Sublinear-Time Adaptive Data Analysis

Benjamin Fish[1], Lev Reyzin[1], and Benjamin I. P. Rubinstein[2]

[1]Dept. of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago
[2]School of Computing & Information Systems, University of Melbourne

September 8, 2017

Abstract

The central aim of most fields of data analysis and experimental scientific investigation is to draw valid conclusions from a given data set. But when past inferences guide future inquiries into the same dataset, reaching valid inferences becomes significantly more difficult. In addition to avoiding the overfitting that can result from adaptive analysis, a data analyst often wants to use as little time and data as possible. A recent line of work in the theory community has established mechanisms that provide low generalization error on adaptive queries, yet there remain large gaps between established theoretical results and how data analysis is done in practice. Many practitioners, for instance, successfully employ bootstrapping and related sampling approaches in order to maintain validity and speed up analysis, but prior to this work, no theoretical analysis existed to justify employing such techniques in this adaptive setting.

In this paper, we show how these techniques can be used to provably guarantee validity while speeding up analysis. Through this investigation, we initiate the study of sub-linear time mechanisms to answer adaptive queries into datasets. Perhaps surprisingly, we describe mechanisms that provide an exponential speed-up per query over previous mechanisms, without needing to increase the total amount of data needed for low generalization error. We also provide a method for achieving statistically-meaningful responses even when the mechanism is only allowed to see a constant number of samples from the data per query.

1 Introduction

The field of data analysis seeks out statistically valid conclusions from data: inferences that generalize to an underlying distribution rather than specialize to the data sample at hand. As a result, classical proofs of statistical efficiency have focused on independence assumptions on data with a pre-determined sequence of analyses [20]. In practice, most data analysis is adaptive: previous inferences inform future analysis. This adaptivity is nigh impossible to avoid when multiple scientists contribute work to an area of study using the same or similar data sets. Unfortunately, adaptivity may lead to 'false discovery,' where the dependence on past analysis may create pervasive overfitting—also known as 'the garden of forking paths' or 'p hacking' [13]. While basing each analysis on new data drawn from the same distribution might appear an appealing solution, repeated data collection and analysis time can be prohibitively costly.

There has been much recent progress in minimizing the amount of data needed to draw generalizable conclusions, without having to make any assumptions about the type of adaptations used by the data analysis. However, the results in this burgeoning field of adaptive data analysis have ignored bootstrapping and related sampling techniques, even though it has enjoyed

Table 1: Time per query

Query Type	Previous Work	This Work
Low-sensitivity queries with $\tilde{O}\left(\frac{\sqrt{k}}{\alpha^2}\right)$ sample complexity	$\tilde{O}\left(\frac{\sqrt{k}}{\alpha^2}\right)$ [1]	$\tilde{O}\left(\frac{\log^2(k)}{\alpha^2}\right)$
Sampling counting queries with $\tilde{O}\left(\frac{\sqrt{k}}{\alpha^2}\right)$ sample complexity	$\tilde{O}\left(\frac{\sqrt{k}}{\alpha^2}\right)$	$\tilde{O}\left(\log\left(\frac{k}{\alpha}\right)\right)$

Summary of our results. k is the number of queries and α is the accuracy rate. Dependence on the probability of failure has been suppressed for ease of reading. For more precise definitions, see Section 2.

widespread and successful use in practice in a variety of settings [19, 29], including in adaptive settings [14]. This is a gap that not only points to an unexplored area of theoretical study, but also opens up the possibility of creating substantially faster algorithms for answering adaptively generated queries.

In this paper, we aim to do just this: we develop strong theoretical results that are exponentially faster than previous approaches, and we open up a host of interesting open problems at the intersection of sublinear-time algorithm design and this important new field. For example, sub-linear time algorithms are a necessary component to establish non-trivial results in property testing. We also enable the introduction of anytime algorithms in adaptive data analysis, by defining mechanisms that provide guarantees on accuracy when the time allotted is restricted.

As in previous literature, a mechanism $\mathcal{M}$ is given an i.i.d. sample S of size n from an unknown distribution D over a finite space X, and is given queries of the form $q : D \to \mathbb{R}$. After each query, the mechanism must respond with an answer a that is close to $q(D)$ up to a parameter α with high probability. Furthermore, each query may be adaptive: The query may depend on the previous queries and answers to those queries.

In previous work, the mechanisms execute in $\Omega(n)$ time per query. In this work, we introduce mechanisms that make an exponential improvement on this bound. Remarkably, we show that these results come at almost no tradeoff—we can obtain these exponential improvements in running time and yet use essentially the same sample sizes.

1.1 Motivation and results

Our results are summarized in Table 1. Our first result, in Section 4, is a method to answer low-sensitivity queries (defined in Section 2) that still has $\tilde{O}(\sqrt{k}/\alpha^2)$ sample complexity (as in previous work) but takes exponentially less time per query than previous approaches (Theorem 10). Moreover, our mechanism to answer a query is simple: given a database S, we first sample ℓ points i.i.d. from S, compute the empirical mean of q on that subsample, and then add Laplacian noise, which guarantees a *differentially-private* mechanism. The intuition behind this approach is that sampling offers two benefits: it can decrease computation time while simultaneously boosting privacy. Privacy yields a strong notion of stability, which in turn allows us to reduce the computation time without sacrificing accuracy. In particular, this mechanism takes only $\tilde{O}(\log^2(k)/\alpha^2)$ time per query and a sample size of $\ell = \tilde{O}(\log(k)/\alpha^2)$, all while matching the previous sample complexity bound $\tilde{O}(\sqrt{k}/\alpha^2)$ for the size of S. Even in the non-adaptive case, it must take $\Omega(\log(k)/\alpha^2)$ samples to answer k such queries [1]. This means our results are tight in ℓ and come close to matching the best known lower-bound for the time complexity of answering such queries adaptively, which is simply $\Omega(\log(k)/\alpha^2)$.

This resampling approach to answering low-sensitivity queries leverages sampling without replacement. However, sampling without replacement is often slower or requires more space than does sampling *with* replacement. We therefore investigate consequences of first sampling with replacement rather than sampling without replacement. We show that in the special case when the low-sensitivity query is a counting query (also defined in Section 2), we obtain the same results using sampling with replacement as without (Theorem 12). This requires a new method, which we provide, for guaranteeing the effects of sampling on differential privacy. This is because a subsample without replacement can only change by a single element under a single element perturbation, while a subsample with replacement can be arbitrarily effected by the same perturbation—this makes analysis significantly more difficult.

While both sampling methods require examining $\ell = \tilde{O}(\log(k)/\alpha^2)$ samples per query, an analyst may wish to control the number of samples used. For example, the analyst might want the answer to a counting query using a very small number of sample points from the database, even just a single sample point. The above methods cannot handle this case gracefully, because the empirical answer to such a query is $\{0,1\}$-valued, while a response using Laplacian noise will not be. Even if one enforces a semantically meaningful response, say using the exponential mechanism, there is another issue: In this model of data analysis mechanisms are permitted to be black-box, where the analyst is not guaranteed any 'internal knowledge' about the mechanism. For example, mechanisms are not required to be anytime algorithms, i.e. have any guarantees if the query is only able to be evaluated on part of the sample set. Indeed, when ℓ is sufficiently small, the guarantees on accuracy (using Definition 2 below) become trivial—we get only that $\alpha = O(1)$, which any mechanism will satisfy. Instead, we want the mechanism to have to return a statistically-meaningful reply even if $\ell = 1$.

To address these issues, we consider an 'honest' setting where the mechanism must always yield a plausible reply to each query (Section 5). This is analogous to the honest version [30] of the statistical query (SQ) setting for learning [2, 17], or the 1-STAT oracle for optimization [12]. Thus we introduce *sampling counting queries*, which imitate the process of an analyst requesting the value of a query on a single random sample. This allows for greater control over how long each query takes, in addition to greater control over the outputs. Namely, we require that for a query of the form $q : X \to \{0,1\}$, the mechanism must output a $\{0,1\}$-valued answer that is accurate in expectation. We show how to answer queries of this form by sampling a single point x from S and then applying a simple differentially-private algorithm to $q(x)$ that has not been used in adaptive data analysis prior to this work (Theorem 16). Finally, in Section 6, we compare sampling counting queries to counting queries.

