

A simple and practical algorithm for differentially private data release

Moritz Hardt*

Katrina Ligett[†]

Frank McSherry

December 19, 2018

Abstract

We present new theoretical results on differentially private data release useful with respect to any target class of counting queries, coupled with experimental results on a variety of real world data sets.

Specifically, we study a simple combination of the multiplicative weights approach of [Hardt and Rothblum, 2010] with the exponential mechanism of [McSherry and Talwar, 2007]. The multiplicative weights framework allows us to maintain and improve a distribution approximating a given data set with respect to a set of counting queries. We use the exponential mechanism to select those queries most incorrectly tracked by the current distribution. Combining the two, we quickly approach a distribution that agrees with the data set on the given set of queries up to small error.

The resulting algorithm and its analysis is simple, but nevertheless improves upon previous work in terms of both error and running time. We also empirically demonstrate the practicality of our approach on several data sets commonly used in the statistical community for contingency table release.

*Center for Computational Intractability, Department of Computer Science, Princeton University. Supported by NSF grants CCF-0426582 and CCF-0832797. Email: mhardt@cs.princeton.edu.

[†]Computer Science Department, Cornell University. Work supported in part by an NSF Computing Innovation Fellowship (NSF Award CNF-0937060) and an NSF Mathematical Sciences Postdoctoral Fellowship (NSF Award DMS-1004416). Email: katrina@cs.cornell.edu.

1 Introduction

Sensitive statistical data on individuals are ubiquitous, and publishable analysis of such private data is an important objective in many settings. When releasing synthetic data or informative statistics based on sensitive data sets, one must balance the inherent tradeoff between the usefulness of the released information and the privacy loss of the affected individuals. Against this backdrop, differential privacy [DN03, DN04, DMNS06] has emerged as a compelling privacy definition that allows us to understand this tradeoff in the context of formal, provable privacy guarantees. In recent years, there has been a flourishing theoretical literature on differential privacy, providing ever stronger techniques for achieving the definition, in an ever growing space of applications.

Applications of the theory of differentially private data analysis to real-world data have received less attention, and have so far met with mixed success. Eager data analysts have on some applications found that the current state of the art (or most accessible) in algorithms for achieving differential privacy add unacceptable levels of noise; such situations are not evidence that differential privacy is an inappropriate definition, but rather that we need to develop and implement better algorithms, perhaps tailored to the application at hand. In many cases, existing theoretical results have been focused on demonstrating good asymptotic worst-case behavior, but with little regard for constant factors or performance on realistic database sizes.

In this work, we present new theoretical results for producing differentially private data useful with respect to a target class of queries, coupled with experimental results on a variety of real world data sets. Our approach uses the exponential mechanism [MT07] to implement randomized query selection in the framework of the multiplicative weights method employed by Hardt and Rothblum [HR10]. The combination of these two techniques achieves improved privacy-utility tradeoffs, improving the state of the art. The resulting algorithm also allows a much simpler analysis than the original multiplicative weights approach.

The multiplicative weights approach maintains a candidate output at all times, represented as a distribution over the space of possible data points. The quality of the distribution when compared with the true database is repeatedly improved by a procedure that selects a query from the target class, and reweights the distribution to improve its fidelity on the selected query. In our algorithm, presented in Section 3, we use the exponential mechanism as a means to bias our selection of the next query towards one that will provide the most improvement in our distribution.

Experimentally, we test our algorithm on four standard data sets, the first three of which are also studied by Fienberg et al. [FRY10] (see their paper for further discussion of the data sets and references to additional work on this data): (1) an epidemiological study of Czechoslovakian car factory workers intended to investigate risk factors for coronary thrombosis; (2) a 1990 genetic study of barley powdery mildew isolates using DNA markers; (3) data relating household characteristics, women’s economic activity, and husband’s unemployment, in households in the city of Rochdale; and (4) the National Long Term Care Survey, a longitudinal study of the health of older Americans, based a sample of tens of thousands of Medicare enrollees.¹ On each of these data sets, we present experimental results that demonstrate the tradeoff between the differential privacy parameter and the accuracy (as measured by relative entropy) of the resulting data, comparing the results when privacy is achieved by each of three different differentially private algorithms: (1) the original approach of Barak et al. [BCD⁺07] for producing synthetic contingency table data (this approach is analogous to the experiments undertaken in Fienberg et al. [FRY10]), (2) our algorithm combining multiplicative weights with the exponential mechanism as described above, with no specialized optimization, and (3) our algorithm, with a number of additional optimizations described in Section 4.

¹See <http://www.nltcs.aas.duke.edu/> for more details on the survey and associated data.

1.1 Our results

We will state our main theorems informally here. A formal statement is given in Section 3.3. To understand the error bounds, note that we normalize a counting query f so that its value on a data set D , denoted $f(D)$, is between 0 and 1.

Theorem 1.1 (informal). *Given a data set $D \subseteq X$ of size $|D| = n$ and a set of counting queries Q , we can compute an ϵ -differentially private data distribution D^* so that with high probability D^* satisfies*

$$\forall f \in Q: |f(D) - f(D^*)| \leq O\left(\frac{\log |X| \log |Q|}{\epsilon n}\right)^{1/3}.$$

The formal statement is given in Theorem 3.3. We can obtain a stronger bound on the error by allowing (ϵ, δ) -differential privacy.

Theorem 1.2 (informal). *Given a data set $D \subseteq X$ of size $|D| = n$ and a set of counting queries Q , we can compute an (ϵ, δ) -differentially private data distribution D^* so that with high probability D^* satisfies*

$$\forall f \in Q: |f(D) - f(D^*)| \leq O\left(\frac{\sqrt{\log |X| \log(1/\delta)} \log |Q|}{\epsilon n}\right)^{1/2}.$$

See Theorem 3.4 for a formal statement. The failure probability in both theorems can be bounded by $1/\text{poly}(|Q|)$ without changing the stated bounds as shown in Section 3.1. It can be reduced even further at a small loss in accuracy. The exact dependence is omitted from the theorems. We also remark that the runtime of our algorithm nearly matches the cryptographic hardness results of [DNR⁺09].

