrt{2k\ln(1/\delta')}, \delta' + k\delta\right)$-differentially private for all $\delta' \in (0, 1)$.

3. Approximate Median

In this section, we present differentially private algorithms for outputting an approximate median of a real-valued dataset. Namely, for $s \in \mathbb{R}^m$, we define an α-approximate median of s to be any element of the set $\mathrm{qi}_s\left(\frac{1-\alpha}{2}, \frac{1+\alpha}{2}\right)$. In our application each real value is obtained by applying the given query function to a single subsample.

Several differentially private algorithms for computing an approximate median are known. All of these algorithms assume that the input elements and the range of the algorithm are restricted to some finite set $T \subseteq \mathbb{R}$. The strongest upper bound was given in a recent work of Bun et al. (2015). They describe an (ε, δ)-differentially private algorithm which, on input $s \in T^m$, outputs an α-approximate median of s as long as $m \geq (2 + o(1))^{\log^* |T|} \cdot O(\log(1/\varepsilon\delta)/\varepsilon\alpha)$.[5]

Bun et al. (2015) also prove a nearly tight lower bound of $m \geq \Omega(\log^* |T|)$ for $\alpha = \varepsilon = 1$ and $\delta = 1/100m^2$. This lower bound implies that privately outputting an approximate median is only possible if we restrict the data points to a finite range. We note that it is also known that for the stricter ε-differential privacy the sample complexity of this problem for constant $\alpha > 0$ is $\theta(\log(|T|)/\varepsilon)$ (*e.g.* Bun et al., 2015, 2017).

Any differentially private algorithm for finding an approximate median can be used in our results. The algorithm in (Bun et al., 2015) is relatively involved and hence we will describe a simple algorithm for the problem that relies on a "folklore" application of the exponential mechanism (McSherry and Talwar, 2007) (*e.g.* Smith, 2011; Nissim and Sheffet, 2014).

Theorem 3.1 *For all $\varepsilon, \alpha, \beta > 0$, finite $T \subset \mathbb{R}$, and all $m \geq 4\ln(|T|/\beta)/\varepsilon\alpha$, there exists an $(\varepsilon, 0)$-differentially private randomized algorithm M that given a dataset $s \in \mathcal{Z}^m$ and a query $\phi : \mathcal{Z} \to T$ outputs an α-approximate median of $\phi(s) \in T^m$ with probability at least $1 - \beta$. The running time of the algorithm is $O(m \cdot \log |T|)$.*

Proof The algorithm is an instantiation of the exponential mechanism (McSherry and Talwar, 2007) with the utility function $c : T \to \mathbb{R}$ defined as

$$c_{\phi(s)}(v) \doteq \max\left\{|\{i \in [m] : \phi(s_i) < v\}|, |\{i \in [m] : \phi(s_i) > v\}|\right\}.$$

The algorithm outputs each $v \in T$ with probability

$$\mathbf{Pr}\left[M(s, \phi) = v\right] = \frac{\exp\left(\frac{-\varepsilon}{2} c_{\phi(s)}(v)\right)}{\sum_{u \in T} \exp\left(\frac{-\varepsilon}{2} c_{\phi(s)}(u)\right)}.$$

[5] Bun et al. (2015) consider the problem of outputting an interior point which is equivalent to our definition of a 1-approximate median. However, by removing the elements of the dataset that are not in $\mathrm{qi}_s\left(\frac{1-\alpha}{2}, \frac{1+\alpha}{2}\right)$ α-approximate median reduces to the interior point problem.

Since $c_{\phi(s)}(v)$ has sensitivity 1 as a function of s, this algorithm is $(\varepsilon, 0)$-differentially private (Dwork and Roth, 2014, Theorem 3.10). Moreover, we have the accuracy guarantee (Dwork and Roth, 2014, Corollary 3.12)

$$\forall s, \beta \qquad \mathop{\mathbf{Pr}}_{V \sim M(s,\phi)} \left[c_{\phi(s)}(V) < \mathrm{OPT}_{\phi(s)} + \frac{2\ln(|T|/\beta)}{\varepsilon} \right] \geq 1 - \beta, \tag{1}$$

where $\mathrm{OPT}_{\phi(s)} \doteq \min_{u \in T} c_{\phi(s)}(u) \leq m/2$. Assuming the event in (1) happens for $V = v$ (that is, $c_{\phi(s)}(v) < \mathrm{OPT}_{\phi(s)} + \frac{2\ln(|T|/\beta)}{\varepsilon}$), we have

$$\mathop{\mathbf{Pr}}_{Y \sim \phi(s)} [Y \leq v] = 1 - \frac{1}{m} |\{ i \in [m] : \phi(s_i) > v \}| > \frac{1}{2} - \frac{2\ln(|T|/\beta)}{\varepsilon m} \geq \frac{1-\alpha}{2},$$

as long as $\alpha \geq 4\ln(|T|/\beta)/\varepsilon m$, which is equivalent to $m \geq 4\ln(|T|/\beta)/\varepsilon\alpha$. Similarly, the event in (1) implies that

$$\mathop{\mathbf{Pr}}_{Y \sim \phi(s)} [Y < v] = \frac{1}{m} |\{ i \in [m] : \phi(s_i) < v \}| < \frac{1}{2} + \frac{2\ln(|T|/\beta)}{\varepsilon m} \leq \frac{1+\alpha}{2}.$$

Thus

$$\mathop{\mathbf{Pr}}_{V \sim M(s,\phi)} \left[V \in \mathrm{qi}_{\phi(s)} \left(\frac{1-\alpha}{2}, \frac{1+\alpha}{2} \right) \right] \geq 1 - \beta$$

as long as $m \geq 4\ln(|T|/\beta)/\varepsilon\alpha$.

To get an upper bound on the running time we observe that using binary search, the elements of T can be split into $m+1$ "intervals" (that is contiguous subsets of T) with all elements of each interval having equal probability. This partition allows us to compute the normalization factor as well as the total probability of all the elements of T in each interval in $O(m \log |T|)$ time. A random point from the desired distribution can now be produced by first picking the interval proportionally to its probability and then outputting a point in that interval randomly and uniformly. (We implicitly assume that the structure of T is simple enough so that such operations can be performed in $O(\log |T|)$ time and ignore the time to evaluate ϕ on each of the elements of s.) ■

We now describe another simple and "folklore" algorithm (*e.g.* Raskhodnikova and Smith, 2010; Feldman, 2016) for finding an approximate median of a distribution that reduces the problem to $O(\log |T|)$ statistical queries. Recall, that an α-accurate response to a statistical query $\psi : \mathcal{Z} \to [-1, 1]$ relative to distribution $\mathcal{D}$ over $\mathcal{Z}$ is any value v such that $|v - \mathcal{D}[\psi]| \leq \alpha$.

Lemma 3.2 *For all $\alpha > 0$, finite $T \subset \mathbb{R}$, a query $\phi : \mathcal{Z} \to T$ and any distribution $\mathcal{D}$ over T, a value $v \in \mathrm{qi}_{\phi(\mathcal{D})} \left(\frac{1-\alpha}{2}, \frac{1+\alpha}{2} \right)$ can be found using $\alpha/4$-accurate responses to at most $2\lceil \log_2(|T|) \rceil$ (adaptively-chosen) statistical queries relative to distribution $\mathcal{D}$.*

Proof Using binary search we find a point $v \in T$ that satisfies the conditions $p_{\leq}(v) > 1/2 - \alpha/4$ and $p_{<}(v) < 1/2 + \alpha/4$, where $p_{\leq}(v)$ (or $p_{<}(v)$) is the response to the statistical query $\psi(z) = \mathbb{1}(v \leq \phi(z))$ ($\psi(z) = \mathbb{1}(v < \phi(z))$, respectively). By the accuracy guarantees of the responses, we have that $|p_{\leq}(v) - \mathbf{Pr}_{Z \sim \mathcal{D}}[\phi(Z) \leq v]| \leq \alpha/4$, and similarly for $p_{<}(v)$.

We choose the next point to test depending on which of the conditions fails (note that we can assume that $p_<(v) \leq p_\leq(v)$ so at most one condition can fail). Further, for the true median point of $\phi(\mathcal{D})$ (that is the point $v^* \in T$ for which $\mathbf{Pr}_{Z\sim\mathcal{D}}[\phi(Z) < v^*] < 1/2$ and $\mathbf{Pr}_{Z\sim\mathcal{D}}[\phi(Z) \leq v^*] \geq 1/2$) both conditions will be satisfied by the accuracy guarantees. Finally, by the accuracy guarantees, any point v' that satisfies both of these conditions is an α-approximate median for $\phi(\mathcal{D})$. ■

To find the α-approximate median of values $\phi(s) \in T^m$ this reduction needs to be applied to the uniform distribution over the elements of $\phi(s)$. Answering statistical queries relative to this empirical distribution (commonly referred to as linear or counting queries) with differential privacy is a well-studied problem. For example, by using the standard Laplace or Gaussian noise addition algorithm one can obtain the following algorithm for finding an α-approximate median (see (Bun and Steinke, 2016) for an analysis of the privacy properties of Gaussian noise addition).

Corollary 3.3 *For all $\varepsilon, \delta, \alpha, \beta \in (0, 1/2)$, finite $T \subset \mathbb{R}$ and all*

$$m \geq \frac{12\sqrt{2\lceil\log_2|T|\rceil \cdot \ln(1/\delta) \cdot \ln(2\lceil\log_2|T|\rceil/\beta)}}{\varepsilon\alpha} = O\left(\frac{\sqrt{\log|T| \cdot \log(1/\delta) \cdot \log\left(\frac{\log|T|}{\beta}\right)}}{\varepsilon\alpha}\right),$$

there exists an (ε, δ)-differentially private randomized algorithm M that given $s \in \mathcal{Z}^m$ and $\phi : \mathcal{Z} \to T$ outputs an α-approximate median of $\phi(s)$ with probability at least $1 - \beta$. The running time of the algorithm is $O(m \cdot \log|T|)$.