1.2 Previous work

Previous work in this area has focused on finding accurate mechanisms with low sample complexity (the size of S) for a variety of queries and settings [1, 6, 7, 24, 25]. Most applicable to our work is that of Bassily et al. [1] who consider, among others, *low-sensitivity queries*, which are merely any function of X^n whose output does not change much when the input is perturbed (for a more precise definition, see below). If the queries are nonadaptive, then only roughly $\log(k)/\alpha^2$ samples are needed to answer k such queries. And if the queries are adaptive but the mechanism simply outputs the empirical estimate of q on S, then the sample complexity is order k/α^2 instead—exponentially worse.

In this paper, we will focus only on computationally efficient mechanisms. It is not necessarily obvious that it is possible to achieve a smaller sample complexity for an efficient mechanism in the adaptive case, but Bassily et al. [1], building on the work of Dwork et al. [7], provide a mechanism with sample complexity $n = \tilde{O}(\sqrt{k}/\alpha^2)$ to answer k low-sensitivity queries. Furthermore, for efficient mechanisms, this bound is tight in k [26]. This literature shows that the key to finding such mechanisms with this quadratic improvement over the naive method is finding stable mechanisms: those whose output does not change too much when the sample is changed by a single element. Much of this literature leverages differential privacy [1, 6, 7, 25], which offers a strong notion of stability.

Since this work uses differentially-private mechanisms after sampling, we are acutely interested in the impact on privacy when sampling. In both theory and practice, sampling in settings where privacy matters has long been deemed useful, in a variety of areas [15, 16, 18].

In our setting, we need an efficient uniform sampling method that not only maintains privacy, but actually boosts it. In particular, for an ϵ-private mechanism on a database of size n, we want to show that if you sample ℓ points uniformly and efficiently from those n points, and then apply the same mechanism, the result is $O\left(\frac{\ell}{n}\epsilon\right)$-private.

In previous literature, a few sampling techniques have been analyzed in the context of differential privacy. First discussed by Chaudhuri and Mishra [5], Li et al. [21] show that Bernoulli sampling boosts privacy. Unfortunately, Bernoulli sampling also takes $\Omega(n)$ time, precluding any sub-linear time algorithm. Ebadi et al. [11] analyze sampling without replacement and fraction sampling, but they only prove that these maintain privacy, rather than boosting it: sampling and then applying an ϵ-private mechanism results in only $O(\epsilon)$-privacy. In the context of their sample-and-aggregate framework, Nissim et al. [23] also show that a version of sampling without replacement maintains privacy. Most closely to what we need, for a particular setting, Lin et al. [22] show that sampling without replacement actually boosts privacy.

We note that their proof method easily generalizes to arbitrary domains and ϵ-private mechanisms. On the other hand, sampling *with* replacement has not been analyzed prior to this work, despite it being one of the simplest and common sampling techniques.

2 Model and preliminaries

In the adaptive data analysis setting we consider, a (possibly stateful) mechanism $\mathcal{M}$ that is given an i.i.d. sample S of size n from an unknown distribution D over a finite space X. The mechanism $\mathcal{M}$ must answer queries from a stateful adversary $\mathcal{A}$. These queries are adaptive: $\mathcal{A}$ outputs a query q_i, to which the mechanism returns a response a_i, and the outputs of $\mathcal{A}$ and $\mathcal{M}$ may depend on all queries $q_1, \ldots, q_{i-1}$ and responses $a_1, \ldots, a_{i-1}$.

2.1 Low-sensitivity queries

In this work, the first type of query we consider is a *low-sensitivity query*, which is specified by a function $q : X^n \to [0,1]$ with the property that for all samples $S, S' \in X^n$ where S and S' differ by at most one element, we have $|q(S) - q(S)'| \leq 1/n$, where we define $q(D) = \mathbb{E}_{S\sim D^n}[q(S)]$. We can now define the accuracy of $\mathcal{M}$.

Definition 1. *A mechanism $\mathcal{M}$ is said to be* (α, β)-accurate *over a sample S on queries $q_1, \ldots, q_k$ if for its responses $a_1, \ldots, a_k$ we have*

$$\mathbb{P}_{\mathcal{M},\mathcal{A}}[\max_i |q_i(S) - a_i| \leq \alpha] \geq 1 - \beta.$$

The key requirement is stronger. Namely, we seek accuracy over the unknown distribution.

Definition 2. *A mechanism $\mathcal{M}$ is* (α, β)-accurate *on distribution D, and on queries $q_1, \ldots, q_k$, if for its responses $a_1, \ldots, a_k$ we have*

$$\mathbb{P}_{\mathcal{M},\mathcal{A}}[\max_i |q_i(D) - a_i| \leq \alpha] \geq 1 - \beta.$$

Using this notation, we seek out mechanisms that are (α, β)-accurate on D with access not to D directly but only an i.i.d. sample S from D. However, we also want mechanisms that are fast. In this work, we consider the time per query $\mathcal{M}$ takes. In particular, we want $\mathcal{M}$ to take no more than approximately $\log(n)$ time for each query. In general, this will not be possible if merely computing $q(S)$ for a query q on a given sample S of size ℓ takes more than $O(\ell)$ time. In this work, we will not consider the extra time it takes to compute q itself, rather just the time it takes to respond using an oracle to q. Thus we assume we will have oracle access to q, which will compute $q(x)$ for a sample point x in unit time (and also $q(S)$ in at most $|S|$ time). This is not a strong assumption: If computing a query actually takes polynomial time in the number of elements in the sample given to q and the size of each element in that sample, then this can add only at most a poly-log factor overhead in n and $|X|$ (as long as we only compute q on a roughly $\log(n)$ size sample, which will turn out to be exactly the case).

2.2 Counting queries and sampling counting queries

In this work we also consider *counting queries*, which ask the question "What proportion of the data satisfies property q?" Counting queries are a simple and important restriction of low-sensitivity queries [3, 4, 25]. More formally, a counting query is specified by a function $q : X \to \{0, 1\}$, where $q(S) = \frac{1}{|S|}\sum_{x\in S} q(x)$ and $q(D) = \mathbb{E}_{x\sim D}[q(x)]$. As in the low-sensitivity setting, an answer to a counting query must be close to $q(D)$ (Definition 2).

This means, however, that answers will not necessarily be counts themselves, nor meaningful in settings where we require ℓ to be small, i.e. very few samples from the database. To this end, we introduce *sampling counting queries.* A sampling counting query (SCQ) is again specified by a function $q : X \to \{0, 1\}$, but this time the mechanism $\mathcal{M}$ must return an answer $a \in \{0, 1\}$. Given these restricted responses, we want such a mechanism to act like what would happen if $\mathcal{A}$ were to take a single random sample point x from D and evaluate $q(x)$. The average value the mechanism returns (over the coins of the mechanism) should be close to the expected value of q. More precisely, we want the following:

Definition 3. *A mechanism $\mathcal{M}$ is (α, β)-accurate on distribution D for k sampling counting queries q_i if for all states of $\mathcal{M}$ and $\mathcal{A}$, when $\mathcal{M}$ is given an i.i.d. sample S from D,*

$$\mathbb{P}_{S,\mathcal{M},\mathcal{A}}\left[\max_i |\mathbb{E}_{\mathcal{M}}[\mathcal{M}(q_i)] - q_i(D)| \le \alpha\right] \ge 1 - \beta.$$

We also define (α, β)-accuracy on a sample S from D analagously. Again, our requirement is that $\mathcal{M}$ be (α, β)-accurate with respect to the unknown distribution D, this time using only around $\log(n)$ time per query.

3 Differential privacy

Differential privacy, first introduced by Dwork et al. [8], provides a strong notion of stability.

Definition 4 (Differential privacy)**.** *Let $\mathcal{M} : X^n \to Z$ a randomized algorithm. We call $\mathcal{M}$ (ϵ, δ)-*differentially private *if for every two samples $S, S' \in X^n$, and every $z \subset Z$,*

$$\mathbb{P}[\mathcal{M}(S) \in z] \le e^{\epsilon} \cdot \mathbb{P}[M(S') \in z] + \delta.$$

If $\mathcal{M}$ is $(\epsilon, 0)$-private, we may simply call it ϵ-private.