Experimental evaluation. Our experimental observations bear out the significance of choosing to take only the most significant measurements, at improved accuracy. On several real datasets, our algorithms yield marked improvement over the prior naive approaches of taking all measurements one seeks preserve. The improvement is most significant when privacy constraints are strong and the query class is rich; in applications where one can afford to simply take all measurements at sufficient accuracy, careful selection is not helpful. Fortunately, the former setting is the most important and challenging for resolving the practical tension between privacy and utility. A detailed discussion is presented in Section 4.

1.2 Comparison to previous work

The work of Barak et al. [BCD⁺07] was the first to address the problem of generating synthetic databases that preserve differential privacy. Their algorithm, which maintains utility with respect to a set of marginals (as opposed to general counting queries), essentially computes the desired noisy marginals and then solves the linear program constrained by these noisy marginals in order to obtain consistent data. This approach identifies all measurements required to reproduce the marginals, and takes each with a uniform level of accuracy. This may make a large number of redundant or uninformative measurements at the expense of accuracy in the more interesting queries. Fienberg et al. [FRY10] observe that, on realistic data sets, the Barak et al. algorithm must add so much noise to preserve differential privacy that the resulting data is no longer useful.

The study of differentially private synthetic data release mechanisms for arbitrary counting queries began with the work of Blum, Ligett, and Roth [BLR08], who gave a computationally inefficient (superpolynomial in $|X|$) algorithm that achieves error that scales only logarithmically with the number of queries. The dependence on the size of the data set n achieved by their algorithm is $n^{-1/3}$.

Li et al. [LHR⁺10] investigated an approach to answering sets of counting queries, which selects an appropriate basis to cover the target set of counting queries and reconstructs answers from this basis. They are primarily concerned with settings of $|Q| \gg |D|$. Our algorithm, when applied to the specific query sets for which they state results, reduces the maximum error rates substantially. For example, for the case of $|Q| = 2^{|X|}$ (corresponding to the set of all $(0, 1)$ -counting queries), the dependence on $|X|$ goes from $|X| \log^2 |X|$ to $|X|^{1/3} \log^{1/3} |X|$.

Since [BLR08], subsequent work has focused on computationally more efficient algorithms (here meaning polynomial in $|X|$). This line of work has yielded error rates of $(|Q|^{o(1)} / \sqrt{n})$ [DNR⁺09] and $\text{poly}(\log |Q|) / \sqrt{n}$ [DRV10] for a relaxed privacy guarantee known as (ϵ, δ) -differential privacy. Hardt and Rothblum [HR10] then introduced the private multiplicative weights mechanism, and showed that it provides a means of achieving error rates of $\log |Q| / \sqrt{n}$ for (ϵ, δ) -differential privacy. A feature of the multiplicative weights algorithm of Hardt and Rothblum [HR10] is that it answers a set of $|Q|$ counting queries in the *interactive* setting.² Still, the algorithm can also be used non-interactively to produce synthetic data. Indeed, by asking the entire set of queries roughly n times repeatedly, one can ensure that on at least one of the iterations there are no large errors. In this case the multiplicative weights distribution represents synthetic data that is correct (up to the desired error) on all $|Q|$ queries.

The error bound stated in Theorem 1.2 improves over [HR10] by a factor of $O(\sqrt{\epsilon^{-1} \log |Q|} \cdot \log^{3/4}(1/\delta))$. We note, however, that their algorithm works in the interactive setting and was not optimized with regards to building synthetic data. We also remark that (even though not stated in their work) the Hardt-Rothblum algorithm can be used to achieve $(\epsilon, 0)$ -differential privacy with an error bound of $O(1/n^{1/3})$ ignoring all other parameters. This gives a bound comparable to that in Theorem 1.1.

The current work builds on [HR10] in providing improved error bounds and runtime for the task of creating synthetic data. (Our results can be interpreted as providing synthetic data, by means of sampling the private distribution we generate.) In addition, our algorithm and its analysis are both significantly simpler. Further, unlike all the previous work mentioned above, our new results are tuned for practical applications, and we arguably provide the first empirical results demonstrating that it is possible to produce useful differentially private synthetic data for real-world statistical applications.

Our algorithm was also inspired by a recent work of Gupta et al. [GHRU10] who showed that agnostic learning algorithms can be used as a subroutine in the multiplicative weights framework to select queries with near maximal error. The use of an agnostic learner in their work is analogous to the use of the exponential mechanism in our algorithm.

2 Preliminaries

We have a data set $D \subseteq X$ of size n over a universe X . We will think of the data set as a distribution over X . A *counting query* is specified by a function $f: X \rightarrow [0, 1]$. With some abuse of notation we define $f(D) = \mathbb{E}_{x \sim D} f(x)$ and refer to f itself as a counting query. With this normalization a counting query has *sensitivity* $1/n$.

A *synthetic data* mechanism is a randomized mapping from $D \subseteq X$ to the space $\Delta(X)$ of distributions over X . We say that such a mechanism M preserves (ϵ, δ) -differential privacy if for every $S \subseteq \Delta(X)$ and for all data sets $D, D' \subseteq X$ differing in only one element, we have

$$\Pr[M(D) \in S] \leq e^\epsilon \Pr[M(D') \in S] + \delta. \quad (1)$$

When $\delta = 0$, we will say that such a mechanism satisfies ϵ -differential privacy.

²Here, the algorithm receives queries one at a time and answers each of them before receiving the next. Subsequent queries may be chosen adaptively based on previous answers. In this setting the work of [RR10] first showed non-trivial utility guarantees when $|Q| \gg n$.