4. Generalization from Differential Privacy

In this section we provide two proofs that differential privacy gives generalization guarantees for statistical queries. The first proof — which we call strong generalization — is most similar to previous work, whereas the second proof — which we call simple generalization — is much simpler, but gives a weaker bound that is only suitable for estimators that are well-concentrated.

4.1. Strong Generalization

Theorem 4.1 in this subsection shows that any differentially private algorithm generalizes with high probability, with a small blowup in the allowed generalization error. This proof closely follows that of Bassily et al. (2016), but is quantitatively sharper. This quantitative sharpening allows us to estimate the probability that a given value v is outside of $\mathrm{qi}_{\phi(\mathcal{D})}(\rho, 1-\rho)$ with higher accuracy (that scales with ρ as in the non-adaptive case). We use this sharper version in Section 5.1; however, for results in this section $\rho = 1/4$ and therefore the bounds from (Bassily et al., 2016) suffice.

The *mean absolute deviation* of a distribution $\mathcal{D}$ over $\mathbb{R}$ is defined as

$$\mathrm{mad}(\mathcal{D}) \doteq \mathop{\mathbf{E}}_{Y'\sim\mathcal{D}}\left[\left|Y' - \mathop{\mathbf{E}}_{Y\sim\mathcal{D}}[Y]\right|\right]. \tag{2}$$

Theorem 4.1 *Fix $\alpha, \beta, \gamma \in (0,1)$ and $m, k \in \mathbb{N}$. Set $\varepsilon = \frac{1}{2}\ln(1+\gamma)$ and $\delta = \alpha\beta/16$. Suppose $m \geq \frac{8}{\varepsilon\alpha}\ln(2k/\beta)$. Let $M : \mathcal{Z}^m \to \mathcal{F}_{[0,1]}^k$ be a (ε, δ)-differentially private algorithm with $\mathcal{F}_{[0,1]}$ being the set of functions $\phi : \mathcal{Z} \to [0,1]$. Let $\mathcal{D}$ be a distribution on $\mathcal{Z}$. Then*

$$\underset{\substack{S\sim\mathcal{D}^m \\ \phi_{[k]}\sim M(S)}}{\mathbf{Pr}} \left[\forall j \in [k] \quad S[\phi_j] - \mathcal{D}[\phi_j] \leq \alpha + \gamma \cdot \mathrm{mad}(\phi_j(\mathcal{D}))\right] \geq 1-\beta.$$

Theorem 4.1 is proved in Appendix A.

Note that, by Jensen's inequality, $\mathrm{mad}(\phi(\mathcal{D})) \leq \mathrm{sd}(\phi(\mathcal{D}))$. Thus Theorem 4.1 gives an error bound that scales with the standard deviation of the query (plus the absolute α term). Also, by the triangle inequality and the fact that $\phi(z) \geq 0$ for all z, it holds that

$$\mathrm{mad}(\phi(\mathcal{D})) \leq 2 \cdot \mathcal{D}[\phi].$$

Thus Theorem 4.1 can also be interpreted as giving a multiplicative accuracy guarantee (plus the additive α). In comparison, the bound of Bassily et al. (2016) can be obtained (up to constants) by substituting the upper bound $\mathrm{mad}(\phi(\mathcal{D})) \leq 1$ into Theorem 4.1. Thus, when $\mathrm{mad}(\phi(\mathcal{D})) \ll 1$, our bound is sharper.

As stated, Theorem 4.1 only applies in the non-adaptive setting and to statistical queries. However, we can easily extend it using the monitor technique of Bassily et al. (2016) and the cumulative probability function:

Theorem 4.2 *Fix $\beta \in (0,1)$ and $k, m \in \mathbb{N}$ with $m \geq 2560\ln(2k/\beta)$. Let M be an $(1/20, \beta/256)$-differentially private interactive algorithm that takes as input $s \in \mathcal{Z}^m$ and provides answers $v_1, \ldots, v_k \in \mathbb{R}$ to an adaptively-chosen sequence of queries $\phi_1, \ldots, \phi_k : \mathcal{Z} \to \mathbb{R}$. Suppose that, for all $s \in \mathcal{Z}^m$ and all interactive algorithms A,*

$$\underset{(\phi_{[k]},v_{[k]})\sim A \rightleftarrows M(s)}{\mathbf{Pr}} \left[\forall j \in [k] \quad v_j \in \mathrm{qi}_{\phi_j(s)}\left(\frac{3}{8}, \frac{5}{8}\right)\right] \geq 1-\beta. \tag{3}$$

Then, for all distributions $\mathcal{D}$ and all interactive algorithms A,

$$\underset{\substack{S\sim\mathcal{D}^m \\ (\phi_{[k]},v_{[k]})\sim A \rightleftarrows M(S)}}{\mathbf{Pr}} \left[\forall j \in [k] \quad v_j \in \mathrm{qi}_{\phi_j(\mathcal{D})}\left(\frac{1}{4}, \frac{3}{4}\right)\right] \geq 1-2\beta.$$

Proof Let $\mathcal{Q}$ be the set of functions $\phi : \mathcal{Z} \to \mathbb{R}$ and let A be an arbitrary algorithm that asks queries in $\mathcal{Q}$. We define $f : \mathcal{Q}^k \times \mathbb{R}^k \to \mathcal{F}_{[0,1]}$ as follows, where $\mathcal{F}_{[0,1]}$ is the set of functions $\psi : \mathcal{Z} \to [0,1]$. Given $(\phi, v) \in \mathcal{Q}^k \times \mathbb{R}^k$, define $\psi_1, \psi_{-1}, \psi_2, \psi_{-2}, \ldots, \psi_k, \psi_{-k} : \mathcal{Z} \to \{0,1\}$ by

$$\psi_j(x) \doteq \mathbb{1}(\phi_j(x) \leq v_j) \quad \text{and} \quad \psi_{-j}(x) \doteq \mathbb{1}(\phi_j(x) \geq v_j)$$

and let

$$f(\phi, v) \doteq \underset{\psi\in\{\psi_1,\psi_{-1},\psi_2,\psi_{-2},\ldots,\psi_k,\psi_{-k}\}}{\mathrm{argmin}} \mathcal{D}[\psi].$$

By the postprocessing property of differential privacy (Theorem 2.3), $f(A \rightleftarrows M(s))$ is a (ε, δ)-differentially private algorithm (relative to its input $s \in \mathcal{Z}^m$). Moreover, by our assumption (3),

$$\forall s \in \mathcal{Z}^m \qquad \underset{\psi\sim f\left(A \rightleftarrows M(s)\right)}{\mathbf{Pr}} \left[s[\psi] \geq \frac{3}{8}\right] \geq 1-\beta.$$

However, by Theorem 4.1,

$$\mathop{\mathbf{Pr}}_{\substack{S\sim\mathcal{D}^m\\ \psi\sim f(A\rightleftarrows M(S))}}\left[S[\psi]-\mathcal{D}[\psi]\le\frac{1}{8}\right]\ge\mathop{\mathbf{Pr}}_{\substack{S\sim\mathcal{D}^m\\ \psi\sim f(A\rightleftarrows M(S))}}\left[S[\psi]-\mathcal{D}[\psi]\le\frac{1}{16}+\frac{1}{8}\cdot\mathrm{mad}(\psi(\mathcal{D}))\right]\ge 1-\beta.$$

Thus, by a union bound and the construction of f,

$$\mathop{\mathbf{Pr}}_{\substack{S\sim\mathcal{D}^m\\ (\phi_{[k]},v_{[k]})\sim A\rightleftarrows M(S)}}\left[\forall j\in[k]\quad v_j\in\mathrm{qi}_{\phi_j(\mathcal{D})}\left(\frac{1}{4},\frac{3}{4}\right)\right]=\mathop{\mathbf{Pr}}_{\substack{S\sim\mathcal{D}^m\\ \psi\sim f(A\rightleftarrows M(S))}}\left[\mathcal{D}[\psi]\ge\frac{1}{4}\right]\ge 1-2\beta.$$

■

Combining generalization (Theorem 4.2) with our approximate median algorithm (Theorem 3.1) and composition (Theorem 2.4) yields our main result, Theorem 1.3. We prove a somewhat more general statement that allows using different range T_j for every query ϕ_j. (The same generalization applies to all our other results, but we do not state it for brevity).

Theorem 4.3 *For any $\beta\in(0,1)$, $t,k,r\in\mathbb{N}$ and $\mathcal{Z}=\mathcal{X}^t$, and with*

$$n\ge n_0=O\left(t\sqrt{k\log(1/\beta)}\cdot\log(kr/\beta)\right)$$

there exists an interactive algorithm M that takes as input a dataset $s\in\mathcal{X}^n$ and provides answers $v_1,\ldots,v_k\in\mathbb{R}$ to adaptively-chosen queries $(T_1,\phi_1),\ldots,(T_k,\phi_k)$, where for all $j\in[k]$, $|T_j|\le r$ and $\phi_j:\mathcal{X}^t\to T_j$ with the following accuracy guarantee. For all interactive algorithms A and distributions $\mathcal{P}$ on $\mathcal{X}$,

$$\mathop{\mathbf{Pr}}_{\substack{S\sim\mathcal{P}^n\\ \left(T_{[k]},\phi_{[k]},v_{[k]}\right)\sim A\rightleftarrows M(S)}}\left[\forall j\in[k]\quad v_j\in\mathrm{qi}_{\phi_j(\mathcal{P}^t)}\left(\frac{1}{4},\frac{3}{4}\right)\right]\ge 1-\beta.$$

Proof The algorithm M promised by Theorem 4.3 is described in Figure 2.

```
Input $S\in\mathcal{X}^{mt}$.
Partition $S$ into $S_1,\ldots,S_m\in\mathcal{X}^t$.
For $j=1,2,\ldots,k$:
    Receive a set $T_j$ and a query $\phi_j:\mathcal{X}^t\to T_j$.
    Run the $(\tilde{\varepsilon},0)$-differentially private 1/4-approximate median algorithm $\tilde{M}$ from
Thm. 3.1 for $T_j$ and with inputs $(S_1,\ldots,S_m)$ and $\phi_j$ to obtain output $v_j\in T_j$.
    Return answer $v_j$.
```

Figure 2: Algorithm for answering adaptive queries.