Differential privacy comes with several guarantees useful for developing new mechanisms.

Proposition 5 (Adaptive composition [9, 10])**.** *Given parameters $0 < \epsilon < 1$ and $\delta > 0$, to ensure $(\epsilon, k\delta' + \delta)$-privacy over k adaptive mechanisms, it suffices that each mechanism is (ϵ', δ')-private, where*

$$\epsilon' = \frac{\epsilon}{2\sqrt{2k\log(1/\delta)}}.$$

A post-processing guarantee states that if the output of a private mechanism is manipulated in any way, without additional access to the input of the mechanism, then this additional processing will maintain privacy.

Lemma 6 (Post-processing [9]). *Let $\mathcal{M} : X^n \to Z$ be an (ϵ, δ)-private mechanism and $f : Z \to Z'$ a (possibly randomized) algorithm. Then $f \circ \mathcal{M}$ is (ϵ, δ)-private.*

In this paper, we use two well-established differentially-private mechanisms: the Laplace and exponential mechanisms. See Dwork & Roth [9] for more on differential privacy and these mechanisms.

3.1 The transfer theorem

A key method of Bassily et al. [1] for answering queries adaptively is a 'transfer theorem,' which states that if a mechanism is both accurate on a sample and differentially private, then it will be accurate on the sample's generating distribution.

Theorem 7 (Bassily et al. [1]). *Let $\mathcal{M}$ be a mechanism that on input sample $S \sim D^n$ answers k adaptively chosen low-sensitivity queries, is $(\frac{\alpha}{64}, \frac{\alpha\beta}{32})$-private for some $\alpha, \beta > 0$ and $(\frac{\alpha}{8}, \frac{\alpha\beta}{16})$-accurate on S. Then $\mathcal{M}$ is (α, β)-accurate on D.*

Their 'monitoring algorithm' proof technique involves a thought experiment in which an algorithm, called the monitor, assesses how accurately an input mechanism replies to an adversary, and remembers the query it does the worst on. It repeats this process some T times, and outputs the query that the mechanism does the worst on over all T rounds. Since the mechanism is private, so too is the monitor; and since privacy implies stability, this will ensure that the accuracy of the worst query is not too bad. For more details see Bassily et al. [1].

This transfer theorem only applies to the notion of accuracy for low-sensitivity queries, rather than accuracy for SCQ's. Consequently, we prove a monitor-based transfer theorem for SCQ's. We need two of Bassily et al.'s results for our transfer theorem. First, that for a monitoring algorithm $\mathcal{W}$, the expected value of the outputted query on the sample will be close to its expected value over the distribution—formalizing a connection between privacy and stability.

Lemma 8 (Bassily et al. [1]). *Let $\mathcal{W} : (X^n)^T \to Q \times [T]$ be (ϵ, δ)-private where Q is the class of counting queries. Let $S_i \sim D^n$ for each of $i \in [T]$ and $\mathbf{S} = \{S_1, \ldots, S_T\}$. Then*

$$|\mathbb{E}_{\mathbf{S},\mathcal{W}}[q(D)|(q,t) = \mathcal{W}(\mathbf{S})] - \mathbb{E}_{\mathbf{S},\mathcal{W}}[q(S_t)|(q,t) = \mathcal{W}(\mathbf{S})]| \leq e^{\epsilon} - 1 + T\delta.$$

We will also use a convenient form of accuracy bound for the exponential mechanism.

Lemma 9 (Bassily et al. [1]). *Let $\mathcal{R}$ be a finite set, $f : \mathcal{R} \to \mathbb{R}$ a function, and $\eta > 0$. Define a random variable X on $\mathcal{R}$ by $\mathbb{P}[X = r] = e^{\eta f(r)}/C$, where $C = \sum_{r \in \mathcal{R}} e^{\eta f(r)}$. Then*

$$\mathbb{E}[f(X)] \geq \max_{r \in \mathcal{R}} f(r) - \frac{1}{\eta} \log |\mathcal{R}|.$$

4 Fast mechanisms using sampling

In this section, we provide simple and fast mechanisms for answering low-sensitivity queries. We start with a mechanism that first samples without replacement. Sampling without replacement may take $O(\log n)$ time per sample, for a total of $O(\ell \log n)$ time over ℓ samples, in several settings. This is sufficient for an exponential-time speed up over the already-existing mechanisms that take $\Omega(n)$ time per query. However, this may come at the cost of space complexity, e.g. by keeping track of which elements have not been chosen so far [28]. Alternatively, there are methods that enjoy optimal space complexity at the cost of worst-case running times, as in rejection sampling [27]. This motivates our consideration of uniform sampling in Section 4.2.

4.1 Sampling without replacement and low-sensitivity queries

Our mechanism $\mathcal{M}_{wr}$ for answering low-sensitivity queries is very simple: Given a data set S of size n and query q, sample some ℓ points uniformly at random from S without replacement, and call this new set S_ℓ. Then the mechanism returns $q(S_\ell) + \mathrm{Lap}\left(\frac{1}{\ell\epsilon}\right)$, where $\mathrm{Lap}(b)$ refers to the zero-mean Laplacian distribution with scale parameter b. We may now state our main theorem for mechanism $\mathcal{M}_{wr}$, using suitable values for ϵ and ℓ.

Theorem 10. *When $\epsilon = O(1/\alpha)$ and $\ell \geq \frac{2\log(4k/\beta)}{\alpha^2}$ for k low-sensitivity queries,*

1. *$\mathcal{M}_{wr}$ takes $\tilde{O}\left(\frac{\log(k)\log(k/\beta)}{\alpha^2}\right)$ time per query.*
2. *$\mathcal{M}_{wr}$ is (α, β)-accurate (on the distribution) so long as $n = \Omega\left(\frac{\sqrt{k}\log k \cdot \log^{3/2}(\frac{1}{\alpha\beta})}{\alpha^2}\right)$.*

As mentioned above, sampling without replacement takes $O(\ell \log(n))$ time, which suffices to prove part one for the values of ℓ and n given. To prove the remaining part two, we must establish that sampling without replacement boosts privacy. If sampling were to only deliver $O(\epsilon)$ instead of $O(\frac{\ell}{n}\epsilon)$ privacy then we would need $\ell > \frac{2\sqrt{2k\log(1/\delta)}\log(2k/\beta)}{\alpha\epsilon}$, which would be undesirable—ℓ then becomes the size of the entire database and sampling yields no time savings over computing $q(S)$. Fortunately Lin et al. [22] provide a proof technique that may be readily adapted to this setting.

Proposition 11 (Adapted from Lin et al. [22])**.** *Given a mechanism $\mathcal{P} : X^\ell \to Y$, $\mathcal{M}$ will be the mechanism that does the following: Sample uniformly at random* without *replacement ℓ points from an input sample $S \in X^n$ of size n, and call this set S_ℓ. Output $\mathcal{P}(S_\ell)$. Then if $\mathcal{P}$ is ϵ-private, then $\mathcal{M}$ is (weakly) $\log(1 + \frac{\ell}{n}(e^\epsilon - 1)) = O\left(\frac{\ell}{n} \cdot \epsilon\right)$ private for $\ell \geq 1$.*

See the appendix for a proof. We may now return to the main theorem:

Proof of Theorem 10. Since the Laplace mechanism receives a sample S_ℓ of size ℓ, output a_q can be bounded with the standard accuracy result for the Laplace mechanism ensuring ϵ''-privacy for any $\epsilon'' > 0$:

$$\mathbb{P}[|a_q - q(S_\ell)| \geq \alpha/2] \leq e^{-\frac{\alpha\epsilon''\ell}{2}}.$$

We can bound this above by $\frac{\beta}{2k}$ provided $\epsilon'' \geq \frac{\log(2k/\beta)}{\ell\alpha}$; and this follows from a Chernoff bound

$$\mathbb{P}[|q(S_\ell) - q(S)| \geq \alpha/2] \leq e^{-\frac{\alpha^2\ell}{2}}.$$

Once again we can bound this above by $\frac{\beta}{2k}$ so long as $\ell \geq \frac{2\log(4k/\beta)}{\alpha^2}$.