The term *synthetic data* is often used to refer to a data set rather than a distribution over X . We note that given a distribution over X we can draw n samples from the distribution in order to obtain an actual data set. This introduces sampling error that in our applications is subsumed by the perturbation needed to ensure privacy.

Given a set of counting queries $Q \subseteq X \rightarrow [0, 1]$, our goal is to produce synthetic data D^* such that

1. D^* satisfies (ϵ, δ) -differential privacy,
2. D^* approximates D on the set of queries in the sense that $\sup_{f \in Q} |f(D) - f(D^*)| \leq \alpha$.

The set Q of queries can be used, for example, to enforce a collection of margins on a contingency table, as well as any other set of counting queries. In many situations where one might wish to fit a statistical model to the private database D , it is desirable to produce D^* that maintains accurate counts for a set A of marginal values, as these values determine the maximum likelihood estimator and minimal sufficient statistics for the model determined by A . In our empirical evaluation of our algorithms, we will evaluate the KL divergence between D^* and D (equivalently, the relative entropy, or likelihood ratio test statistic) for a variety of values of the privacy parameter ϵ . A line of work in statistics beginning with Duncan et al. [DFK⁺01] refers to this type of evaluation as the “risk-utility trade-off”.

3 Multiplicative weights update with exponential mechanism

Our algorithm is presented in Figure 1. Its utility guarantees are stated and analyzed in Section 3.1. The privacy analysis follows in Section 3.2. Finally, we state and prove our privacy-utility trade-off theorems in Section 3.3.

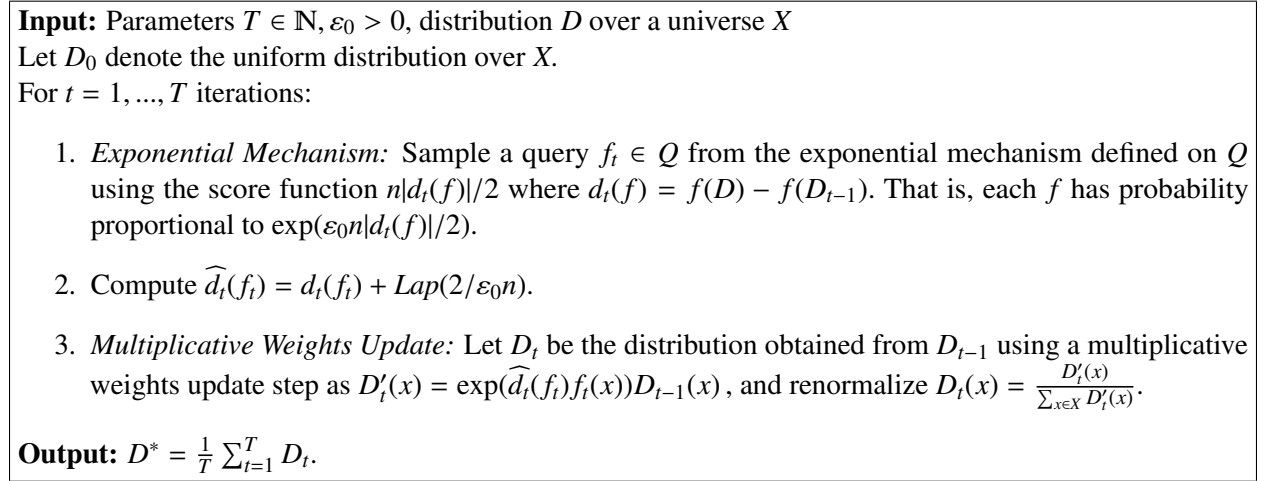


Figure 1: Multiplicative weights update with exponential mechanism (Algorithm 1)

3.1 Utility Analysis

We denote the worst-case (or maximum) error of our output over all queries by

$$\text{maxerr}(D^*, Q) \stackrel{\text{def}}{=} \max_{f \in Q} |f(D^*) - f(D)|. \quad (2)$$

The utility guarantee of our algorithm is captured by the following lemma that gives a bound on the expected value of $\text{maxerr}(D^*, Q)$.

Lemma 3.1 (Expected Maximum Error).

$$\mathbb{E} \maxerr(D^*, Q) \leq O\left(\sqrt{\frac{\log |X|}{T}} + \frac{\log |Q|}{\varepsilon_0 n}\right) \quad (3)$$

Here the expectation is taken over the randomness of Algorithm 1.

We can show a large deviation bound on $\maxerr(D^*, Q)$ as stated in the following lemma.

Lemma 3.2 (Concentration). *Let α denote the right hand side in Equation 3. Then, there is a constant $C > 0$ such that for every $\ell > 0$,*

$$\Pr\left\{\maxerr(D^*, Q) > \alpha + \frac{\ell}{\varepsilon_0 n}\right\} \leq T \exp(-C\ell). \quad (4)$$

In typical settings $\log T = O(\log |Q|)$. In this case by setting $\ell = O(\log |Q|)$, the lemma allows us to bound the failure probability by $1/\text{poly}(|Q|)$ without increasing the error by more than a constant factor. The proof of this lemma is deferred to Appendix A. For simplicity we will not include the failure probability as an explicit parameter in our theorems later.

Proof of Lemma 3.1. Let us use the shorthand $\maxerr_t \stackrel{\text{def}}{=} \maxerr(D_t, Q) = \max_{f \in Q} |d_t(f)|$ to denote the worst-case error at step t of our algorithm. The next lemma says that the exponential mechanism at step t selects a query whose error nearly matches \maxerr_t .