Let $\mathcal{Z}\doteq\mathcal{X}^t$ and assume that for some r fixed in advance, $r\ge\max_{j\in[k]}|T_j|$. Theorem 3.1 says that if $m\ge 4\ln(kr/\beta)/(\alpha\tilde{\varepsilon})$, then each execution of the median algorithm is $(\tilde{\varepsilon},0)$-differentially private and outputs an α-approximate median with probability at least $1-\beta/k$.

Here $\alpha = 1/4$, so this rearranges to $\tilde{\varepsilon} = 16\ln(kr/\beta)/m$. This implies that for all interactive algorithms A and every $s \in \mathcal{Z}^m$,

$$\Pr_{(T_{[k]},\phi_{[k]},v_{[k]})\sim A \underset{\leftarrow}{\rightarrow} M(s)} \left[\forall j \in [k] \quad v_j \in \mathrm{qi}_{\phi_j(s)}\left(\frac{3}{8},\frac{5}{8}\right)\right] \geq 1-\beta. \tag{4}$$

Interactive composition (Theorem 2.5) implies that M is (ε,δ)-differentially private for any $\delta \in (0,1)$ and

$$\varepsilon = \frac{k}{2}\left(\frac{16\ln(kr/\beta)}{m}\right)^2 + \frac{16\ln(kr/\beta)}{m}\sqrt{2k\ln(1/\delta)}. \tag{5}$$

By Theorem 4.2, if, in addition to (4), we have $\varepsilon \leq 1/20$ for $\delta = \beta/256$ in (5) and $m \geq 2560\ln(2k/\beta)$, then, for all distributions $\mathcal{P}$, $\mathcal{D} \doteq \mathcal{P}^t$ and all interactive algorithms A,

$$\Pr_{\substack{S\sim\mathcal{D}^m \\ \left(T_{[k]},\phi_{[k]},v_{[k]}\right)\sim A \underset{\leftarrow}{\rightarrow} M(S)}} \left[\forall j \in [k] \quad v_j \in \mathrm{qi}_{\phi_j(\mathcal{D})}\left(\frac{1}{4},\frac{3}{4}\right)\right] \geq 1-2\beta,$$

which is our desired conclusion.

It only remains to find the appropriate bound on the parameter m. We need $m \geq 2560\ln(2k/\beta)$ and

$$\frac{1}{20} \geq \varepsilon = \frac{k}{2}\left(\frac{16\ln(kr/\beta)}{m}\right)^2 + \frac{16\ln(kr/\beta)}{m}\sqrt{2k\ln(256/\beta)}.$$

Setting $m = 640\sqrt{\max\{k,16\}\cdot\ln(256/\beta)}\cdot\ln(kr/\beta)$ achieves this. ■

We now state a simple corollary of Theorem 4.3 that converts the $(1/4,3/4)$-quantile interval guarantees to explicit additive error guarantees. The error will be measured in terms of the mean absolute deviation of the query ϕ on inputs sampled from $\mathcal{P}^t$ (eq. (2)). For normalization purposes we will also assume that queries are scaled by the analyst is such a way that both $\mathcal{P}^t[\phi] \in [-1,1]$ and $\mathrm{mad}(\phi(\mathcal{P}^t)) \leq 1$. Note that this assumption is implied by ϕ having range $[-1,1]$ and, in general, allows ϕ to have an infinite range.

Corollary 4.4 *For $t \in \mathbb{N}$ and a distribution $\mathcal{P}$ over $\mathcal{X}$, let $\mathcal{F}_{\mathcal{P},t}$ denote the set of functions $\phi : \mathcal{X}^t \to \mathbb{R}$ such that $\mathcal{P}^t[\phi] \in [-1,1]$ and $\mathrm{mad}(\phi(\mathcal{P}^t)) \leq 1$. For all $\zeta > 0$, $\beta > 0$, $k \in \mathbb{N}$, and $n \geq n_0 = O\left(t\sqrt{k\log(1/\beta)}\cdot\log(k/(\zeta\beta))\right)$, there exists an efficient algorithm M which takes a dataset $s \in \mathcal{X}^n$ as an input and provides answers $v_1,\ldots,v_k \in \mathbb{R}$ to an adaptively-chosen sequence of queries $\phi_1,\ldots,\phi_k : \mathcal{X}^t \to \mathbb{R}$ satisfying: for all interactive algorithms A and distributions $\mathcal{P}$ over $\mathcal{X}$,*

$$\Pr_{\substack{S\sim\mathcal{P}^n \\ (\phi_{[k]},v_{[k]})\sim A \underset{\leftarrow}{\rightarrow} M(S)}} \left[\forall j \in [k] \text{ s.t. } \phi_j \in \mathcal{F}_{\mathcal{P},t} : \quad \left|v_j - \mathcal{P}^t[\phi_j]\right| \leq 4\cdot(\phi_j(\mathcal{P}^t)) + \zeta\right] \geq 1-\beta.$$

Proof We first observe that by Markov's inequality,

$$\Pr_{Z\sim\mathcal{P}^t}\left[|\phi(Z) - \mathcal{P}^t[\phi]| \geq 4\cdot\mathrm{mad}(\phi(\mathcal{P}^t))\right] \leq 1/4.$$

Therefore

$$\mathrm{qi}_{\phi(\mathcal{P}^t)}\left(\frac{1}{4},\frac{3}{4}\right) \subseteq \left[\mathcal{P}^t[\phi] - 4\cdot \mathrm{mad}(\phi(\mathcal{P}^t)), \mathcal{P}^t[\phi] + 4\cdot \mathrm{mad}(\phi(\mathcal{P}^t))\right]. \tag{6}$$

Hence for all $j \in [k]$ such that $\phi_j \in \mathcal{F}_{\mathcal{P},t}$, we have that $\mathrm{qi}_{\phi_j(\mathcal{P}^t)}(1/4, 3/4) \subseteq [-5,5]$. Now we define T to be the interval $[-5,5]$ discretized with step ζ, or $T \doteq \{r\cdot\zeta : r \in \mathbb{Z}\} \bigcap [-5,5]$. To answer a query ϕ_j we define $\phi'_j : \mathcal{X}^t \to T$ as $\phi'_j(z) \doteq \mathrm{argmin}_{v\in T} |v - \phi_j(z)|$ and then use the algorithm from Theorem 4.3 to answer the query ϕ'_j. The projection of the values of ϕ to T simultaneously truncates the range to $[-5,5]$ and discretizes it. The $(1/4, 3/4)$-quantile interval of $\phi_j(\mathcal{P}^t)$ is inside the interval $[-5,5]$ and therefore is not affected by the truncation step. The discretization can affect this interval by at most ζ. Combining this with (6) we obtain that if $\phi_j \in \mathcal{F}_{\mathcal{P},t}$ then

$$\mathrm{qi}_{\phi'_j(\mathcal{P}^t)}\left(\frac{1}{4},\frac{3}{4}\right) \subseteq \left[\mathcal{P}^t[\phi_j] - 4\cdot \mathrm{mad}(\phi_j(\mathcal{P}^t)) - \zeta, \mathcal{P}^t[\phi_j] + 4\cdot \mathrm{mad}(\phi_j(\mathcal{P}^t)) + \zeta\right].$$

Therefore the value v_j returned by the algorithm from Theorem 4.3 to query ϕ'_j satisfies:

$$|v_j - \mathcal{P}^t[\phi_j]| \leq 4\cdot \mathrm{mad}(\phi_j(\mathcal{P}^t)) + \zeta.$$

Now to obtain the claimed bound on the sample complexity we observe that $|T| \leq 10/\zeta$. ■

Remark 1 *Note that mean absolute deviation of ϕ is upper-bounded by the standard deviation of ϕ. Therefore Corollary 4.4 also holds with $\mathrm{mad}(\phi_j(\mathcal{P}^t))$ replaced by $\mathrm{sd}(\phi_j(\mathcal{P}^t))$ both in the definition of $\mathcal{F}_{t,\mathcal{P}}$ and the accuracy bound (with the constant factor 4 being replaced by 2 since Chebyshev's inequality can be used instead of Markov's). The obtained statement generalizes Theorem 1.2 we stated in the introduction.*

The quantile-based guarantees of our other algorithms can be converted to additive error guarantees in an analogous way.

We remark that somewhat sharper (asymptotic) bounds can be obtained by using the approximate median algorithm based on linear queries (Lemma 3.2) together with the algorithm for answering linear queries by Steinke and Ullman (2016). Specifically, this algorithm can solve the problem given $n = O\left(t\sqrt{k\cdot\log(1/\zeta)\cdot\log(1/\beta)\cdot\log(\log(k\log(1/\zeta))/\beta)}\right)$ samples.