Thus we have, for all q, $\mathbb{P}[|a_q - q(S)| \geq \alpha] \leq \mathbb{P}[|a_q - q(S_\ell)| \geq \alpha/2] + \mathbb{P}[|q(S_\ell) + q(S)| \geq \alpha/2] \leq \beta/k$. The union bound immediately yields (α, β)-accuracy over all k queries. From Proposition 11, we also have $\left(\frac{\ell}{n}\epsilon''\right)$-privacy, where $\frac{\ell}{n}\epsilon'' = \frac{\log(2k/\beta)}{n\alpha}$. Equivalently, we have ϵ'-privacy when $n \geq \frac{\log(2k/\beta)}{\epsilon'\alpha}$. With adaptive composition (Proposition 5), we can answer k queries with (ϵ, δ)-privacy when $\epsilon' = \frac{\epsilon}{2\sqrt{2k\log(1/\delta)}}$, resulting in (α, β)-accuracy and (ϵ, δ)-privacy on S so long as $n > \frac{2\sqrt{2k\log(1/\delta)}\log(2k/\beta)}{\alpha\epsilon}$. The proof is concluded by applying Theorem 7. □

4.2 Sampling with replacement and counting queries

In case sampling without replacement is prohibitive or simply unnecessary, we provide an alternative mechanism. We can sample with replacement instead, with mechanism $\mathcal{M}_r$ defined as: For each query q, sample some ℓ points uniformly at random with replacement, and call this new set S_ℓ. Then $\mathcal{M}_r$ returns $q(S_\ell) + \mathrm{Lap}\left(\frac{1}{\ell\epsilon}\right)$. Under counting queries, this mechanism works just as well as sampling without replacement.

Theorem 12. *When* $\epsilon = O(1/\alpha)$ *and* $\ell \geq \frac{2\log(4k/\beta)}{\alpha^2}$ *for* k *counting queries,*

1. $\mathcal{M}_r$ *takes* $\tilde{O}\left(\frac{\log(k)\log(k/\beta)}{\alpha^2}\right)$ *time per query.*
2. $\mathcal{M}_r$ *is* (α, β)*-accurate (on the distribution) so long as* $n = \Omega\left(\frac{\sqrt{k}\log k \cdot \log^{3/2}(\frac{1}{\alpha\beta})}{\alpha^2}\right)$.

Again, sampling takes $O(\ell \log(n))$ time, and othewise the proof remains the same as in Theorem 10, so it remains to prove the crux of this theorem: that sampling with replacement can not only maintain but boost privacy.

Proposition 13. *For any counting query* q*,* $\mathcal{M}_r$ *is* $\ell \log(1 + \frac{1}{n}(e^\epsilon - 1)) = O\left(\frac{\ell}{n} \cdot \epsilon\right)$ *private, for* $\ell \geq 1$.

In order to prove this, we need to bound the privacy loss incurred by the mechanism. Let S be the database of size n before sampling. For convenience, define $p = q(S)$. With $\mathcal{M}_r$ defined as above, we have

$$\mathbb{P}[\mathcal{M}_r(S) = z|S] = \sum_{y=0}^{\ell} \binom{\ell}{y} p^y (1-p)^{\ell-y} e^{-\epsilon\ell|z-y/\ell|}.$$

We will want to upper bound the log privacy loss by ϵ, and to do so, we will write the privacy loss as a function of z:

$$\mathrm{PL}(z) := \frac{\mathbb{P}[\mathcal{M}_r(S) = z|S]}{\mathbb{P}[\mathcal{M}_r(S') = z|S']} = \frac{\sum_{y=0}^{\ell} \binom{\ell}{y} p^y (1-p)^{\ell-y} e^{-\epsilon\ell|z-y/\ell|}}{\sum_{y=0}^{\ell} \binom{\ell}{y} p'^y (1-p')^{\ell-y} e^{-\epsilon\ell|z-y/\ell|}},$$

where $p = q(S)$ and $p' = q(S')$ for neighboring databases S and S'. Note we have $|p - p'| \leq 1/n$.

The proof of Proposition 13 proceeds in two parts: first, we show that $\mathrm{PL}(z)$ is monotonic as a function of z. This allows us to reduce the problem to showing that $\log(\mathrm{PL}(z)) \leq \epsilon$ on the boundary, namely when $z \leq 0$ or $z \geq 1$, which is much easier to compute.

Lemma 14. PL *is a monotonically non-increasing function of* z *(as* z *increases) if* $p' \geq p$ *and a monotonically non-decreasing function if* $p' \leq p$.

Proof. We examine when $\mathrm{PL}(z) \leq \mathrm{PL}(z - \delta)$ for $\delta > 0$. In particular, we need to show that $\mathrm{PL}(z) \leq \mathrm{PL}(z - \delta)$ when $p' > p$ and $\mathrm{PL}(z) \geq \mathrm{PL}(z - \delta)$ when $p' < p$. To do this, we split the sum into two pieces, a lower part and an upper part:

$$\begin{aligned} L_p(z, \delta) &:= \sum_{y=0}^{\lfloor \ell(z-\delta) \rfloor} \binom{\ell}{y} p^y (1-p)^{\ell-y} e^{-\epsilon\ell|z-y/\ell|} \\ U_p(z, \delta) &:= \sum_{y=\lfloor \ell(z-\delta) \rfloor + 1}^{\ell} \binom{\ell}{y} p^y (1-p)^{\ell-y} e^{-\epsilon\ell|z-y/\ell|}. \end{aligned}$$

If $\lfloor \ell(z-\delta) \rfloor < 0$, define $L_p(z, \delta) = 0$ and if $\lfloor \ell(z-\delta) \rfloor > \ell$, define $L_p(z, \delta)$ as the entire sum, i.e. $L_p(z, \delta) = P[\mathcal{M}_r(S) = z|S]$. Similarly, if $\lfloor \ell(z-\delta) \rfloor + 1 > \ell$, define $U_p(z, \delta) = 0$ and if $\lfloor \ell(z-\delta) \rfloor + 1 < 0$, define $U_p(z, \delta) = P[\mathcal{M}_r(S) = z|S]$. We can thus rewrite $\mathrm{PL}(z)$, for any δ, as

$$\mathrm{PL}(z) = \frac{L_p(z, \delta) + U_p(z, \delta)}{L_{p'}(z, \delta) + U_{p'}(z, \delta)}.$$

Writing PL(z) in this manner also allows us to write down PL($z-\delta$) as a function of L and U. In particular we can write $\mathbb{P}[\mathcal{M}_r(S) = z-\delta | S]$ as the following:

$$\begin{aligned}
&\sum_{y=0}^{\ell} \binom{\ell}{y} p^y (1-p)^{\ell-y} e^{-\epsilon|z-\delta-y/\ell|} \\
&= \sum_{y=0}^{\lfloor \ell(z-\delta) \rfloor} \binom{\ell}{y} p^y (1-p)^{\ell-y} e^{-\epsilon\ell(z-\delta-y/\ell)} + \sum_{y=\lfloor \ell(z-\delta)\rfloor+1}^{\ell} \binom{\ell}{y} p^y (1-p)^{\ell-y} e^{-\epsilon\ell(y/\ell-z+\delta)} \\
&= e^{\delta\epsilon\ell} \left(\sum_{y=0}^{\lfloor \ell(z-\delta) \rfloor} \binom{\ell}{y} p^y (1-p)^{\ell-y} e^{-\epsilon\ell(z-y/\ell)} \right) + e^{-\delta\epsilon\ell} \left(\sum_{y=\lfloor \ell(z-\delta)\rfloor+1}^{\ell} \binom{\ell}{y} p^y (1-p)^{\ell-y} e^{-\epsilon\ell(y/\ell-z)} \right) \\
&= e^{\delta\epsilon\ell} L_p(z,\delta) + e^{-\delta\epsilon\ell} U_p(z,\delta).
\end{aligned}$$