Lemma 3.3 ([MT07]). *For every $t \in \{1, \dots, T\}$,*

$$\Pr\{|d_t(f_t)| \leq \maxerr_t - r\} \leq \frac{|Q| \exp(-\varepsilon_0 r n / 2)}{|\{f \in Q : d_t(f) > \maxerr_t - r\}|}.$$

Lemma 3.4.

$$\mathbb{E} |d_t(f_t)| \geq \maxerr_t - \frac{2 \log |Q| + 1}{\varepsilon_0 n}.$$

Proof. Note that the denominator in the RHS of Lemma 3.3 is always at least 1. Hence, by Lemma 3.3,

$$\Pr\left\{|d_t(f_t)| \leq \maxerr_t - \frac{2 \log |Q| + \ell}{\varepsilon_0 n}\right\} \leq \exp(-\ell). \quad (5)$$

On the other hand $\int_0^\infty \ell \exp(-\ell) d\ell = 1$. □

We will next analyze the convergence of our algorithm to a good distribution using a potential argument, following the argument of [HR10]. Specifically, we show that while our error bounds are not met, each update results in a significant decrease in the relative entropy of D_t and D , which is initially at most $\log |X|$ and always at least 0. This bounds the number of rounds before the error bounds become satisfied.

For two distributions P, Q on a universe X the *relative entropy* is $\text{RE}(P||Q) = \sum_{x \in X} P(x) \log(P(x)/Q(x))$. We consider the potential $\Psi_t = \text{RE}(D||D_t)$. The following two properties follow from non-negativity of entropy, and Jensen's Inequality:

Fact 3.5. $\Psi_t \geq 0$

Fact 3.6. $\Psi_0 \leq \log |X|$

The next lemma gives a lower bound on the expected potential drop.

Lemma 3.7. *In expectation over the Laplacian random variable at step t , we have*

$$\mathbb{E} [\Psi_{t-1} - \Psi_t] \geq \frac{1}{2} |d_t(f_t)|^2 - \frac{1}{4} \mathbb{E} \left| \widehat{d}_t(f_t) \right|^2$$

Proof. A lemma from [HR10] shows that

$$\Psi_{t-1} - \Psi_t \geq \eta d_t(f_t) - \eta^2$$

where η is a scaling parameter that appears in the multiplicative weights update. We chose $\eta = 1/2\widehat{d}_t(f_t)$ so that

$$\Psi_{t-1} - \Psi_t \geq \frac{1}{2} \widehat{d}_t(f_t) \cdot d_t(f_t) - \frac{1}{4} \left| \widehat{d}_t(f_t) \right|^2.$$

Taking expectations over the Laplacian random variable in step t , we get

$$\mathbb{E} \left[\widehat{d}_t(f_t) \cdot d_t(f_t) \right] = d_t(f_t) \mathbb{E} \widehat{d}_t(f_t) = d_t(f_t) \cdot d_t(f_t). \quad \square$$

Lemma 3.8. $\mathbb{E} \left| \widehat{d}_t(f_t) \right|^2 \leq |d_t(f_t)|^2 + \frac{8}{\varepsilon_0^2 n^2}$

Proof. Recall that $\widehat{d}_t(f_t) = d_t(f_t) + \text{Lap}(2/\varepsilon_0 n)$. The claim now follows from the fact that $\mathbb{E} \text{Lap}(\sigma)^2 = 2\sigma^2$ \square

We will now compute the expected potential drop where this time the expectation is taken over the entire randomness of our algorithm. This will allow us to sum the total expected potential drop over all steps of our algorithm.

Combining the previous two lemmas, we get

$$\mathbb{E} [\Psi_{t-1} - \Psi_t] \geq \frac{1}{4} \mathbb{E} |d_t(f_t)|^2 - \frac{2}{\varepsilon_0^2 n^2}. \quad (6)$$

On the other hand, by Jensen's inequality and Lemma 3.4,

$$\mathbb{E} |d_t(f_t)|^2 \geq (\mathbb{E} |d_t(f_t)|)^2 \geq \left(\mathbb{E} \maxerr_t - \frac{2 \log |Q| + 1}{\varepsilon_0 n} \right)^2 \quad (7)$$

Combining (6) with (7), we get

$$\mathbb{E} [\Psi_{t-1} - \Psi_t] \geq \frac{1}{4} \left(\mathbb{E} \maxerr_t - \frac{2 \log |Q| + 1}{\varepsilon_0 n} \right)^2 - \frac{2}{\varepsilon_0^2 n^2}. \quad (8)$$

By linearity of expectation, Fact 3.5, and Fact 3.6, we have

$$\sum_{t=1}^T \mathbb{E} [\Psi_{t-1} - \Psi_t] = \mathbb{E} \left[\sum_{t=1}^T \Psi_{t-1} - \Psi_t \right] = \mathbb{E} [\Psi_0 - \Psi_T] \leq \log |X|.$$

Therefore,

$$\sum_{t=1}^T \left(\mathbb{E} \maxerr_t - \frac{2 \log |Q| + 1}{\varepsilon_0 n} \right)^2 \leq 4 \log |X| + \frac{8}{\varepsilon_0^2 n^2}. \quad (9)$$

On the other hand, by Cauchy-Schwarz ($\sum a_t^2 \geq \frac{1}{T} (\sum a_t)^2$),

$$\sum_{t=1}^T \left(\mathbb{E} \maxerr_t - \frac{2 \log |Q| + 1}{\varepsilon_0 n} \right)^2 \geq \frac{1}{T} \left(\sum_{t=1}^T \left[\mathbb{E} \maxerr_t - \frac{2 \log |Q| + 1}{\varepsilon_0 n} \right] \right)^2. \quad (10)$$

Combining (9) with (10) and rearranging, we get

$$\sum_{t=1}^T \mathbb{E} \maxerr_t \leq \sqrt{4T \log |X| + \frac{8T}{\varepsilon_0^2 n^2}} + \frac{T(2 \log |Q| + 1)}{\varepsilon_0 n}.$$

Lemma 3.1 now follows by observing that

$$\mathbb{E} \maxerr(D^*, Q) \leq \sum_{t=1}^T \frac{\mathbb{E} \maxerr_t}{T} \leq \sqrt{\frac{4 \log |X|}{T} + \frac{8}{T \varepsilon_0^2 n^2}} + \frac{2 \log |Q| + 1}{\varepsilon_0 n}. \quad (11)$$

We can further simplify this bound by noting that $T \geq 1$ and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. Hence, the additional term of $8/T\varepsilon_0^2 n^2$ under the square root is bounded by $O(1/\varepsilon_0 n) \leq O(\log |Q|/\varepsilon_0 n)$. This concludes the proof of Lemma 3.1.

