Differential Privacy and the Fat-Shattering Dimension of Linear Queries

Aaron Roth[*]

April 17, 2010

Abstract

In this paper, we consider the task of answering linear queries under the constraint of differential privacy. This is a general and well-studied class of queries that captures other commonly studied classes, including predicate queries and histogram queries. We show that the accuracy to which a set of linear queries can be answered is closely related to its *fat-shattering dimension*, a property that characterizes the learnability of real-valued functions in the agnostic-learning setting.

1 Introduction

The administrator of a database consisting of sensitive, but valuable information faces two conflicting objectives. Because the data is valuable, she would like to make statistical information about it available to the public. However, because the data is sensitive, she must take care not to release information that exposes the data of any particular individual in the data set. The central question in the field of *private data analysis* is how these two objectives can be traded off, and more specifically, how many queries of what type can be answered to given degrees of accuracy, while still preserving privacy.

Recent work on *differential privacy* provides a mathematical framework to reason about such questions. Informally, a probabilistic function f from a database D to some range $\mathcal{R}$ is α-differentially private if adding or removing a single individual from the dataset does not change the probability that $f(D) = r$ for any outcome $r \in \mathcal{R}$ by more than an e^α factor. The intuition behind this definition is that an individual's privacy should not be considered to have been violated by some event r, if r would have been almost as likely to occur even without the individual's data.

In this paper, we consider databases D which are real valued vectors, and the class of queries that we consider correspond to linear combinations of the entries of D. Formally, we consider databases $D \in \mathbb{R}^n_+$, and queries of the form $q \in [0,1]^n$. The answer to query q on database D is simply the dot-product of the two vectors: $q(D) = q \cdot D$. This model has previously been considered ([DN03, DMT07, DY08, HT10]), and generalizes the class of *count queries* or *predicate queries*, which has also been well studied ([DMNS06, BLR08, DNR$^+$09, RR10, UV10]).

The *fat-shattering dimension* (FSD) of a class of real-valued functions C over some domain is a generalization of the Vapnik-Chervonenkis dimension, and characterizes a distribution-free convergence property of the mean value of each $f \in C$ to its expectation. The fat-shattering dimension of a class of functions C is known to characterize the sample complexity necessary to PAC learn C in the agnostic framework [ABDCBH97, BLW94]: that is, ignoring computation, the sample complexity that is both necessary and sufficient to learn C in the agnostic framework is polynomially related to the fat-shattering dimension of C.

Our main result is a similar information theoretic characterization of the magnitude of the noise that must be added to the answer to each query in some class C in terms of the fat-shattering dimension of C, $\text{FSD}(C)$. We show polynomially related information theoretic upper and lower bounds on the noise that must be added to each query in C in terms of $\text{FSD}(C)$. This generalizes the results of [BLR08] to linear queries, and to our knowledge gives the first analysis of generic linear queries using some parameter other than their cardinality. This yields the first mechanism capable of answering a possibly infinite set of generic linear queries, and the first non-trivial lower bound for infinite classes of non-boolean linear queries. As a consequence, we extend results of Kasiviswanathan et al. and Blum et al. [**?**, BLR08] relating the sample complexity necessary for agnostic PAC learning and private agnostic PAC learning from classes of boolean valued functions to classes of real valued functions.

1.1 Related Work and Our Results

Dinur and Nissim studied the special case of linear queries for which both the database and the query are elements of the boolean hypercube $\{0,1\}^n$ [DN03]. Even in this special case, they showed that there cannot be any private mechanism that answers n queries with error $o(\sqrt{n})$, because an adversary could use any such mechanism to reconstruct a $1 - o(1)$ fraction of the original database, a condition which they called *blatant non-privacy*. This result was strengthened by several subsequent papers [DMT07, DY08].

Dwork et al. gave the original definition of differential privacy, as well as the Laplace mechanism, which is capable of answering any k "low sensitivity" queries (including linear queries) up to error $O(k)$. A more refined analysis of the relationship between the laplace mechanism and function sensitivity was later given by [NRS07].

In a different setting, Blum Ligett and Roth considered the question of answering *predicate queries* over a database drawn from some domain X [BLR08]. This can be viewed as a special case of linear queries in which the queries are restricted to lie on the boolean hypercube, and the database must be integer valued: $D \in \mathbb{Z}_+^n$. They give a mechanism for answering every query in some class C with noise that depends linearly on the VC-dimension of the class of queries. This is a quantity that is at most $\log|C|$ for finite classes C, and can be finite even for *infinite* classes. Roth and Roughgarden later gave a mechanism which achieved similar bounds in the online model, in which the mechanism does not know the set of queries that must be answered ahead of time, and instead must answer them as they arrive [RR10]. We generalize the technique of [BLR08, RR10] to apply to general linear queries. VC-dimension is no longer an appropriate measure of query complexity in this setting, but we show that a quantity known as Fat-Shattering dimension plays an analogous role.