Now PL(z) $\le$ PL($z-\delta$) if and only if

$$\frac{L_p(z,\delta)+U_p(z,\delta)}{L_{p'}(z,\delta)+U_{p'}(z,\delta)} \le \frac{e^{\delta\epsilon\ell} L_p(z,\delta) + e^{-\delta\epsilon\ell} U_p(z,\delta)}{e^{\delta\epsilon\ell} L_{p'}(z,\delta) + e^{-\delta\epsilon\ell} U_{p'}(z,\delta)}.$$

Observing that $e^{\delta\epsilon\ell} - e^{-\delta\epsilon\ell} > 0$, canceling implies this in turn is equivalent to

$$L_{p'}(z,\delta) U_p(z,\delta) \le L_p(z,\delta) U_{p'}(z,\delta). \tag{1}$$

If it's not the case that $0 \le \lfloor \ell(z-\delta) \rfloor \le \ell - 1$, then either $L_{p'}(z,\delta) = L_p(z,\delta) = 0$ or $L_{p'}(z,\delta) = L_{p'}(z,\delta) = 0$, and we're done, so we may assume that $0 \le \lfloor \ell(z-\delta) \rfloor \le \ell-1$. Expanding, we can write the terms of (1) as

$$L_{p'}(z,\delta) U_p(z,\delta) = \sum_{\substack{a,b: \\ 0 \le a \le \lfloor \ell(z-\delta) \rfloor \\ \lfloor \ell(z-\delta)\rfloor+1 \le b \le \ell}} \binom{\ell}{a}\binom{\ell}{b} s_{a,b} \qquad \text{and} \qquad L_p(z,\delta) U_{p'}(z,\delta) = \sum_{\substack{a,b: \\ 0 \le a \le \lfloor \ell(z-\delta) \rfloor \\ \lfloor \ell(z-\delta)\rfloor+1 \le b \le \ell}} \binom{\ell}{a}\binom{\ell}{b} t_{a,b},$$

where

$$s_{ab} = \left((1-p')^{\ell-a} {p'}^a e^{-\epsilon\ell|z-a/\ell|} \right) \left((1-p)^{\ell-b} p^b e^{-\epsilon\ell|z-b/\ell|} \right),$$

and

$$t_{ab} = \left((1-p)^{\ell-a} p^a e^{-\epsilon\ell|z-a/\ell|} \right) \left((1-p')^{\ell-b} {p'}^b e^{-\epsilon\ell|z-b/\ell|} \right).$$

Note that by definition we have $a < \ell$ and $b > 0$. This allows us to quickly dispatch a few edge cases: If $p = 0$ (so necessarily $p' > p$) then $s_{a,b} = p^b = 0$, which in turn means we have $s_{a,b} \le t_{a,b}$. If $p = 1$ then $t_{a,b} = (1-p)^{\ell-a} = 0$, as desired. Likewise, if $p' = 0$ then $t_{a,b} = 0$ and if $p' = 1$ then $s_{a,b} = 0$, again as desired. Now we may assume $0 < p, p' < 1$, in which case we may cancel to get $\binom{\ell}{a}\binom{\ell}{b} s_{ab} \le \binom{\ell}{a}\binom{\ell}{b} t_{ab}$ if and only if

$$\left(\frac{p'(1-p)}{p(1-p')} \right)^a \le \left(\frac{p'(1-p)}{p(1-p')} \right)^b. \tag{2}$$

$\frac{p'(1-p)}{p(1-p')} \ge 1$ if and only if $p' \ge p$, and since $a < b$, (2) holds if and only if $p' \ge p$, completing the proof. □

We can now prove Proposition 13.

Proof of Proposition 13. With notation as above, if z is sufficiently small or large, then we can just use the binomial theorem:

$$\mathbb{P}[\mathcal{M}_r(S) = z|S] = \begin{cases} e^{\epsilon\ell z}(pe^{-\epsilon} + 1 - p)^\ell & \text{if } z \leq 0; \\ e^{-\epsilon\ell z}(pe^{\epsilon} + 1 - p)^\ell & \text{if } z \geq 1. \end{cases}$$

When $p' > p$, Lemma 14 implies for any z, $\mathrm{PL}(z) \leq \mathrm{PL}(\min(z, 0))$ and for $p' < p$, $\mathrm{PL}(z) \leq \mathrm{PL}(\max(z, 1))$.

First suppose $p' > p$. Then

$$\mathrm{PL}(z) \leq \mathrm{PL}(\min(z, 0)) \leq \frac{e^{\epsilon\ell z}(pe^{-\epsilon} + 1 - p)^\ell}{e^{\epsilon\ell z}(p'e^{-\epsilon} + 1 - p')^\ell} \leq \left(\frac{pe^{-\epsilon} + 1 - p}{p'e^{-\epsilon} + 1 - p'}\right)^\ell.$$

Recall that $p' \leq p + \frac{1}{n}$ and because $p = q(S)$, $p = \frac{i}{n}$ for some non-negative integer i. Then $p \leq \frac{n-1}{n}$ (because $p < p' \leq 1$). Hence

$$\begin{aligned} \log \mathrm{PL}(z) \leq \log\left(\left(\frac{pe^{-\epsilon} + 1 - p}{p'e^{-\epsilon} + 1 - p'}\right)^\ell\right) &= \ell \log\left(\frac{e^\epsilon - p(e^\epsilon - 1)}{e^\epsilon - p'(e^\epsilon - 1)}\right) \\ \leq \ell \log\left(\frac{e^\epsilon - p(e^\epsilon - 1)}{e^\epsilon - (p + 1/n)(e^\epsilon - 1)}\right) &= \ell \log\left(1 + \frac{\frac{1}{n}(e^\epsilon - 1)}{e^\epsilon - (p + 1/n)(e^\epsilon - 1)}\right) \\ \leq \ell \log\left(1 + \frac{1}{n}(e^\epsilon - 1)\right). \end{aligned}$$

The case when $p' < p$ proceeds in the same manner, except now $\mathrm{PL}(z) \leq \mathrm{PL}(\max(z, 1))$. □

5 Sampling counting queries

We now turn to sampling counting queries. Unlike in the previous section, we cannot leverage an existing transfer theorem, so instead we establish a new one.

Theorem 15. *Let $\mathcal{M}$ be a mechanism that on input sample $S \sim D^n$ answers k adaptively chosen sampling counting queries, is $(\frac{\alpha}{64}, \frac{\alpha\beta}{16})$-private for some $\alpha, \beta > 0$ and $(\alpha/2, 0)$-accurate on S. Suppose further that $n \geq \frac{1024\log(k/\beta)}{\alpha^2}$. Then $\mathcal{M}$ is (α, β)-accurate on D.*

This allows us to answer sampling counting queries:

Theorem 16. *There is a mechanism $\mathcal{M}$ that satisfies the following:*

1. *$\mathcal{M}$ takes $\tilde{O}\left(\log\left(\frac{k\log(\frac{1}{\beta})}{\alpha}\right)\right)$ time per query.*
2. *$\mathcal{M}$ is (α, β)-accurate on k sampling counting queries, where*

$$n \geq \Omega\left(\max\left(\frac{\sqrt{k\log(\frac{1}{\alpha\beta})}}{\alpha^2}, \frac{\log(k/\beta)}{\alpha^2}\right)\right) = \Omega\left(\frac{\sqrt{k}\log(\frac{1}{\alpha\beta})}{\alpha^2}\right).$$

We prove our transfer theorem using the following monitoring algorithm, which takes as input T sample sets, and outputs a query with probability proportional to how far away the query is on the sample as opposed to the distribution.

Definition 17 (Monitor with exponential mechanism)**.** *Define a monitoring algorithm $\mathcal{W}_D$ as the following: Given input $\mathbf{S} = \{S_1, \ldots, S_T\}$, for each of $t \in [T]$, simulate $\mathcal{M}(S_t)$ and $\mathcal{A}$ interacting, and let $q_{t,1}, \ldots, q_{t,k}$ be the queries of $\mathcal{A}$.*

Let $\mathcal{R} = \{(q_{t,i}, t)\}_{t\in[T], i\in[k]}$. Abusing notation, for each t and $i \in [k]$, consider the corresponding element $r_{t,i}$ of $\mathcal{R}$ and define the utility of $r_{t,i}$ as $u(\mathbf{S}, r_{t,i}) = |q_{t,i}(S_t) - q_{t,i}(D)|$. Release $r \in \mathcal{R}$ with probability proportional to $\exp\left(\frac{\epsilon \cdot n \cdot u(\mathbf{S}, r)}{2}\right)$.