3.2 Privacy Analysis

Our privacy analysis can be derived easily from known composition theorems for differential privacy.

Theorem 3.1 ([DMNS06]). *For every $\varepsilon_0 > 0$, and $T \in \mathbb{N}$, the class of ε_0 -differentially private mechanisms is ε -differentially private under T -fold adaptive composition, for $\varepsilon = T\varepsilon_0$.*

If willing to accept small additive slack in the privacy guarantee, we can get the following guarantee which is better in terms of the privacy guarantee ε by roughly a \sqrt{T} factor provided that $\varepsilon \leq 1/\sqrt{T}$.

Theorem 3.2 ([DRV10]). *For every $\varepsilon_0 > 0, \delta > 0$ and $T \in \mathbb{N}$, the class of ε_0 -differentially private mechanisms is (ε, δ) -differentially private under T -fold adaptive composition, for $\varepsilon = \sqrt{2T \ln(1/\delta)}\varepsilon_0 + T\varepsilon_0 \cdot (e^{\varepsilon_0} - 1)$.*

Lemma 3.9. *A single time step iteration of Algorithm 1 satisfies ε_0 -differential privacy.*

Proof. The exponential mechanism as defined satisfies $\varepsilon_0/2$ -differential privacy [MT07]. On the other hand, $d_t(f_t)$ has sensitivity at most $1/n$ (the counting query $f(D)$ minus a public quantity $f(D_t)$). Hence, $\hat{d}_t(f_t)$ satisfies $\varepsilon_0/2$ -differential privacy as well. The claim now follows from Theorem 3.1. \square

Corollary 3.10 (Privacy). *Algorithm 1 satisfies*

1. $(\varepsilon_0 T)$ -differential privacy,
2. $(\varepsilon_0 \sqrt{2T \log(1/\delta)} + T\varepsilon_0(e^{\varepsilon_0} - 1), \delta)$ -differential privacy.

3.3 Minimizing error while maintaining privacy

In this section we give two theorems that are each obtained directly from our previous analysis by minimizing the error of Algorithm 1, while maintaining either ε -differential privacy or (ε, δ) -differential privacy. In each case, the required setting of the parameter T and the privacy parameter is omitted from the theorem and instead stated explicitly only in the proof of each theorem.

Theorem 3.3. *Let $\varepsilon > 0$. Given a data set of size n over a universe X and a set of counting queries Q , Algorithm 1 produces synthetic data D^* satisfying ε -differential privacy and*

$$\mathbb{E} \maxerr(D^*, Q) \leq O\left(\frac{\log |X| \log |Q|}{\varepsilon n}\right)^{1/3}.$$

Proof. To achieve ε -differential privacy over all we will run our algorithm with $\varepsilon_0 = \varepsilon/T$ and invoke Theorem 3.1. Now, let

$$\alpha(T) = \sqrt{\frac{\log |X|}{T}} + \frac{T \log |Q|}{\varepsilon n}.$$

By Lemma 3.1, $\mathbb{E} \max_{\text{err}}(D^*, Q) \leq O(\alpha(T))$. We would therefore like to minimize $\alpha(T)$ as a function of T . Writing $\alpha(T) = aT^{-1/2} + bT$ and taking the derivative, we see that the minimum obtains at $T^* = (a/2b)^{2/3}$, giving $\alpha^* = \alpha(T^*) = 2a^{2/3}b^{1/3}$, which corresponds to

$$T^* = \left(\frac{\varepsilon n \sqrt{\log |X|}}{2 \log |Q|} \right)^{2/3}, \quad \alpha^* = 2 \left(\frac{\log |X| \log |Q|}{\varepsilon n} \right)^{1/3}.$$

This concludes the theorem. \square

We can obtain a stronger error bound by passing from ε -differential privacy to (ε, δ) -differential privacy.

Theorem 3.4. *Let $\varepsilon > 0, \delta > 0$. Given a data set of size n over a universe X and a set of counting queries Q , Algorithm 1 produces synthetic data D^* satisfying (ε, δ) -differential privacy and*

$$\mathbb{E} \max_{\text{err}}(D^*, Q) \leq O \left(\frac{\sqrt{\log |X| \log(1/\delta)} \cdot \log |Q|}{\varepsilon n} \right)^{1/2}.$$

Proof. To achieve (ε, δ) -differential privacy overall, by Corollary 3.10, we will run Algorithm 1 with privacy parameter ε_0 satisfying $\varepsilon = \varepsilon_0 \sqrt{2T \log 1/\delta} + T\varepsilon_0(e^{\varepsilon_0} - 1)$.

Choosing $\varepsilon_0 = \varepsilon/C \sqrt{T \log(1/\delta)}$ for some constant $C > 0$ is sufficient. Let

$$\alpha(T) = \sqrt{\frac{\log |X|}{T}} + \frac{\sqrt{T \log(1/\delta)} \log |Q|}{\varepsilon n}.$$

By Lemma 3.1, $\mathbb{E} \max_{\text{err}}(D^*, Q) \leq O(\alpha(T))$. Again, we would like to minimize $\alpha(T)$ as a function of T . This is achieved for

$$T^* = \Theta \left(\frac{\sqrt{\log |X|} \varepsilon n}{\sqrt{\log(1/\delta)} \log |Q|} \right),$$

giving $\alpha(T^*)$ that matches the bound stated in the theorem. \square

4 Implementation and Experimentation

In this section we consider an application of our general framework to the problem of contingency table release. We choose this particular problem because it exhibits interesting correlations between queries, as well as having a significant role in the practice of official statistics.