4.2. Simple Generalization

We now describe a simple and seemingly weak generalization result that shows the output of a differentially private algorithm cannot "overfit" its input dataset. Namely, if an (ε,δ) differentially private algorithm outputs a function $\phi : \mathcal{Z} \to \mathbb{R}$ on a dataset $s \in \mathcal{Z}^m$ sampled from $\mathcal{D}^m$, then the value of ϕ on any element of the dataset is within the $(\rho, 1-\rho)$-quantile interval of $\phi(\mathcal{D})$ with probability at least $1 - (2e^\varepsilon\rho + \delta)$. To obtain meaningful guarantees about the whole dataset from such generalization result, ρ must be relatively small (much smaller than the desired 1/4). The good news is that for an estimator that is well-concentrated around its mean, even values within $(\rho, 1-\rho)$-quantile interval for small ρ are close to the mean. Note that, in principle, any estimator can be amplified by sampling and taking a median before being used in this analysis and hence we can obtain generalization guarantees from such simple analysis even in the general case (although the algorithm in this case would need to use two median steps).

Theorem 4.5 *Let $M : \mathcal{Z}^m \to \mathcal{F}_{[0,1]}$ be a (ε, δ)-differentially private algorithm with $\mathcal{F}_{[0,1]}$ being the set of functions $\phi : \mathcal{Z} \to [0, 1]$. Let $\mathcal{D}$ be a distribution on $\mathcal{Z}$. Then for all $i \in [m]$,*

$$\mathop{\mathbf{Pr}}_{\substack{S\sim\mathcal{D}^m \\ \phi\sim M(S)}} \left[\phi(S_i) \notin \mathrm{qi}_{\phi(\mathcal{D})}\left(\rho, 1-\rho\right)\right] \leq 2\rho e^{\varepsilon} + \delta$$

Proof By differential privacy, for all $i \in [m]$,

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{\substack{S\sim\mathcal{D}^m \\ \phi\sim M(S)}} \left[\phi(S_i) \notin \mathrm{qi}_{\phi(\mathcal{D})}\left(\rho, 1-\rho\right)\right] \leq & e^{\varepsilon} \mathop{\mathbf{Pr}}_{\substack{(S,Z)\sim\mathcal{D}^m\times\mathcal{D} \\ \phi\sim M(S_{-i},Z)}} \left[\phi(S_i) \notin \mathrm{qi}_{\phi(\mathcal{D})}\left(\rho, 1-\rho\right)\right] + \delta \\
= & e^{\varepsilon} \mathop{\mathbf{Pr}}_{\substack{(S,Z)\sim\mathcal{D}^m\times\mathcal{D} \\ \phi\sim M(S)}} \left[\phi(Z) \notin \mathrm{qi}_{\phi(\mathcal{D})}\left(\rho, 1-\rho\right)\right] + \delta \\
\leq & e^{\varepsilon} 2\rho + \delta,
\end{aligned}$$

where the equalities follow from the fact that the pairs (S, Z) and $((S_{-i}, Z), Z_i)$ are identically distributed and the definition of the $(\rho, 1-\rho)$-quantile interval. ■

Now for ρ and δ that are sufficiently small, Theorem 4.5 ensures that with probability at least $1-\beta$, for all $i \in [m]$, $\phi(s_i) \in \mathrm{qi}_{\phi(\mathcal{D})}\left(\rho, 1-\rho\right)$. This means that to get a value in $\mathrm{qi}_{\phi(\mathcal{D})}\left(\rho, 1-\rho\right)$, we can use an algorithm that outputs a value that is in between the smallest and the largest values of ϕ on the elements of a dataset s. Such value is referred to as an interior point of $\phi(s)$ (and is equivalent to a 1-approximate median).

This argument gives the following theorem.

Theorem 4.6 *For any $\beta \in (0, 1)$, $t, k \in \mathbb{N}$, a finite set $T \subset \mathbb{R}$ and $\mathcal{Z} = \mathcal{X}^t$, and with*

$$n \geq n_0 = O\left(t \cdot \sqrt{k} \cdot \log(|T|/\beta) \cdot \log^{1/2}(k \log(|T|)/\beta)\right)$$

there exists an interactive algorithm M that takes as input a dataset $s \in \mathcal{X}^n$ and provides answers $v_1, \ldots, v_k \in T$ to adaptively-chosen queries $\phi_1, \ldots, \phi_k : \mathcal{X}^t \to T$ such that, for all interactive algorithms A and distributions $\mathcal{P}$ on $\mathcal{X}$,

$$\mathop{\mathbf{Pr}}_{\substack{S\sim\mathcal{P}^n \\ (\phi_{[k]},v_{[k]})\sim A \rightleftarrows M(S)}} \left[\forall j \in [k] \quad v_j \in \mathrm{qi}_{\phi_j(\mathcal{P}^t)}\left(\rho, 1-\rho\right)\right] \geq 1-\beta,$$

where $\rho = \beta \cdot t/(4kn) = \tilde{\Omega}(\beta/(k^{3/2} \cdot \log|T|))$.

Proof We use the algorithm given in Figure 2 but with 1-approximate median, instead of 1/4. As in the proof of Theorem 1.3 we let $\mathcal{Z} \doteq \mathcal{X}^t$ and $\mathcal{D} \doteq \mathcal{P}^t$. Theorem 3.1 says that if $m \geq 4\ln(2k|T|/\beta)/\tilde{\varepsilon}$, then each execution of the median algorithm is $(\tilde{\varepsilon}, 0)$-differentially private for every input query $\phi_j : \mathcal{Z} \to T_j$. This implies that for all interactive algorithms A and every $s \in \mathcal{Z}^m$,

$$\underset{(\phi_{[k]}, v_{[k]}) \sim A \rightleftarrows M(s)}{\mathbf{Pr}} \left[\forall j \in [k] \quad v_j \in \mathrm{qi}_{\phi_j(s)}\left(0,1\right) \right] \geq 1 - \frac{\beta}{2}. \tag{7}$$

The interactive composition (Theorem 2.5) implies that M is (ε, δ)-differentially private for any $\delta \in (0,1)$ and

$$\varepsilon = \frac{k}{2} \left(\frac{4 \ln(2|T|/\beta)}{m} \right)^2 + \frac{4 \ln(2|T|/\beta)}{m} \sqrt{2k \ln(1/\delta)}.$$

Now applying Theorem 4.5 and a union bound we obtain that

$$\underset{\substack{S \sim \mathcal{D}^m \\ (\phi_{[k]}, v_{[k]}) \sim A \rightleftarrows M(S)}}{\mathbf{Pr}} \left[\exists j \in [k],\ i \in [m] \quad \phi_j(S_i) \notin \mathrm{qi}_{\phi_j(\mathcal{D})}\left(\rho, 1-\rho\right) \right] \leq km(e^{\varepsilon} 2\rho + \delta).$$

Combining this with (7) we obtain that

$$\underset{\substack{S \sim \mathcal{D}^m \\ (\phi_{[k]}, v_{[k]}) \sim A \rightleftarrows M(S)}}{\mathbf{Pr}} \left[\forall j \in [k] \quad v_j \in \mathrm{qi}_{\phi_j(\mathcal{D})}\left(\rho, 1-\rho\right) \right] \geq 1 - \frac{\beta}{2} - km(e^{\varepsilon} 2\rho + \delta).$$

Setting $m = 8 \log(2|T|/\beta)\sqrt{2k \ln(1/\delta)}$ ensures that $\varepsilon \leq \ln 2$. Hence for $\delta = \beta/(10km)$ and $\rho = \beta/(10km)$ we obtain that $km(e^{\varepsilon} 2\rho + \delta) \leq \beta/2$ thus establishing the claim. ■

For example, if each ϕ_j is $(1/\sqrt{t})$-subgaussian with the mean $\mathcal{P}^t[\phi_j] \in [-1,1]$ then for every $\alpha > 0$, setting $t = \tilde{O}(\log(k/\beta)/\alpha^2)$ ensures that $\mathrm{qi}_{\phi_j(\mathcal{P}^t)}\left(\rho, 1-\rho\right) \subseteq [\mathcal{P}^t[\phi_j] - \alpha, \mathcal{P}^t[\phi_j] + \alpha]$. This implies that the means can be estimated with accuracy α given $\tilde{O}\left(\sqrt{k} \cdot \log^2(1/\beta)/\alpha^2\right)$ samples. Note that low-sensitivity queries are $(1/\sqrt{t})$-subgaussian and therefore the sample complexity of our algorithm given by this simple analysis is comparable to the best known for this problem.

As pointed out above, this analysis can also be used to deal with general estimators by adding an additional amplification step. Namely, computing the estimator on several independent subsamples and taking (the exact) median. The resulting algorithm would have sample complexity that is identical to that obtained in Theorem 4.3 up to an additional logarithmic factor (which can be removed with careful calibration of parameters).

5. Dealing with a Large Number of Queries

In this section we briefly cover ways to use our approach when the number of queries that needs to be answered is (relatively) large. Namely, we provide an algorithm for answering verification queries and an algorithm whose complexity scales as $\log k$, rather than $\sqrt{k}$.

5.1. Verification Queries

Another application of techniques from differential privacy given by Dwork et al. (2014) is an algorithm that given a statistical query and a proposed estimate of the expectation of this query, verifies the estimate. This problem requires less data if most proposed answers are correct. Specifically, the number of samples needed by this algorithm is (just) logarithmic in the number of queries k but also scales linearly in $\sqrt{\ell}$, where ℓ is the number of queries that fail the verification step. Dwork et al. (2015a) extended this result to low-sensitivity queries using the results from (Bassily et al., 2016). In addition, Dwork et al. (2015a) describe a query verification algorithm that can handle arbitrary queries (not just real-valued) which however has sample complexity with linear dependence on ℓ. Its analysis is based on a simple description length-based argument.