Lemma 18. *If $\mathcal{M}$ is (ϵ, δ)-private for k queries, then $\mathcal{W}_D$ is $(2\epsilon, \delta)$-private.*

Proof. A single pertubation to $\mathbf{S}$ can only change one S_t, for some t. Then since $\mathcal{M}$ on S_t is (ϵ, δ)-private, $\mathcal{M}$ remains (ϵ, δ)-private over the course of the T simulations. Since $\mathcal{A}$ uses only the outputs of $\mathcal{M}$, $\mathcal{A}$ is just post-processing $\mathcal{M}$, and therefore it is (ϵ, δ)-private as well: releasing all of $\mathcal{R}$ remains (ϵ, δ)-private.

Since the sensitivity of u is $\Delta = 1/n$, the monitor is just using the exponential mechanism to release some $r \in \mathcal{R}$, which is ϵ-private. Using the standard composition theorem finishes the proof. ☐

We can now bound the probability that the query that the monitor outputs on the sample are far away from the distribution on both sides, if $\mathcal{M}$ is not accurate, by using both Lemmas 8 and 9.

Proof of Theorem 15. Consider the results for simulating T times the interaction between $\mathcal{M}$ and $\mathcal{A}$. Suppose for the sake of contradiction that $\mathcal{M}$ is not (α, β)-accurate on D. Then for every i in $[k]$ and t in T, since $|\mathbb{E}_{\mathcal{M}}[\mathcal{M}(q_{t,i})] - q(S_t)| \leq \alpha/2$, we have

$$\mathbb{P}_{S_t, \mathcal{M}, \mathcal{A}}\left[\max_i |q_{t,i}(S_t) - q_{t,i}(D)| > \alpha/2\right] > \beta.$$

Call some q and t that achieves the maximum $|q(S_t) - q(D)|$ over the T independent rounds of $\mathcal{M}$ and $\mathcal{A}$ interacting, as $\mathcal{W}_D$ does, by q_w and t_w. Since each round t is independent, the probability that $|q_w(S_{t_w}) - q_w(D)| \leq \alpha/2$ is then no more than $(1-\beta)^T$. Then using Markov's inequality immediately grants us that

$$\mathbb{E}_{\mathbf{S}, \mathcal{W}_D}\left[|q_w(S_{t_w}) - q_w(D)|\right] > \frac{\alpha}{2}(1 - (1-\beta)^T). \tag{3}$$

Let $\Gamma = \mathbb{E}_{\mathbf{S}, \mathcal{W}_D}\left[|q^*(S_{t^*}) - q^*(D)| : (q^*, t^*) = \mathcal{W}_D(\mathbf{S})\right]$.

Setting $f(r) = u(\mathbf{S}, r)$, Lemma 9 implies that under the exponential mechanism, we have

$$\mathbb{E}[|q^*(S_{t^*}) - q^*(D)| : (q^*, t^*) = \mathcal{W}_D(\mathbf{S})] \geq |q_w(S_{t_w}) - q_w(D)| - \frac{2}{\epsilon n}\log(kT).$$

Taking the expected value of both sides with respect to $\mathbf{S}$ and the randomness of the rest of $\mathcal{W}_D$, we obtain

$$\begin{aligned} \Gamma &\geq \mathbb{E}_{\mathbf{S}, \mathcal{W}_D}[|q_w(S_{t_w}) - q_w(D)|] - \frac{2}{\epsilon n}\log(kT) \\ &> \frac{\alpha}{2}(1 - (1-\beta)^T) - \frac{2}{\epsilon n}\log(kT), \end{aligned} \tag{4}$$

which follows from employing Equation (3). On the other hand, suppose that $\mathcal{M}$ is (ϵ, δ)-private for some $\epsilon, \delta > 0$. Then by Lemma 18, $\mathcal{W}_D$ is $(2\epsilon, \delta)$-private, and then in turn Lemma 8 implies that

$$\Gamma \leq e^{2\epsilon} - 1 + T\delta. \tag{5}$$

We will now ensure (4) $\geq \alpha/8$ and (5) $\leq \alpha/8$, a contradiction. Set $T = \lfloor \frac{1}{\beta} \rfloor$ and $\delta = \frac{\alpha\beta}{16}$. Then

$$e^{2\epsilon} - 1 + T\delta \leq e^{2\epsilon} - 1 + \alpha/16 \leq \alpha/8$$

when $e^{2\epsilon} - 1 \leq \alpha/16$, which in turn is satisfied when $\epsilon \leq \alpha/64$, since $0 \leq \alpha \leq 1$.

On the other side, $1 - (1-\beta)^{\lfloor \frac{1}{\beta} \rfloor} \geq 1/2$. Then it suffices to set ϵ such that $\frac{2}{\epsilon n}\log(kT) \leq \alpha/8$. Thus we need ϵ such that

$$\frac{16\log(k/\beta)}{\alpha n} \;\leq\; \epsilon \;\leq\; \alpha/64.$$

Such an ϵ exists, since we explicitly required $n \geq \frac{1024\log(k/\beta)}{\alpha^2}$. □

With a transfer theorem in hand, we now introduce a private mechanism that is accurate on a sample for answering sampling counting queries.

Lemma 19 (SCQ mechanism)**.** *For $\epsilon \leq 1$, There is an (ϵ, δ)-private mechanism to release k SSQs that is $(\alpha, 0)$-accurate, for $\alpha \leq 1/2$, with respect to a fixed sample S of size n so long as*

$$n > \frac{2\sqrt{2k\log(1/\delta)}}{\alpha\epsilon}.$$

Proof. We design a mechanism $\mathcal{M}$ to release a $(\alpha, 0)$-accurate SCQ for $n > \frac{1}{\alpha\epsilon}$ and then use Proposition 5. The mechanism is simple: sample x i.i.d. from S. Then release $q(x)$ with probability $1-\alpha$ and $1-q(x)$ with probability α. Let $i = \sum_{x\in S} q(x)$. Then $\mathbb{E}_{\mathcal{M}}[\mathcal{M}(q)] = \frac{(1-\alpha)i + \alpha(n-i)}{n} = \frac{i}{n} + \alpha\left(\frac{n-2i}{n}\right)$, so $\frac{i}{n} - \alpha \leq \mathbb{E}_{\mathcal{M}}[\mathcal{M}(q)] \leq \frac{i}{n} + \alpha$, implying that $\mathcal{M}$ is $(\alpha, 0)$-accurate on S.

Now let S' differ from S on one element x, where $q(x) = 0$ but for $x' \in S'$, $q(x') = 1$. Consider

$$\frac{\mathbb{P}[\mathcal{M}(S) = 1]}{\mathbb{P}[\mathcal{M}(S') = 1]} = \frac{(1-\alpha)\frac{i+1}{n} + \alpha(\frac{n-i+1}{n})}{(1-\alpha)\frac{i}{n} + \alpha(\frac{n-i}{n})} = 1 + \frac{1-2\alpha}{i - 2\alpha i + \alpha n},$$

for $i = 0$ to $i = n-1$. The other cases are similar. Note this is at least 1 since $1 - 2\alpha \geq 0$. Thus it suffices to show when this is upper-bounded by e^{ϵ}. By computing the partial derivative with respect to i, it is easy to see that it suffices to consider the cases when $i = 0$ or $i = n-1$. When $i = 0$,

$$\log\left(\frac{\mathbb{P}[\mathcal{M}(S) = 1]}{\mathbb{P}[\mathcal{M}(S') = 1]}\right) \leq \frac{1-2\alpha}{\alpha n} \leq \frac{1}{\alpha n} \leq \epsilon$$

when $n \geq \frac{1}{\epsilon\alpha}$. When $i = n-1$,

$$\log\left(\frac{\mathbb{P}[\mathcal{M}(S) = 1]}{\mathbb{P}[\mathcal{M}(S') = 1]}\right) \leq \frac{1-2\alpha}{n(1-\alpha) - (1-2\alpha)} \leq \epsilon$$

when $n \geq \frac{(1-2\alpha)(\epsilon+1)}{(1-\alpha)\epsilon}$ but because $\frac{1-2\alpha}{1-\alpha} \leq 1$, it suffices to set $n \geq 1 + \frac{1}{\epsilon}$. The proof is completed by noting that $\frac{1}{\epsilon\alpha} \geq 1 + \frac{1}{\epsilon}$ because $\epsilon \leq 1$. □

We now use this mechanism to answer sampling counting queries.