A contingency table reflects a set of k discrete attributes, where each record in the table has a setting of each attribute. A contingency table is commonly represented by enumerating the list of all possible settings of the attributes and reporting the counts of the number of records with the associated setting. We can also do the same for a subset of the attributes, reporting the counts for each possible setting of the attributes in the subset, which is referred to as a *marginal*.

When statistical inference is performed over contingency tables, statisticians seek sets of *low-order* marginals, those containing relatively few attributes at a time, that explain the data well. Our goal, in releasing contingency tables, is to release data so that these low-order marginals are accurately preserved.

In previous work, Barak et al. [BCD⁺07] describe an approach to differentially private contingency table release through the Fourier transformation. If we view a contingency table as vector, coordinates ordered

lexicographically by [binary] attribute settings, the Fourier transformation corresponds to multiplication by the Hadamard matrix, defined recursively as

$$H_{n+1} = \begin{bmatrix} H_n & H_n \\ H_n & -H_n \end{bmatrix}, \quad H_1 = [1].$$

Each result in the multiplication corresponds to the measurement of a Fourier coefficient. Intuitively, there is one such measurement for each subset of attributes, reflecting the counts of records with even and odd parities of attribute values among the subset of attributes.

These coefficients are interesting in that all low order marginals can be exactly recovered by examination of relatively few entries in the transformed vector, the measurements corresponding to subsets of size at most k . Rather than explain the production of marginals from these Fourier measurements (details can be found in Barak et al. [BCD⁺07]) we simply exploit the connection by considering the corresponding count queries, and insisting on producing a distribution that respects them. The marginals can then be derived directly from the distribution.³

4.1 Experimental Set-up

For our experiments, we consider several datasets used in the statistical literature. The datasets range from relatively small (70 records) to substantial (21k records). We have avoided enormous datasets as they seem to occur less frequently in practice (truly large datasets are invariably segmented into subpopulations before analysis), and are not especially good indicators of algorithm performance (even the simplest algorithm works well). The challenge with differential privacy is getting it to work on smaller datasets, rather than larger. The datasets we consider are detailed in Table 1. There are several ways to measure the quality of our approaches.

	records	attributes	non-zero / total cells
mildew	70	6	22 / 64
czech	1841	6	63 / 64
rochdale	665	8	91 / 256
nlts	21574	16	3152 / 65536

Table 1: Structural details of the four datasets we consider.

One that we will focus on is relative entropy, or KL divergence. This measurement has appealing properties for statistical inference, and is used in previous statistical work on the problem. Other measurements, for example directly measuring the error in marginal tables, certainly exist, but our goal is ultimately learning models that fit the data well, and it is not immediately clear how accuracy in these measurements result in statistical quality of fit.

Our experiments are intended both to compare our approach to the prior work of Barak et al. [BCD⁺07] as well as to evaluate it in absolute terms. For the purposes of our experiments, Barak et al. will simply be represented by the approach that takes all low order Fourier measurements with a uniform level of accuracy; their approach involved an additional linear programming step, which we found hurts its performance with respect to relative entropy. For the absolute comparison, we invoke the work of Fienberg et al. [FRY10] on several of these datasets where they report absolute numbers for quality of fit (in terms of relative entropy) without privacy constraints.

³There is nothing wrong with explicitly using the Fourier transform to return to the marginals, but it is exciting to note that we do not need to specify the relationship between the measurements we take and the quantities of interest; we only need the relationships to exist. This is helpful when the dependence is complicated and/or inexact.

All of our experiments are done with ϵ -differential privacy, that is, $\delta = 0$. The absolute numbers improve in the privacy-utility trade-off if we permit a non-zero δ . However, the relationships between the curves can change; the improvement in ϵ one would see with a non-zero δ depends on the technique and the dataset in a way we have not measured.

4.2 Improvements

We now consider several variations on the simple approach presented in Section 3 that can lead to noticeably improved performance. Although the worst case bounds do not improve, there is theoretical motivation for each of the improvements, which we also detail.

Iterating the Update On each iteration we select a query to measure based on the amount of error exhibited between our approximating distribution and the true data. The selected query is measured, and corrected. However, over the course of the algorithm, measurements may drift again. There is no privacy cost to re-processing a previous measurement, so we can take advantage of this to further decrease our potential function without re-interrogating the data.

Initialization Our potential function starts at the logarithm of the universe size, because our best guess at the outset is of a uniform distribution. This can sometimes be improved by taking a histogram of the values; by simply counting (with noise) the number of occurrences of each type of record, we can identify values that occur with substantial frequency and update the prior accordingly. This works well if there are several values with high frequency (as they contribute most to the potential function) but it does consume from the privacy budget, and reduces the accuracy allowed in the query measurement stage.

Adapting the Number of Rounds The number of rounds to conduct is an important parameter. Setting it too low results in not enough information extracted about the data, but setting it too high causes each round to give very noisy measurements, of little value. Instead, we can set the number adaptively, by starting with a very small epsilon value and asking queries until the observed signal drops below noise levels. At this point, if privacy budget still remains, we double epsilon and restart. As epsilon increases we will only drill deeper, each round asking at least as many questions as the last at twice the privacy cost, causing the cumulative cost to telescope and be within a factor of two of the final cost.