A natural way to apply such algorithms is the reusable holdout technique of Dwork et al. (2015b). In this technique the dataset is split into two disjoint parts: the "training" set s_t and the holdout set s_h. The analyst then uses the training set to answer queries and perform other arbitrary analyses. The holdout set is used solely to check whether the answers that were obtained on the training set generalize. Another application proposed by Dwork et al. (2014) is an algorithm referred to as `EffectiveRounds`. This algorithm splits the dataset into several disjoint subsets and at each time uses only one of the subsets to answer queries. An algorithm for verifying answers to queries is used to switch to a new subset of samples whenever a query fails the verification step (and uses its own subset of samples).

Here we demonstrate, an algorithm for verifying answers to queries about general estimators. Formally, our algorithm satisfies the following guarantees.

Theorem 5.1 *Fix $\rho > \alpha > 0$, $\beta > 0$, $\ell, t, k \in \mathbb{N}$, and $n \geq n_0 = O(t\sqrt{\ell \log(1/\alpha\beta)}\log(k/\beta)\rho/\alpha^2)$. There exists an interactive algorithm M that takes as input $s \in \mathcal{X}^n$ and provides answers $a_1, \ldots, a_k \in \{\mathrm{Y}, \mathrm{N}, \perp\}$ to adaptively-chosen queries $(\phi_1, v_1), \ldots, (\phi_k, v_k)$ (where $\phi_j : \mathcal{X}^t \to \mathbb{R}$ and $v_j \in \mathbb{R}$ for all $j \in [k]$) satisfying the following: for all interactive algorithms A and distributions $\mathcal{P}$ over $\mathcal{X}$,*

$$\mathop{\mathbf{Pr}}_{\substack{S \sim \mathcal{P}^n \\ (\phi_{[k]}, v_{[k]}, a_{[k]}) \sim A \rightleftarrows M(S)}} \left[\forall j \in [k] \quad \begin{array}{rcl} v_j \in \mathrm{qi}_{\phi_j(\mathcal{P}^t)}(\rho, 1-\rho) & \implies & a_j \in \{\mathrm{Y}, \perp\} \\ v_j \notin \mathrm{qi}_{\phi_j(\mathcal{P}^t)}(\rho-\alpha, 1-\rho+\alpha) & \implies & a_j \in \{\mathrm{N}, \perp\} \\ |\{j' \in [j-1] : a_{j'} = \mathrm{N}\}| = \ell & \iff & a_j = \perp \end{array} \right] \geq 1-\beta.$$

First, we explain the accuracy promise of this algorithm. Each query is specified by a function $\phi_j : \mathcal{X}^t \to \mathbb{R}$ as well as a "guess" $v_j \in \mathbb{R}$. The guess is "good" if $v_j \in \mathrm{qi}_{\phi_j(\mathcal{P}^t)}(\rho, 1-\rho)$ and "bad" if $v_j \notin \mathrm{qi}_{\phi_j(\mathcal{P}^t)}(\rho-\alpha, 1-\rho+\alpha)$. Essentially, the guarantee is that, if the guess is good, the algorithm answers Y and, if the guess is bad, the algorithm answers N. However, there are two caveats to this guarantee — (i) if the guess is neither good nor bad, then the algorithm may output either Y or N and, (ii) once the algorithm has given the answer N to ℓ queries, its failure budget is "exhausted" and it only outputs $\perp$.

Note that this algorithm handles only the verification and does not provide correct answers to queries that failed the verification step. To obtain correct responses one can run an instance of the query-answering algorithm from Theorem 1.3 in parallel with the verification algorithm. The query-answering algorithm would be used only ℓ times and hence the dataset size it would require would be independent of k and scale linearly in $\sqrt{\ell}$.

(The two algorithms can either be run on disjoint subsets of data or on the same dataset since differential privacy composes.)

Our proof is a simple reduction of the verification step to verification of answers to statistical queries (relative to $\mathcal{P}^t$). Hence we could directly apply results from (Dwork et al., 2015a) to analyze our algorithm. However, as we mentioned above, for small values of ρ the existing algorithm has suboptimal dependence of sample complexity on ρ and α. Specifically, the dependence is $1/\alpha^2$ instead of ρ/α^2 which, in particular, is quadratically worse for the typical setting of $\alpha = \Theta(\rho)$. Note that the dataset size grows with ℓ – the number of times that "No" is returned to a verification query. Therefore choosing a small ρ is useful for ensuring that "No" is returned only when overfitting is substantial enough to require correction.

To improve this dependence we use our sharper generalization result (Theorem 4.2). But first we need to state the stability properties of the sparse vector technique that is the basis of this algorithm (Dwork et al., 2009) (see Dwork and Roth (2014, §3.6) for a detailed treatment).

Theorem 5.2 ((Dwork et al., 2009)) *For all $\alpha, \beta, \varepsilon, \delta > 0$, $m, k, \ell \in \mathbb{N}$ with*

$$m \geq m_0 = O(\sqrt{\ell \log(1/\delta)} \log(k/\beta)/\varepsilon\alpha),$$

there exists an interactive (ε, δ)-differentially private algorithm $\tilde{M}$ that takes as input $s \in \mathcal{Z}^m$ and provides answers $b_1, \ldots, b_k \in \{\mathrm{Y}, \mathrm{N}, \perp\}$ to adaptively chosen queries $(\psi_1, u_1), \ldots, (\psi_k, u_k)$ (where $\psi_j : \mathcal{Z} \to \mathbb{R}$ and $u_j \in \mathbb{R}$ for all $j \in [k]$) with the following accuracy guarantee. For all interactive algorithms A and all $s \in \mathcal{Z}^m$,

$$\underset{(\psi_{[k]}, u_{[k]}, b_{[k]}) \sim A \underset{\leftarrow}{\rightarrow} \tilde{M}(s)}{\mathbf{Pr}} \left[\forall j \in [k] \quad \begin{array}{rcl} s[\psi_j] > u_j & \Longrightarrow & b_j \in \{\mathrm{Y}, \perp\} \\ s[\psi_j] \leq u_j - \alpha & \Longrightarrow & b_j \in \{\mathrm{N}, \perp\} \\ |\{j' \in [j-1] : b_{j'} = \mathrm{N}\}| = \ell & \Longleftrightarrow & b_j = \perp \end{array} \right] \geq 1 - \beta.$$

Proof [of Theorem 5.1.] As before, we let $\mathcal{Z} \doteq \mathcal{X}^t$, $\mathcal{D} \doteq \mathcal{P}^t$ and view the input dataset as an element of $\mathcal{Z}^m$ sampled from $\mathcal{D}^m$.

We use the sparse vector algorithm of Theorem 5.2. We use Theorem 4.1 to convert the empirical guarantee into a guarantee relative to the data distribution, as in Theorem 4.2. However, we must convert every verification query into two statistical queries to check the two "ends" of the quantile interval. That is, given a verification query (ϕ_j, v_j) for which we want to know whether $v_j \in \mathrm{qi}_{\phi_j(\mathcal{D})}(\rho, 1-\rho)$ or $v_j \notin \mathrm{qi}_{\phi_j(\mathcal{D})}(\rho - \alpha, 1 - \rho + \alpha)$, we ask two statistical queries ψ_{2j-1} and ψ_{2j} for which we want to know if $\mathcal{D}[\psi] > \rho$ or $\mathcal{D}[\psi] < \rho - \alpha$ (with $\psi \in \{\psi_{2j-1}, \psi_{2j}\}$). Formally, we have the following reduction.

We convert the sparse vector algorithm $\tilde{M}$ of Theorem 5.2 into an algorithm M of the desired form of Theorem 5.1: Each query (ϕ_j, v_j) to M is converted into two queries (ψ_{2j-1}, u_{2j-1}) and (ψ_{2j}, u_{2j}) to $\tilde{M}$, where

$$\psi_{2j-1}(z) \doteq \mathbb{1}(\phi_j(x) \leq v_j), \qquad \psi_{2j}(z) \doteq \mathbb{1}(\phi_j(x) \geq v_j), \qquad u_{2j-1} \doteq u_{2j} \doteq \rho - \alpha/3.$$

These queries have the following key property: (†) If $v_j \in \mathrm{qi}_{\phi_j(\mathcal{D})}(\rho, 1-\rho)$, then $\mathcal{D}[\psi_{2j-1}] > \rho$ and $\mathcal{D}[\psi_{2j}] > \rho$. If $v_j \notin \mathrm{qi}_{\phi_j(\mathcal{P}^t)}(\rho - \alpha, 1 - \rho + \alpha)$, then either $\mathcal{D}[\psi_{2j-1}] \leq \rho - \alpha$ or $\mathcal{D}[\psi_{2j}] \leq \rho - \alpha$ (but not both).

Let b_{2j-1} and b_{2j} be the answers produced by $\tilde{M}$ to (ψ_{2j-1}, u_{2j-1}) and (ψ_{2j}, u_{2j}) respectively. If $b_{2j-1} = b_{2j} = \mathrm{Y}$, then M returns $a_j = \mathrm{Y}$. If $b_{2j-1} = \perp$ or $b_{2j} = \perp$ (or both), then M returns $a_j = \perp$. Otherwise M returns $a_j = \mathrm{N}$.

Note that $\tilde{M}$ must answer twice as many queries as M; thus $\tilde{M}$ must be instantiated with the value k being twice as large as for M. We also instantiate $\tilde{M}$ with α reduced by a factor of 3 and β reduced by a factor of 2. The values ε and δ used by $\tilde{M}$ will be determined later in this proof.