Proof of Theorem 16. We use the mechanism of Lemma 19. This gives an (ϵ, δ)-private mechanism that is $(\alpha/2, 0)$-accurate so long as $n \geq \frac{4\sqrt{2k\log(1/\delta)}}{\alpha\epsilon}$.

Setting ϵ and δ as required by Theorem 15 implies that we need $n \geq \Omega\left(\sqrt{k\log(\frac{1}{\alpha\beta})}/\alpha^2\right)$.

Note to use Theorem 15 we also need $n \geq \Omega\left(\log(k/\beta)/\alpha^2\right)$. The sample complexity follows. This mechanism samples a single random point, which takes $O(\log(n))$ time, completing the proof. □

6 Comparing counting and sampling counting queries

A natural arises: is one of our methods redundant? Can we use a mechanism for SCQ's to simulate a mechanism for counting queries, or vice-versa? We now show that the natural approach to simulate a counting query with SCQ's results in an extra $O(1/\alpha)$ factor (although enjoying a slightly better dependence on k). This represents a $O(1/\alpha)$ overhead in order to ensure that the mechanism returns meaningful results for all sample sizes ℓ.

Proposition 20. *Using ℓ SCQ's to estimate each counting query is an (α, β)-accurate mechanism for k counting queries if $\ell \geq \frac{2\log(4k/\beta)}{\alpha^2}$ and $n = \Omega\left(\frac{\sqrt{k\log k}\log^{3/2}(\frac{1}{\alpha\beta})}{\alpha^3}\right)$.*

Proof. The mechanism, for each query q, will query the SCQ mechanism $\mathcal{M}$ described in Section 5 ℓ times with the query q, and return the average, call this a_q. Note that $\mathbb{E}[a_q] = \mathbb{E}[\mathcal{M}(q)]$. Since each SCQ is independent of each other, a Chernoff bound gives $\mathbb{P}[|a_q - \mathbb{E}[a_q]| \geq \alpha/2] \leq 2e^{-\ell\alpha^2/2} \leq \beta/2k$ when $\ell \geq \frac{2\log(4k/\beta)}{\alpha^2}$. Using Theorem 16, as long as

$$n = \Omega\left(\frac{\sqrt{k\ell}\log(\frac{1}{\alpha\beta})}{\alpha^2}\right),$$

we have that $\mathbb{P}[\max_q |\mathbb{E}[\mathcal{M}(q)] - q(D)| \geq \alpha/2] \leq \beta/2$, over all $k\ell$ of them. Then the union bound implies that

$$\begin{aligned}
\mathbb{P}[\max_q |a_q - q(D)| \geq \alpha] &\leq \mathbb{P}[\max_q |a_q - \mathbb{E}[\mathcal{M}(q)]| + |\mathbb{E}[\mathcal{M}(q)] - q(D)| \geq \alpha] \\
&\leq \beta/2 + \beta/2 \leq \beta.
\end{aligned}$$

Plugging in ℓ into the above expression for n completes the proof. □

On the other hand, even if willing to ignore time constraints one might attempt to use a mechanism for counting queries to attempt to answer SCQ's. Indeed, there is the naive approach of first computing $q(S)$, adding noise to obtain a value $\tilde{a_q}$, and then returning 1 with probability $\tilde{a_q}$ and 0 otherwise. For this mechanism we obtain an (ϵ, δ)-private mechanism to release k SCQ's that is (α, β)-accurate with respect to a fixed sample S of size n so long as

$$n > \frac{2\sqrt{2k\log(1/\delta)}\log(1/\beta)}{\alpha\epsilon},$$

which is strictly worse than the mechanism for SCQ's we actually use. This motivates our approach to SCQ's.

7 Future work

In this paper, we have introduced new faster mechanisms that take advantage of sampling's simultaneous ability to boost privacy while decreasing running time. This prompts several interesting questions: In what other adaptive settings can sampling help as much as it does in this work? This would presumably mean expanding results on the effect of sampling on privacy. Does sampling with replacement boost privacy for all low-sensitivity queries? Do our results that say pure privacy ($\delta = 0$) can be boosted by fast sampling methods generalize to approximate privacy?

More broadly, our work opens up the area of designing sublinear algorithms for adaptive data analysis. Such analysis is frequently achieved in settings where fast algorithms are frequently required, such as sublinear algorithms for streaming, online, on large data sets, etc. How can fast algorithms for adaptive analysis be developed in these types of settings?

8 Acknowledgements

Benjamin Fish was supported in part by the NSF EAPSI fellowship and NSF grant IIS-1526379. Lev Reyzin was supported in part by NSF grant IIS-1526379. Benjamin Rubinstein acknowledges support of the Australian Research Council (DP150103710).

References

[1] Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pages 1046–1059, 2016.

[2] Avrim Blum, Merrick L. Furst, Jeffrey C. Jackson, Michael J. Kearns, Yishay Mansour, and Steven Rudich. Weakly learning DNF and characterizing statistical query learning using fourier analysis. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing, 23-25 May 1994, Montréal, Québec, Canada*, pages 253–262, 1994.

[3] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing, Victoria, British Columbia, Canada, May 17-20, 2008*, pages 609–618, 2008.

[4] Mark Bun, Jonathan Ullman, and Salil P. Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the 46th Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 1–10, 2014.

[5] Kamalika Chaudhuri and Nina Mishra. When random sampling preserves privacy. In *Proceedings of the 26th Annual International Cryptology Conference, CRYPTO 2006, Santa Barbara, California, USA, August 20-24, 2006, Proceedings*, pages 198–213, 2006.

[6] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems 28, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2350–2358, 2015.

[7] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 117–126, 2015.

[8] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006*, pages 265–284, 2006.

[9] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[10] Cynthia Dwork, Guy N. Rothblum, and Salil P. Vadhan. Boosting and differential privacy. In *51th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2010, October 23-26, 2010, Las Vegas, Nevada, USA*, pages 51–60, 2010.

[11] Hamid Ebadi, Thibaud Antignac, and David Sands. Sampling and partitioning for differential privacy. In *14th Annual Conference on Privacy, Security and Trust, PST 2016, Auckland, New Zealand, December 12-14, 2016*, pages 664–673, 2016.

[12] Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Srinivas Vempala, and Ying Xiao. Statistical algorithms and a lower bound for detecting planted cliques. *Journal of the ACM*, 64(2):8:1–8:37, 2017.

[13] Andrew Gelman and Eric Loken. The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *American Scientist*, 102(6):460–465, 2014.

[14] Nadav Golbandi, Yehuda Koren, and Ronny Lempel. Adaptive bootstrapping of recommender systems using decision trees. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, pages 595–604, 2011.

[15] Zach Jorgensen, Ting Yu, and Graham Cormode. Conservative or liberal? Personalized differential privacy. In *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, pages 1023–1034, 2015.

[16] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. What can we learn privately? In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, October 25-28, 2008, Philadelphia, PA, USA*, pages 531–540, 2008.

[17] Michael J. Kearns. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6):983–1006, 1998.

[18] Georgios Kellaris and Stavros Papadopoulos. Practical differential privacy via grouping and smoothing. *PVLDB*, 6(5):301–312, 2013.

[19] Ariel Kleiner, Ameet Talwalkar, Purnamrita Sarkar, and Michael I. Jordan. The big data bootstrap. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

[20] Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927, 06 2016.