4.3 Small Datasets

We first evaluated our approach on several small datasets in common use by statisticians. Our findings here were fairly uniform across the datasets: the ability to measure only those queries that are informative about the dataset results in substantial savings over taking all possible measurements. We evaluated both our theoretically pure algorithm and its heuristic improvement as discussed in the previous section, against a modified version of the algorithm of Barak et al. [BCD⁺07] (integrating the multiplicative weights of Hardt-Rothblum [HR10]), and the accepted "good" non-private relative entropy values from Fienberg et al. [FRY10]. The trade-off between relative entropy and ϵ for three datasets appears in Figure 2. In each case, we see that we noticeably improve on the algorithm of Barak et al., and in many cases our heuristic approach matches the good non-private values.

4.4 Large Dataset

We also consider a larger dataset, the National Long-Term Care Study (NLTC) in Figure 3. This dataset contains orders of magnitudes more records, and has 16 binary attributes. For our initial settings, maintaining

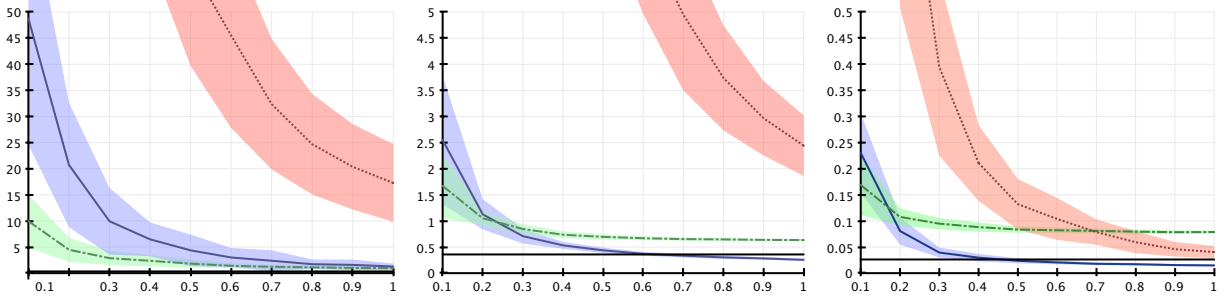


Figure 2: Curves describing the behavior of algorithms on the mildew, rochdale, and czech datasets, respectively. The x-axis is the value of epsilon guaranteed, and the y-axis is the relative entropy between the produced distribution and actual dataset. The lines represent averages across 100 runs, and the corresponding shaded areas one standard deviation in each direction. Red (dashed) represents the modified Barak et al. algorithm, green (dot-dashed) represents the unoptimized use of the exponential mechanism to select queries for the multiplicative weights algorithm, and blue (solid) represents the optimized version thereof. The solid black horizontal line is the stated relative entropy values from Fienberg et al.

all three-way Fourier measurements, we see similar behavior as above: the ability to choose the measurements that are important allows substantially higher accuracy on those that matter.

However, we see that the algorithm of Barak et al. [BCD⁺07] is substantially more competitive in the regime where we are interested in querying all two-way marginals, rather than the default three we have been using. In this case, for values of epsilon at least 0.1, it seems that there is enough signal present to simply measure all such Fourier coefficients; each is sufficiently informative that measuring substantially fewer at higher accuracy imparts less information, rather than more.

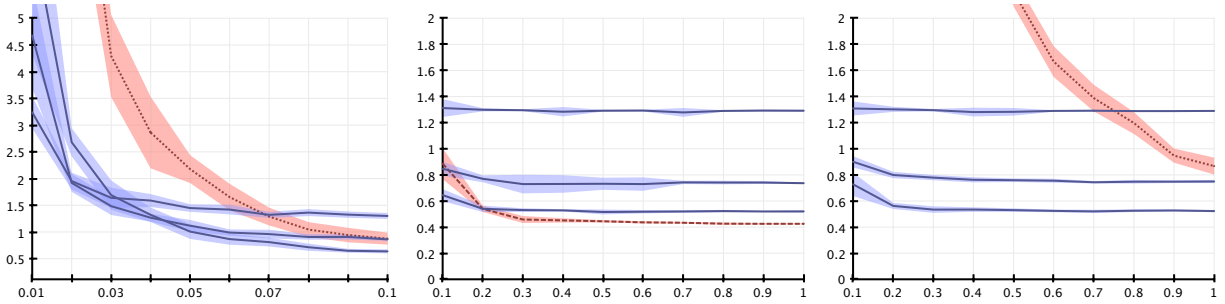


Figure 3: Curves comparing our approach with that of Barak et al on the National Long Term Care Survey. The red (dashed) curve represents Barak et al, and the multiple blue (solid) curves represent multiplicative weights combined with the exponential mechanism, with 20, 30, and 40 queries (top to bottom, respectively). From left to right, the first two figures correspond to degree 2 Fourier coefficients, and the third to degree 3 Fourier coefficients. We see that the exponential mechanism improves accuracy when the privacy requirements are strong relative to the number of measurements (the first and third graphs). The middle graph demonstrates that with few measurements and sufficient privacy budget, one does best by simply measuring everything. As before, the x-axis is the value of epsilon guaranteed, and the y-axis is the relative entropy between the produced distribution and actual dataset. The lines represent averages across only 10 runs, owing to the high complexity of Barak et al on this many-attributed dataset, and the corresponding shaded areas one standard deviation in each direction.

For every dataset and query set, there is some sufficiently high epsilon level where the judicious selection of queries is no longer required. In such regimes, the approach we present in this paper does not provide an improvement over more naive approaches. The impact of our approach returns if we increase the order of

marginal that must be preserved (dramatically increasing the number of measurements Barak et al. would take) or if we decrease epsilon to a level such that the majority of two-way Fourier coefficients are not above the noise level. However, the analyst’s goal should be to get the right output for the analysis task at hand, under the supplied privacy constraints. In some cases this may not require the use of our advanced query selection.