In particular, $\tilde{M}$ is instantiated to achieve the following accuracy guarantee. For all interactive algorithms A and all $s \in \mathcal{Z}^m$,

$$\underset{(\psi_{[k]}, u_{[k]}, b_{[k]}) \sim A \rightleftarrows \tilde{M}(s)}{\mathbf{Pr}} \left[\forall j \in [k] \quad \begin{array}{rcl} s[\psi_j] > u_j & \Longrightarrow & b_j \in \{\mathrm{Y}, \perp\} \\ s[\psi_j] \le u_j - \alpha/3 & \Longrightarrow & b_j \in \{\mathrm{N}, \perp\} \\ |\{j' \in [j-1] : b_{j'} = \mathrm{N}\}| = \ell & \Longleftrightarrow & b_j = \perp \end{array} \right] \ge 1 - \frac{\beta}{2}.$$

Now we prove that M satisfies the promised accuracy requirement relative to the distribution $\mathcal{D}$ rather than relative to the empirical values. Given the key property (†) above, it suffices to show that, with probability at least $1 - \beta/2$ over a random choice of $S \sim \mathcal{D}^m$, for all $j \in [2k]$ and $(\psi_{[k]}, u_{[k]}, b_{[k]}) \sim A \rightleftarrows \tilde{M}(S)$, we have

$$\mathcal{D}[\psi_j] > \rho \implies S[\psi_j] > u_j = \rho - \alpha/3 \quad \text{and} \quad \mathcal{D}[\psi_j] \le \rho - \alpha \implies S[\psi_j] \le u_j - \alpha/3 = \rho - 2\alpha/3. \tag{8}$$

Furthermore, to prove (8), it suffices to have

$$\mathcal{D}[\psi_j] \cdot \frac{\rho - \alpha/4}{\rho} - \frac{\alpha}{12} \le S[\psi_j] \le \mathcal{D}[\psi_j] \cdot \frac{\rho - 3\alpha/4}{\rho - \alpha} + \frac{\alpha}{12},$$

which is, in turn, implied by

$$|S[\psi_j] - \mathcal{D}[\psi_j]| \le \frac{\alpha}{4\rho} \cdot \mathcal{D}[\psi_j] + \frac{\alpha}{12}. \tag{9}$$

We will now use Theorem 4.1 to prove that (9) holds simultaneously for all $j \in [2k]$ with probability at least $1 - \beta/2$, as required to complete the proof.

First we define $f : \mathcal{F}_{\{0,1\}}^{2k} \times \mathbb{R}^{2k} \times \{\mathrm{Y}, \mathrm{N}, \perp\}^{2k} \to \mathcal{F}_{\{0,1\}}^{4k}$ by $f(\psi, u, b) = (\psi_1, 1 - \psi_1, \psi_2, 1 - \psi_2, \ldots, \psi_{2k}, 1 - \psi_{2k})$. By postprocessing (Theorem 2.3), $f(A \rightleftarrows M(s))$ is an (ε, δ)-differentially private function of $s \in \mathcal{Z}^m$ for all interactive algorithms A. The output of $f(A \rightleftarrows M(s))$ is $4k$ functions mapping $\mathcal{Z}$ to $\{0,1\}$. Set $\varepsilon = \frac{1}{2} \ln\left(1 + \frac{\alpha}{8\rho}\right)$ and $\delta = \alpha\beta/(16 \cdot 2 \cdot 12)$. Suppose $m \ge \frac{8 \cdot 12}{\varepsilon\alpha} \ln(16k/\beta)$. By Theorem 4.1,

$$\underset{\substack{S \sim \mathcal{D}^m \\ \hat{\psi} \sim f(A \rightleftarrows M(S))}}{\mathbf{Pr}} \left[\forall j \in [2k] \quad \begin{array}{rl} S[\psi_j] - \mathcal{D}[\psi_j] & \le \frac{\alpha}{12} + \frac{\alpha}{8\rho} \cdot \mathrm{mad}(\psi_j(\mathcal{D})) \\ S[1 - \psi_j] - \mathcal{D}[1 - \psi_j] & \le \frac{\alpha}{12} + \frac{\alpha}{8\rho} \cdot \mathrm{mad}(\psi_j(\mathcal{D})) \end{array} \right] \ge 1 - \frac{\beta}{2}. \tag{10}$$

Note that $\max\{S[\psi_j] - \mathcal{D}[\psi_j], S[1-\psi_j] - \mathcal{D}[1-\psi_j]\} = |S[\psi_j] - \mathcal{D}[\psi_j]|$ and mad of ψ_j is equal to mad of $1 - \psi_j$. Since $\mathrm{mad}(\psi_j(\mathcal{D})) \le 2 \cdot \mathcal{D}[\psi_j]$, the generalization bound (10) implies the desired bound (9) holds simultaneously for all $j \in [2k]$ with probability at least $1 - \beta/2$, as required.

It only remains to work out the parameters. We have $\varepsilon \geq \alpha/18\rho$. Theorem 4.1 requires $m \geq m_1$ where $m_1 = \frac{8 \cdot 12}{\varepsilon\alpha} \ln(16k/\beta) \leq 8 \cdot 12 \cdot 18 \frac{\rho}{\alpha^2} \ln(16k/\beta)$, while Theorem 5.2 requires

$$m \geq m_0 = O(\sqrt{\ell \log(1/\delta)} \log(k/\beta)/\varepsilon\alpha) = O(\sqrt{\ell \log(1/\alpha\beta)} \log(k/\beta)\rho/\alpha^2).$$

Thus the final sample complexity is $\max\{m_0, m_1\}$, as required. ■

5.2. Private Multiplicative Weights

Now we present a result in which the dependence of the required dataset size on the number of queries k is logarithmic (at the expense of some additional terms and computational efficiency).[6]

Our result follows from a direct combination of an algorithm for answering statistical queries from (Dwork et al., 2014) and the reduction from the approximate median problem to the problem of answering statistical queries relative to distribution $\mathcal{D} = \mathcal{P}^t$ (given in Lemma 3.2).

Specifically, we rely on the following result from (Dwork et al., 2014; Bassily et al., 2016) that is based on the private multiplicative weights algorithm of Hardt and Rothblum (2010) (see also (Dwork and Roth, 2014, §4.2) for further exposition).

Theorem 5.3 ((Bassily et al., 2016, Corollary 6.3)) *For all $\alpha, \beta \in (0,1)$ and $m, k \in \mathbb{N}$ with*

$$m \geq m_0 = O(\sqrt{\log|\mathcal{Z}|} \cdot \log k \cdot \log^{3/2}(1/(\alpha\beta))/\alpha^3),$$

there exists an interactive algorithm M that takes as input a dataset $s \in \mathcal{Z}^m$ and provides answers $v_1, \ldots, v_k \in [-1,1]$ to adaptively-chosen queries $\psi_1, \ldots, \psi_k : \mathcal{Z} \to [-1,1]$ such that, for all interactive algorithms A and distributions $\mathcal{D}$ over $\mathcal{Z}$,

$$\underset{\substack{S \sim \mathcal{D}^m \\ (\psi_{[k]}, v_{[k]}) \sim A \rightleftarrows M(S)}}{\mathbf{Pr}} [\forall j \in [k] \quad |v_j - \mathcal{D}[\psi_j(\mathcal{D})] \leq \alpha] \geq 1 - \beta.$$

Now, by Lemma 3.2, for $\mathcal{Z} = \mathcal{X}^t$ and $\mathcal{D} = \mathcal{P}^t$ and any query $\phi : \mathcal{Z} \to T$, responses to $2\lceil \log_2 |T| \rceil$ statistical queries relative to $\mathcal{D}$ with accuracy $1/8$ can be used to find a value $v \in \mathrm{qi}_{\phi(\mathcal{D})}(1/4, 3/4)$. By plugging this reduction into Theorem 5.3 we get the following result.

Theorem 5.4 *For any $\beta \in (0,1)$, $t, k \in \mathbb{N}$, a finite set $T \subset \mathbb{R}$ and $\mathcal{Z} = \mathcal{X}^t$ and with*

$$n \geq n_0 = O\left(t^{3/2} \cdot \sqrt{\log|\mathcal{X}|} \cdot \log(k \log|T|) \cdot \log^{3/2}(1/\beta)\right)$$

there exists an interactive algorithm M that takes as input a dataset $s \in \mathcal{P}^n$ and provides answers $v_1, \ldots, v_k \in T$ to adaptively-chosen queries $\phi_1, \ldots, \phi_k : \mathcal{X}^t \to T$ such that, for all interactive algorithms A and distributions $\mathcal{P}$ over $\mathcal{X}$,

$$\underset{\substack{S \sim \mathcal{P}^n \\ (\phi_{[k]}, v_{[k]}) \sim A \rightleftarrows M(S)}}{\mathbf{Pr}} \left[\forall j \in [k] \quad v_j \in \mathrm{qi}_{\phi_j(\mathcal{P}^t)}\left(\frac{1}{4}, \frac{3}{4}\right)\right] \geq 1 - \beta.$$

[6] It is known that in this general setting, the dependence on the data universe size and the loss of computational efficiency are unavoidable (Hardt and Ullman, 2014; Steinke and Ullman, 2015).