[21] Ninghui Li, Wahbeh H. Qardaji, and Dong Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *7th ACM Symposium on Information, Compuer and Communications Security, ASIACCS '12, Seoul, Korea, May 2-4, 2012*, pages 32–33, 2012.

[22] Bing-Rong Lin, Ye Wang, and Shantanu Rane. On the benefits of sampling in privacy preserving statistical analysis on distributed databases. *arXiv preprint arXiv:1304.4613*, 2013.

[23] Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA, June 11-13, 2007*, pages 75–84, 2007.

[24] Ryan M. Rogers, Aaron Roth, Adam D. Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *57th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 487–494, 2016.

[25] Thomas Steinke and Jon Ullman. Between pure and approximate differential privacy. In *Theory and Practice of Differential Privacy (TPDP 2015), London, UK*, 2015.

[26] Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Proceedings of the 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, pages 1588–1628, 2015.

[27] Jeffrey Scott Vitter. Faster methods for random sampling. *Communications of the ACM*, 27(7):703–718, 1984.

[28] C. K. Wong and Malcolm C. Easton. An efficient method for weighted sampling without replacement. *SIAM Journal of Computing*, 9(1):111–113, 1980.

[29] Houping Xiao, Jing Gao, Qi Li, Fenglong Ma, Lu Su, Yunlong Feng, and Aidong Zhang. Towards confidence in the truth: A bootstrapping based truth discovery approach. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1935–1944, 2016.

[30] Ke Yang. On learning correlated Boolean functions using statistical queries. In *Proceedings of the 12th International Conference on Algorithmic Learning Theory ALT 2001, Washington, DC, USA, November 25-28, 2001*, pages 59–76, 2001.

A Proof of Proposition 11

Let S and S' be samples of size that differ only on the kth index. Consider the set of all subsamples of the indices $[n]$ of size ℓ:

$$\mathcal{R} = \{\pi : \pi = \{i_1, \ldots, i_k\} \subset [n]\}.$$

Under uniform sampling without replacement, we choose uniformly at random a subsample π from $\mathcal{R}$. For any index k, either k is in π or π differs in exactly one element from some subsample that includes k. In particular, any π not including k is distance one away from exactly ℓ subsamples with k: namely, the subsamples that replace each element of π with k. Abusing notation, we will identify the subsample of indices with the corresponding subsample of S (and likewise with S'), so that $\mathbb{P}[\mathcal{P}(\pi) = z|S]$ refers to the probability that mechanism $\mathcal{P}$ outputs z given the subsample π of S. Then for any output z in Y, where $d(\pi, \pi') = 1$ denotes two subsamples differing in exactly one element,

$$\begin{aligned}
\mathbb{P}[\mathcal{M}(S) = z|S] &= \frac{1}{|\mathcal{R}|} \sum_{\pi \in \mathcal{R}} \mathbb{P}[\mathcal{P}(\pi) = z|S] \\
&= \frac{1}{|\mathcal{R}|} \left(\sum_{\pi \in \mathcal{R}: k \in \pi} \mathbb{P}[\mathcal{P}(\pi) = z|S] + \frac{1}{\ell} \sum_{\pi \in \mathcal{R}: k \in \pi} \sum_{\pi': k \notin \pi, d(\pi,\pi')=1} \mathbb{P}[\mathcal{P}(\pi') = z|S] \right) \\
&= \frac{1}{|\mathcal{R}|} \sum_{\pi \in \mathcal{R}: k \in \pi} \left(\mathbb{P}[\mathcal{P}(\pi) = z|S] + \frac{1}{\ell} \sum_{\pi': k \notin \pi, d(\pi,\pi')=1} \mathbb{P}[\mathcal{P}(\pi') = z|S] \right).
\end{aligned}$$

That is,

$$\begin{aligned}
\frac{\mathbb{P}[\mathcal{M}(S) = z|S]}{\mathbb{P}[\mathcal{M}(S') = z|S']} &= \frac{\sum_{\pi \in \mathcal{R}: k \in \pi} \left(\mathbb{P}[\mathcal{P}(\pi) = z|S] + \frac{1}{\ell} \sum_{\pi': k \notin \pi, d(\pi,\pi')=1} \mathbb{P}[\mathcal{P}(\pi') = z|S] \right)}{\sum_{\pi \in \mathcal{R}: k \in \pi} \left(\mathbb{P}[\mathcal{P}(\pi) = z|S'] + \frac{1}{\ell} \sum_{\pi': k \notin \pi, d(\pi,\pi')=1} \mathbb{P}[\mathcal{P}(\pi') = z|S'] \right)} \\
&\le \max_{\pi \in \mathcal{R}: k \in \pi} \frac{\mathbb{P}[\mathcal{P}(\pi) = z|S] + \frac{1}{\ell} \sum_{\pi': k \notin \pi, d(\pi,\pi')=1} \mathbb{P}[\mathcal{P}(\pi') = z|S]}{\mathbb{P}[\mathcal{P}(\pi) = z|S'] + \frac{1}{\ell} \sum_{\pi': k \notin \pi', d(\pi,\pi')=1} \mathbb{P}[\mathcal{P}(\pi') = z|S']},
\end{aligned}$$

where we bound the ratio of sums by the maximum ratio. Name π^* the maximizer of this ratio. Now we bound the numerator and denominator via privacy:

Fix $\pi \in \mathcal{R}$ with $k \in \pi$. Firstly, we have $\mathbb{P}[\mathcal{P}(\pi) = z|S] \le e^{\epsilon} \mathbb{P}[\mathcal{P}(\pi) = z|S']$. Secondly, for π' such that $k \notin \pi'$ but $d(\pi, \pi') = 1$ we have $\mathbb{P}[\mathcal{P}(\pi) = z|S] \le e^{\epsilon} \mathbb{P}[\mathcal{P}(\pi') = z|S]$. Finally, $\mathbb{P}[\mathcal{P}(\pi') = z|S] = \mathbb{P}[\mathcal{P}(\pi') = z|S']$ since $k \notin \pi'$. Thus

$$
\begin{aligned}
\frac{\mathbb{P}[\mathcal{M}(S) = z|S]}{\mathbb{P}[\mathcal{M}(S') = z|S']} &\leq \frac{\mathbb{P}[\mathcal{P}(\pi^*) = z|S] + \frac{1}{\ell}\sum_{\pi':k\notin\pi',d(\pi^*,\pi')=1}\mathbb{P}[\mathcal{P}(\pi') = z|S]}{\mathbb{P}[\mathcal{P}(\pi^*) = z|S'] + \frac{1}{\ell}\sum_{\pi':k\notin\pi',d(\pi^*,\pi')=1}\mathbb{P}[\mathcal{P}(\pi') = z|S]} \\
&= 1 + \frac{\mathbb{P}[\mathcal{P}(\pi^*) = z|S] - \mathbb{P}[\mathcal{P}(\pi^*) = z|S']}{\mathbb{P}[\mathcal{P}(\pi^*) = z|S'] + \frac{1}{\ell}\sum_{\pi':k\notin\pi',d(\pi^*,\pi')=1}\mathbb{P}[\mathcal{P}(\pi') = z|S]} \\
&\leq 1 + \frac{\mathbb{P}[\mathcal{P}(\pi^*) = z|S] - e^{-\epsilon}\mathbb{P}[\mathcal{P}(\pi^*) = z|S]}{e^{-\epsilon}\mathbb{P}[\mathcal{P}(\pi^*) = z|S] + \frac{1}{\ell}\sum_{\pi':k\notin\pi',d(\pi^*,\pi')=1}e^{-\epsilon}\mathbb{P}[\mathcal{P}(\pi^*) = z|S]} \\
&= 1 + \frac{1 - e^{-\epsilon}}{e^{-\epsilon} + \frac{1}{\ell}\sum_{\pi':k\notin\pi',d(\pi^*,\pi')=1}e^{-\epsilon}} \\
&= 1 + \frac{\ell}{n}(e^{\epsilon} - 1),
\end{aligned}
$$

where the second inequality uses privacy and the last equality follows from the fact that there are $n - \ell$ such π' where $k \notin \pi'$ but $d(\pi^*, \pi') = 1$. This completes the proof.