5 Conclusions

We have studied a simple algorithm for releasing data maintaining a high fidelity to the protected source data, as well as differential privacy with respect to the records. The approach builds upon the multiplicative weights approach of [HR10], by introducing the exponential mechanism [MT07] as a more judicious approach to determining which measurements to take. The theoretical analysis improves upon previous work in the area, and experimentally we have evidence that for many interesting parameters it represents a substantial improvement over existing techniques.

As well as improving on theoretical bounds and experimental error, the algorithm is both simple to implement and simple to use. An analyst does not require a complicated mathematical understanding of the nature of the queries (as the community has for Fourier coefficients and marginal tables), but rather only needs to enumerate those measurements that should be preserved. We hope that this generality leads to a broader class of high fidelity differentially-private data releases for a variety of data domains.

References

- [BCD⁺07] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proc. ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 273–282. ACM, 2007.
- [BLR08] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *Proc. ACM Symposium on Theory of Computing (STOC)*, pages 609–618, 2008.
- [DFK⁺01] G.T. Duncan, S.E. Fienberg, R. Krishnan, R. Padman, and S.F. Roehrig. Disclosure limitation methods and information loss for tabular data. In P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, editors, *Theory and Practical Applications for Statistical Agencies*, pages 135–166. 2001.
- [DMNS06] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proc. Theory of Cryptography Conference*, pages 265–284, 2006.
- [DN03] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 202–210. ACM Press New York, NY, USA, 2003.
- [DN04] Cynthia Dwork and Kobbi Nissim. Privacy-preserving datamining on vertically partitioned databases. In *In CRYPTO*, pages 528–544. Springer, 2004.
- [DNR⁺09] C. Dwork, M. Naor, O. Reingold, G.N. Rothblum, and S. Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proc. ACM Symposium on Theory of Computing (STOC)*, pages 381–390, 2009.
- [DRV10] Cynthia Dwork, Guy Rothblum, and Salil Vadhan. Boosting and differential privacy. In *Proc. IEEE Symposium on Foundations of Computer Science (FOCS)*, 2010.

- [FRY10] Stephen E. Fienberg, Alessandro Rinaldo, and Xiolin Yang. Differential privacy and the risk-utility tradeoff for multi-dimensional contingency tables. In *Privacy in Statistical Databases*, 2010.
- [GHRU10] Anupam Gupta, Moritz Hardt, Aaron Roth, and Jon Ullman. Privately releasing submodular functions and the statistical query barrier. *CoRR*, abs/1011.1296, 2010.
- [HR10] Moritz Hardt and Guy Rothblum. A multiplicative weights mechanism for interactive privacy-preserving data analysis. In *Proc. IEEE Symposium on Foundations of Computer Science (FOCS)*, 2010.
- [LHR⁺10] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor. Optimizing linear counting queries under differential privacy. In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 123–134. ACM, 2010.
- [MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proc. IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 94–103, 2007.
- [RR10] Aaron Roth and Tim Roughgarden. The median mechanism: Interactive and efficient privacy with multiple queries. In *Proc. ACM Symposium on Theory of Computing (STOC)*, 2010.

A A large deviation bound

It is not difficult to show that $\text{maxerr}(D^*, Q)$ is unlikely to be significantly larger than its expectation. For convenience we will restate Lemma 3.2 below.

Lemma A.1. *Let $\alpha = \Theta(\sqrt{\log |X|/T} + \log |Q|/\varepsilon_0 n)$. Then, there is a constant $C > 0$ such that for every $\ell > 0$,*

$$\Pr \left\{ \text{maxerr}(D^*, Q) > \alpha + \frac{\ell}{\varepsilon_0 n} \right\} \leq T \exp(-C\ell).$$

Proof (sketch). Let X_t denote the deviation of the exponential mechanism from maxerr_t in round t of our algorithm. That is $X_t = \text{maxerr}_t - d_t(f_t)$. Note that $X_t \geq 0$. As was argued in Equation 5, we have that $\Pr(X_t > (2 \log |Q| + \ell)/\varepsilon_0 n) \leq \exp(-\ell)$. Likewise, let $Y_t \geq 0$ denote the deviation of the Laplace mechanism from its mean at step t . That is $Y_t = |d_t(f_t) - \widehat{d}_t(f_t)|$. By basic properties of the Laplace distribution, we have $\Pr(Y_t > \ell/\varepsilon_0 n) \leq \exp(-\ell/10)$. Finally, let $Z_t = X_t + Y_t$.

With this notation we can follow the proof of Lemma 3.1 step by step, but without taking expectations. We can replace Equation 9 by

$$\sum_{t=1}^T (\text{maxerr}_t - X_t)^2 \leq 4 \log |X| + 4 \sum_{t=1}^T Y_t^2. \quad (12)$$

From this we conclude,

$$\sum_{t=1}^T \text{maxerr}_t \leq \sqrt{4 \log |X| + 4 \sum_{t=1}^T Y_t^2} + \sum_{t=1}^T X_t \leq \sqrt{4 \log |X|} + 2 \sum_{t=1}^T Y_t + \sum_{t=1}^T X_t, \quad (13)$$

using the fact that $\sqrt{\sum_t Y_t^2} \leq \sum_t Y_t$. Therefore,

$$\text{maxerr}(D^*, Q) \leq O \left(\sqrt{\frac{\log |X|}{T}} + \frac{1}{T} \sum_{t=1}^T Z_t \right). \quad (14)$$

Combining this with our previous observations and taking the union bound over all variables Z_t , there is a constant $C' = O(1)$ such that

$$\Pr \left\{ \max_{\text{err}}(D^*, Q) > C' \left(\sqrt{\frac{\log |X|}{T}} + \frac{\log |Q| + \ell}{\varepsilon_0 n} \right) \right\} \leq T \exp(-\ell).$$

Comparing this with our bound on $\mathbb{E} \max_{\text{err}}(D^*, Q)$ (as given in Lemma 3.1), the claim follows. \square