For example we can use this algorithm to obtain a new algorithm for answering a large number of low-sensitivity queries (that is queries $\phi : \mathcal{X}^t \to [-1, 1]$ such that $\Delta(\phi) = 1/t$). To answer queries with accuracy α we can use $t = 16/\alpha^2$ and set T that is the interval $[-1, 1]$ discretized with step $\alpha/2$. Thus the number of samples that our algorithm uses is $n = O\left(\sqrt{\log|\mathcal{X}|} \cdot \log(k/\alpha) \cdot \log^{3/2}(1/\beta)/\alpha^3\right)$. For comparison, the best previously known algorithm for this problem uses

$$n = O\left(\log|\mathcal{X}| \cdot \log(k/\alpha) \cdot \log^{3/2}(1/\beta)/\alpha^4\right)$$

(Bassily et al., 2016) (a different bound is stated there in error). Although, as pointed out in the introduction, the setting in which each query is applied to the entire dataset is more general than ours.

References

Raef Bassily and Yoav Freund. Typicality-based stability and privacy. *CoRR*, abs/1604.03336, 2016. URL http://arxiv.org/abs/1604.03336.

Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *STOC*, pages 1046–1059, 2016.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer Berlin Heidelberg, 2016.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 634–649. IEEE, 2015.

Mark Bun, Thomas Steinke, and Jonathan Ullman. Make up your mind: The price of online queries in differential privacy. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1306–1325. Society for Industrial and Applied Mathematics, 2017.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006a.

C. Dwork, M. Naor, O. Reingold, G. Rothblum, and S. Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *STOC*, pages 381–390, 2009.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *STOC*, pages 371–380, 2009.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*, volume 9. 2014. URL http://dx.doi.org/10.1561/0400000042.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer Berlin Heidelberg, 2006b.

Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *FOCS*, pages 51–60, 2010.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. *CoRR*, abs/1411.2664, 2014. Extended abstract in STOC 2015.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, pages 2350–2358, 2015a.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349 (6248):636–638, 2015b. doi: 10.1126/science.aaa9375. URL http://www.sciencemag.org/content/349/6248/636.abstract.

Vitaly Feldman. Dealing with range anxiety in mean estimation via statistical queries. *arXiv*, abs/1611.06475, 2016. URL http://arxiv.org/abs/1611.06475.

M. Hardt and G. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *FOCS*, pages 61–70, 2010.

M. Hardt and J. Ullman. Preventing false discovery in interactive data analysis is hard. In *FOCS*, pages 454–463, 2014.

M. Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.

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

Kobbi Nissim and Or Sheffet. Topics in cryptography and privacy - differential privacy. hw 1. http://isites.harvard.edu/fs/docs/icb.topic1475289.files/hwk1-kobbi.pdf, 2014.

Kobbi Nissim and Uri Stemmer. Concentration bounds for high sensitivity functions through differential privacy. *CoRR*, abs/1703.01970, 2017. URL http://arxiv.org/abs/1703.01970.

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

Sofya Raskhodnikova and Adam D. Smith. Algorithmic challenges in data privacy. hw 1. http://www.cse.psu.edu/ads22/privacy598/handouts/hw1.pdf, 2010.

Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pages 487–494. IEEE, 2016.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS*, 2016.

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 813–822. ACM, 2011.

Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *COLT*, pages 1588–1628, 2015. URL http://jmlr.org/proceedings/papers/v40/Steinke15.html.

Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 2016.

Thomas Steinke and Jonathan Ullman. Subgaussian tail bounds via stability arguments. *CoRR*, abs/1701.03493, 2017. URL http://arxiv.org/abs/1701.03493.

Appendix A. Proof of Theorem 4.1

Recall that, for a distribution $\mathcal{D}$ on $\mathbb{R}$, we define its mean absolute deviation by

$$\mathrm{mad}(\mathcal{D}) \doteq \underset{X\sim\mathcal{D}}{\mathbf{E}}[|X - \underset{Y\sim\mathcal{D}}{\mathbf{E}}[Y]|].$$

We first give a bound relating differential privacy to expectations:

Lemma A.1 *Fix $\mu, \varepsilon, \delta, \Delta \in \mathbb{R}$. Let X and Y be random variables supported on $[\mu - \Delta, \mu + \Delta]$. Suppose that X and Y are (ε, δ)-indistinguishable, that is*

$$e^{-\varepsilon}\left(\mathbf{Pr}\left[X \in E\right] - \delta\right) \le \mathbf{Pr}\left[Y \in E\right] \le e^{\varepsilon}\,\mathbf{Pr}\left[X \in E\right] + \delta$$

for all $E \subseteq \mathbb{R}$. Then

$$|\mathbf{E}\left[X\right] - \mathbf{E}\left[Y\right]| \le (e^{\varepsilon} - 1)\mathbf{E}\left[|X - \mu|\right] + 2\delta\Delta.$$

Thus, if $M : \mathcal{Z}^m \to [0,1]$ satisfies (ε, δ)-differential privacy, then for any datasets $s, s' \in \mathcal{Z}^m$ differing on a single entry, we have

$$|\mathbf{E}\left[M(s)\right] - \mathbf{E}\left[M(s')\right]| \le (e^{\varepsilon} - 1) \inf_{\mu\in[0,1]} \mathbf{E}\left[|M(s) - \mu|\right] + 2\delta \le (e^{\varepsilon} - 1)\cdot \mathrm{mad}(M(s)) + 2\delta.$$

In contrast, Bassily et al. (2016) use a bound corresponding to $|\mathbf{E}\left[M(S)\right] - \mathbf{E}\left[M(S')\right]| \le e^{\varepsilon} - 1 + \delta$.

Proof We use three facts: (i) $x = \max\{x,0\} - \max\{-x,0\}$ and $|x| = \max\{x,0\} + \max\{-x,0\}$ for all $x \in \mathbb{R}$, (ii) if $X \geq 0$, then $\mathbf{E}[X] = \int_0^\infty \mathbf{Pr}[X \geq t]\,\mathrm{d}t$, and (iii) $e^\varepsilon - 1 \geq 1 - e^{-\varepsilon}$.

$$\begin{aligned}
\mathbf{E}[Y-\mu] =& \mathbf{E}[\max\{Y-\mu,0\} - \max\{\mu-Y,0\}] \\
=& \int_0^\infty \mathbf{Pr}[Y-\mu \geq t] - \mathbf{Pr}[\mu - Y \geq t]\,\mathrm{d}t \\
=& \int_0^\Delta \mathbf{Pr}[Y-\mu \geq t] - \mathbf{Pr}[\mu - Y \geq t]\,\mathrm{d}t \\
\leq& \int_0^\Delta (e^\varepsilon \mathbf{Pr}[X-\mu \geq t] + \delta) - e^{-\varepsilon}(\mathbf{Pr}[\mu - X \geq t] - \delta)\mathrm{d}t \\
=& \int_0^\Delta \mathbf{Pr}[X-\mu \geq t] - \mathbf{Pr}[\mu - X \geq t]\,\mathrm{d}t \\
&+ \int_0^\Delta (e^\varepsilon - 1)\mathbf{Pr}[X-\mu \geq t] + \delta + (1-e^{-\varepsilon})\mathbf{Pr}[\mu - X \geq t] + e^{-\varepsilon}\delta\mathrm{d}t \\
=& \mathbf{E}[X-\mu] + (e^\varepsilon-1)\mathbf{E}[\max\{X-\mu,0\}] + (1-e^{-\varepsilon})\mathbf{E}[\max\{\mu-X,0\}] + (1+e^{-\varepsilon})\delta\Delta \\
\leq& \mathbf{E}[X-\mu] + (e^\varepsilon-1)\mathbf{E}[\max\{X-\mu,0\} + \max\{\mu-X,0\}] + 2\delta\Delta \\
=& \mathbf{E}[X-\mu] + (e^\varepsilon-1)\mathbf{E}[|X-\mu|] + 2\delta\Delta.
\end{aligned}$$

Thus $\mathbf{E}[Y] - \mathbf{E}[X] \leq (e^\varepsilon - 1)\mathbf{E}[|X-\mu|] + 2\delta\Delta$. To obtain the other half of the result, replace X, Y, and μ with their negations in the above. ■

Intuitively, the following lemma says the following. Suppose a differentially private algorithm is given ℓ independent samples $S^1, \dots, S^\ell \sim \mathcal{D}^m$. The algorithm picks one of the ℓ samples and produces a statistical query. The algorithm's "goal" is to "overfit" — that is, to produce a query whose empirical value on the chosen sample differs from the expected value on the population. The lemma says that this cannot happen in expectation. The reason for the ℓ-fold repetition is probability amplification: The lemma says the mechanism cannot overfit in expectation, given ℓ chances to do so. Consequently, if the mechanism is given only one chance to overfit (i.e. one sample $S \sim \mathcal{D}^m$), then with high probability it cannot. The ℓ repetitions mean that, if the mechanism can overfit with probability $1/\ell$ per sample, then it can overfit with constant probability given ℓ samples.

Lemma A.2 *Let $M : (\mathcal{Z}^m)^\ell \to [\ell] \times \mathcal{F}_{[-1,1]}$ be a (ε, δ)-differentially private algorithm with $\mathcal{F}_{[-1,1]}$ being the set of functions $\phi : \mathcal{Z} \to [-1,1]$. Let $\mathcal{D}$ be a distribution on $\mathcal{Z}$. Then*

$$\underset{\substack{S^1,\dots,S^\ell \sim \mathcal{D}^m \\ (k,\phi)\sim M(S)}}{\mathbf{E}} \left[S^k[\phi] - \mathcal{D}[\phi]\right] \leq (e^\varepsilon - 1) \underset{\substack{(S,Z)\sim(\mathcal{D}^m)^\ell \times \mathcal{D} \\ (k,\phi)\sim M(S)}}{\mathbf{E}} [|\phi(Z)|] + 2\delta m.$$

Proof For $j \in [\ell]$, define $f_j : [\ell] \times \mathcal{F}_{[-1,1]} \times \mathcal{Z} \to [-1,1]$ by $f_j(k,\phi,x) = \phi(x) \cdot \mathbb{1}(k=j)$.