Hardt and Talwar give matching upper and lower bounds on the noise that must be added when answering $k \le n$ linear queries of roughly $\Theta(\frac{\sqrt{k}\log(n/k)}{\alpha})$, and [KRSU10] give similar lower bounds. In contrast, we prove bounds in terms of different parameters, and can handle arbitrarily (even infinitely) large values of k. For finite sets of k queries, our mechanism adds noise roughly $O\left(||D||_1^{2/3} \cdot \left(\frac{\log k \log n}{\alpha}\right)^{1/3}\right)$. Note that this is significantly less noise than even the lower bound of [HT10] for $k \ge \Omega(||D||_1^{4/3})$, and to achieve low *relative* error η (i.e. error $\epsilon = \eta||D||_1$), our mechanism requires only that $||D||_1$ be poly-logarithmic in k, rather than polynomial in k. For infinite classes of queries $|C|$, the $\log k$ in our bound can be replaced with the fat shattering dimension of the class C. We also show a lower bound in terms of the fat shattering dimension of the class C, which is the first non-trivial lower bound for infinite classes of non-boolean linear queries.

2   Preliminaries

A database is some vector $D \in \mathbb{R}_+^n$, and a query is some vector $q \in [0,1]^n$. We write that the evaluation of q on D is $q(D) = q \cdot D$. We write $||D||_1 = \sum_{i=1}^n D_i$ to denote the ℓ_1 norm of D, and note that for any query q, $q(D) \in [0, ||D||_1]$. We let C denote a (possibly infinite) class of queries. We are interested in mechanisms that are able to provide answers a_i for each $q_i \in C$ so that the maximum error, defined to be $\max_{i \in C} |q_i(D) - a_i|$ is as small as possible. Without loss of generality, we restrict our attention to mechanisms which actually output some synthetic database: mechanisms with range $\mathcal{R} = \mathbb{R}_+^n$. That is, if our mechanism outputs some synthetic database D', we take a_i to be $q_i(D')$ for each i.[1]

We formalize our notion of utility and relative utility for a randomized mechanism M:

Definition 2.1 (Usefulness and Relative Usefulness). A mechanism $M : \mathbb{R}_+^n \to \mathbb{R}_+^n$ is (ϵ, δ)-useful with respect to a class of queries C if with probability at least $1 - \delta$ (over the internal coins of the mechanism), it outputs a synthetic database D' such that:

$$\sup_{q_i \in C} |q_i(D) - q_i(D')| \leq \epsilon$$

For $0 < \eta \leq 1$, M is (η, δ)-relatively useful with respect to C if there exists a finite m such that it is $(\eta ||D||_1, \delta)$-useful with respect to C for all input databases D with $||D||_1 \geq m$.

That is, useful mechanisms should have low error for each query in C. We now define differential privacy:

Definition 2.2 (Differential Privacy [DMNS06]). A mechanism $M : \mathbb{R}_+^n \to \mathbb{R}_+^n$ is α-differentially private, if for any two databases D_1, D_2 such that $||D_1 - D_2||_1 \leq 1$, and for any $S \subseteq \mathbb{R}_+^n$:

$$\Pr[M(D_1) \in S] \leq e^\alpha \Pr[M(D_2) \in S]$$

The standard notion of differential privacy need only hold for mechanisms defined over integer valued databases $D_1, D_2 \in \mathbb{N}^n$, which is a weaker condition. Our upper bounds will hold for the stronger notion of differential privacy, and our lower bounds for the weaker notion. A useful observation is that arbitrary (database independent) functions of differentially private mechanisms are also differentially private:

Fact 1. If $M : \mathbb{R}_+^n \to \mathbb{R}_+^n$ is α-differentially private, and if $f : \mathbb{R}_+^n \to \mathbb{R}_+^n$ is a (possibly randomized) function, then $f(M)$ is α-differentially private.

2.1   Fat Shattering Dimension

Fat-shattering-dimension is a combinatorial property describing classes of functions of the form $f : X \to [0,1]$ for some domain X. It is a generalization of the Vapnik-Chervonenkis-dimension, which is a property only of classes of boolean valued functions of the form $f : X \to \{0,1\}$. In this section, we generalize these concepts slightly to classes of linear queries, where we view our linear queries as linear combinations of functions $f : X \to [0,1]$, where we let X be the set of standard basis vectors of $\mathbb{R}^n$.