We use two facts:

(i) For $(S,Z) \sim (\mathcal{D}^m)^\ell \times \mathcal{D}$, the pair (S, S_i^j) has the same distribution as the pair $((S_{-i}^{-j}, Z), Z)$, where (S_{-i}^{-j}, Z) denotes S with the $(i,j)^{\text{th}}$ entry S_i^j replaced by Z.

(ii) By differential privacy of M, for all fixed $(s, z) \in (\mathcal{Z}^m)^\ell \times \mathcal{Z}$, the distribution of $(M(s_{-i}^{-j}, z), z)$ is (ε, δ)-indistinguishable from $(M(s), z)$. Hence, for a random pair $(S, Z) \sim (\mathcal{D}^m)^\ell \times \mathcal{D}$, $(M((S_{-i}^{-j}, Z), Z)$ is also (ε, δ)-indistinguishable from $(M(S), Z)$.

Now

$$
\begin{aligned}
\mathop{\mathbf{E}}_{\substack{S\sim(\mathcal{D}^m)^\ell \\ (k,\phi)\sim M(S)}} \left[S^k[\phi] - \mathcal{D}[\phi]\right] &= \sum_{j\in[\ell]} \frac{1}{m} \sum_{i\in[m]} \mathop{\mathbf{E}}_{\substack{S\sim(\mathcal{D}^m)^\ell \\ (k,\phi)\sim M(S)}} \left[(\phi(S_i^j) - \mathcal{D}[\phi]) \cdot \mathbb{1}(k=j)\right] \\
&= \sum_{j\in[\ell]} \frac{1}{m} \sum_{i\in[m]} \mathop{\mathbf{E}}_{\substack{S\sim(\mathcal{D}^m)^\ell \\ Z\sim\mathcal{D}}} \left[f_j(M(S), S_i^j) - f_j(M(S), Z)\right] \\
\text{(Fact (i))} \qquad &= \sum_{j\in[\ell]} \frac{1}{m} \sum_{i\in[m]} \mathop{\mathbf{E}}_{\substack{S\sim(\mathcal{D}^m)^\ell \\ Z\sim\mathcal{D}}} \left[f_j(M(S_{-i}^{-j}, Z), Z) - f_j(M(S), Z)\right] \\
\text{(Fact (ii) and Lemma A.1)} \qquad &\leq \sum_{j\in[\ell]} \frac{1}{m} \sum_{i\in[m]} (e^\varepsilon - 1) \mathop{\mathbf{E}}_{\substack{S\sim(\mathcal{D}^m)^\ell \\ Z\sim\mathcal{D}}} \left[|f_j(M(S), Z)|\right] + 2\delta \\
&= (e^\varepsilon - 1) \mathop{\mathbf{E}}_{\substack{(S,Z)\sim(\mathcal{D}^m)^\ell\times\mathcal{D} \\ (k,\phi)\sim M(S)}} \left[|\phi(Z)|\right] + 2\delta\ell.
\end{aligned}
$$

■

Now we can state and prove the "transfer theorem" which relates differential privacy to generalization. This result improves the previous bound (Bassily et al., 2016) by having a bound in terms of the mean absolute deviation, rather than the sensitivity, (i.e. the previous bound can be obtained (up to constants) by replacing $\mathrm{mad}(\cdot)$ in the expression by 1). Also, the previous bound is only stated for $k = 1$.

Theorem A.3 (Theorem 4.1) *Fix $\beta, \varepsilon, \delta \in (0, 1)$ and $m, k \in \mathbb{N}$. Let $M : \mathcal{Z}^m \to \mathcal{F}_{[0,1]}^k$ be a (ε, δ)-differentially private algorithm with $\mathcal{F}_{[0,1]}$ being the set of functions $\phi : \mathcal{Z} \to [0, 1]$. Let $\mathcal{D}$ be a distribution on $\mathcal{Z}$. Then*

$$
\Pr_{\substack{S\sim\mathcal{D}^m \\ \phi\sim M(S)}} \left[\exists j \in [k] \qquad S[\phi_j] - \mathcal{D}[\phi_j] > \alpha + \gamma \cdot \mathrm{mad}\left(\phi_j(\mathcal{D})\right)\right] \leq \beta
$$

for $\alpha = \frac{4}{\varepsilon m} \ln\left(2\frac{k}{\beta}\right) + 8\frac{\delta}{\beta}$, $\gamma = e^{2\varepsilon} - 1$ and $\beta \in (0, 1)$ arbitrary.

We note that the statement of Theorem 4.1 has been rearranged to make α and γ independent parameters and ε and δ dependent parameters, rather than the reverse.

Proof Let $\ell = \lceil 1/\beta \rceil$. Define $M' : (\mathcal{Z}^m)^\ell \to [\ell] \times \mathcal{F}_{[-1,1]}$ as follows. On input S it runs ℓ copies of M on $S^1, \ldots, S^\ell$ and obtains outputs $\phi^1, \ldots, \phi^\ell \in \mathcal{F}_{[0,1]}^k$. Also define ϕ_0^0 to be the constant 0 function and S^0 to be an arbitrary fixed element of $\mathcal{Z}^m$. Now M' randomly samples $(I, J) \in ([\ell] \times [k]) \cup \{(0, 0)\}$ with

$$
\Pr\left[I = i \wedge J = j\right] \propto \exp\left(\frac{\varepsilon m}{2}\left(S^i[\phi_j^i] - \mathcal{D}[\phi_j^i] - \gamma \cdot \mathrm{mad}\left(\phi_j^i(\mathcal{D})\right)\right)\right).
$$

Finally, M' returns $(\max\{I, 1\}, \phi^*)$ where $\phi^*(x) = \phi_J^I(x) - \mathcal{D}[\phi_J^I]$.

Firstly, M' satisfies $(2\varepsilon, \delta)$-differential privacy: The choice of I and J is $(\varepsilon, 0)$-differentially private, as it is an instantiation of the exponential mechanism (McSherry and Talwar, 2007). Simple composition property of differential privacy (Dwork and Roth, 2014, Theorem 3.16) then implies this privacy bound.

Moreover, by the properties of the exponential mechanism (Bassily et al., 2016, Lemma 7.1)

$$\begin{aligned}&\mathop{\mathbf{E}}_{I,J}\left[S^I[\phi^I_J] - \mathcal{D}[\phi^I_J] - \gamma \cdot \mathrm{mad}\left(\phi^I_J(\mathcal{D})\right)\right] \\ &\qquad \geq \max_{(i,j)\in[\ell]\times[k]\cup\{(0,0)\}} S^i[\phi^i_j] - \mathcal{D}[\phi^i_j] - \gamma \cdot \mathrm{mad}\left(\phi^i_j(\mathcal{D})\right) - \frac{2}{\varepsilon m}\ln(\ell\cdot k+1). \end{aligned} \tag{11}$$

On the other hand, by Lemma A.2,

$$\mathop{\mathbf{E}}_{\substack{S\sim(\mathcal{D}^m)^\ell \\ (I,\phi^*)\sim M'(S)}}\left[S^I[\phi^*] - \mathcal{D}[\phi^*]\right] \leq (e^{2\varepsilon}-1) \mathop{\mathbf{E}}_{\substack{(S,Z)\sim(\mathcal{D}^m)^\ell\times\mathcal{D} \\ (I,\phi^*)\sim M'(S)}}[|\phi^*(Z)|] + 2\delta m. \tag{12}$$

For the sake of contradiction, assume that

$$\mathop{\mathbf{Pr}}_{\substack{S\sim\mathcal{D}^m \\ \phi\sim M(S)}}\left[\exists j \in [k] \quad S[\phi_j] - \mathcal{D}[\phi_j] > \alpha + \gamma\cdot\mathrm{mad}\left(\phi_j(\mathcal{D})\right)\right] > \beta.$$

It follows that

$$\mathop{\mathbf{Pr}}_{\substack{S^1,\ldots,S^\ell\sim\mathcal{D}^m \\ \phi^1,\ldots,\phi^\ell\sim M(S^1),\ldots,M(S^\ell)}}\left[\max_{(i,j)\in[\ell]\times[k]} S^i[\phi^i_j] - \mathcal{D}[\phi^i_j] - \gamma\cdot\mathrm{mad}\left(\phi^i_j(\mathcal{D})\right) > \alpha\right] > 1-(1-\beta)^\ell$$

and, hence,

$$\mathop{\mathbf{E}}_{\substack{S^1,\ldots,S^\ell\sim\mathcal{D}^m \\ \phi^1,\ldots,\phi^\ell\sim M(S^1),\ldots,M(S^\ell)}}\left[\max_{(i,j)\in[\ell]\times[k]\cup\{(0,0)\}} S^i[\phi^i_j] - \mathcal{D}[\phi^i_j] - \gamma\cdot\mathrm{mad}\left(\phi^i_j(\mathcal{D})\right)\right] > \alpha(1-(1-\beta)^\ell). \tag{13}$$

Combining (11), (12), (13), and $\gamma = e^{2\varepsilon}-1$ yields

$$\alpha(1-(1-\beta)^\ell) < \frac{2}{\varepsilon m}\ln(\ell\cdot k+1) + 2\delta\ell.$$

Since $\ell \geq 1/\beta$, $1-(1-\beta)^\ell \geq 1-e^{-1} \geq 1/2$. We also have $\ell \leq 1/\beta + 1 \leq 2/\beta$. Thus $\frac{\alpha}{2} < \frac{2}{\varepsilon m}\ln\left(2\frac{k}{\beta}\right) + 4\frac{\delta}{\beta}$ — a contradiction. ■