Let $B = \{e_i\}_{i=1}^n$ denote the set of n standard basis vectors of $\mathbb{R}^n$ (e_i is the vector with a 1 in the i'th coordinate, and a 0 in all other coordinates). For any $S \subseteq B$ of size $|S| = d$, we say that S is γ-shattered by C if there exists a vector $r \in [0,1]^d$ such that for every $b \in \{0,1\}^d$, there exists a query $q_b \in C$ such that for each $e_i \in S$:

$$q_b(e_i) \begin{cases} \geq r_i + \gamma, & \text{if } b_i = 1; \\ \leq r_i - \gamma, & \text{if } b_i = 0. \end{cases}$$

Note that since the range of each query is $[0,1]$, γ can range from 0 to $1/2$.

---

[1]This is without loss of generality, because given a different representation for each answer a_i to error ϵ, it is possible to compute a synthetic database D' with error at most 2ϵ using the linear program of [DNR+09].

Definition 2.3 (Fat Shattering Dimension [BLW94, KS94]). The γ-fat-shattering dimension of a class of linear queries C is:

$$\text{FSD}_\gamma(C) = \max\{d \in \mathbb{N} : C \ \gamma - \text{shatters some } S \subseteq B \text{ with } |S| = d\}$$

In the special case when $\gamma = r_i = 1/2$ for all i, note that the fat shattering dimension of a class of boolean valued functions is equal to its VC-dimension.

For finite classes C, we will let $k = |C|$ denote the cardinality of C. The following observation follows immediately from the definition of fat-shattering dimension:

Observation 2.4. *For finite classes C, $FSD_\gamma(C) \leq \log k$ for all $\gamma > 0$, where $k = |C|$.*

3   Lower Bound

In this section, we show that any α-differentially private mechanism that answers every linear query in some class C must add noise at least linear in the fat-shattering dimension of C at any scale.

Theorem 3.1. *For any δ bounded away from 1 by a constant, let M be a mechanism M that is (ϵ, δ) useful with with respect to some class of linear queries C. If M preserves α-differential privacy, then*

$$\epsilon \geq \Omega\left(\sup_{0 < \gamma \leq 1/2} \frac{\gamma^2 \cdot FSD_\gamma(C)}{e^\alpha}\right)$$

Given some class of linear queries C and any $\gamma > 0$, let $S \subseteq B$ be a collection of basis-vectors of size $\text{FSD}_\gamma(C)$ that are γ-shattered by C, and let $r \in [0,1]^{\text{FSD}_\gamma(C)}$ be the corresponding vector as in the definition of fat-shattering dimension. We now partition S into $1/\gamma$ pieces. For each $j \in \{1, \ldots, 1/\gamma\}$, let:

$$S^j = \{e_i \in S : (j-1) \cdot \gamma < r_i \leq j \cdot \gamma\}$$

Since the sets $\{S_j\}$ partition S, By the pigeon-hole principle, there exists some j^* such that $|S^{j^*}| \geq \gamma \cdot |S| = \gamma \cdot \text{FSD}_\gamma(C)$. Let $d = |S^{j^*}|$.

We consider subsets $T \subset S^{j^*}$ of size $|T| = d/2$. For each such subset, we consider the database $D_T = \sum_{e_i \in T} e_i$. Let $b^T \in \{0,1\}^d$ be the vector guaranteed by the definition of fat shattering dimension such that:

$$b_i^T = \begin{cases} 1, & e_i \in T; \\ 0, & \text{otherwise.} \end{cases}$$

Let $q_T \in C$ be the query that corresponds to b^T as in the definition of fat shattering dimension, and let $C_{S^{j^*}} = \{q_T : T \subseteq S^{j^*}, |T| = d/2\}$.

We first show that each function q_T takes its highest value on D_T and cannot take large values on databases $D_{T'}$ for sets T' that differ significantly from T.

Lemma 3.2. *For all $q_T \in C_{S^{j^*}}$ and for all $T' \subseteq S^{j^*}$ with $|T'| = d/2$:*

$$q_T(D_T) - q_T(D_{T'}) \geq \frac{\gamma}{2} \cdot |T \bigtriangleup T'|$$

Proof.

$$\begin{aligned}
q_T(D_T) - q_T(D_{T'}) &= \sum_{e_i \in T} q_T(e_i) - \sum_{e_i \in T'} q_T(e_i) \\
&= \left(\sum_{e_i \in T \cap T'} q_T(e_i) - q_T(e_i) \right) + \sum_{e_i \in T \setminus T'} q_T(e_i) - \sum_{e_i \in T' \setminus T} q_T(e_i) \\
&\geq \left(\sum_{e_i \in T \setminus T'} r_i + \gamma \right) - \left(\sum_{e_i \in T' \setminus T} r_i - \gamma \right) \\
&\geq 2\gamma \cdot |T \setminus T'| - \left(\max_{i \in T' \setminus T} r_i - \min_{i \in T \setminus T'} r_i \right) \cdot |T \setminus T'| \\
&\geq \gamma \cdot |T \setminus T'|
\end{aligned}$$

where the last inequality follows from the fact that $T, T' \subset S^{j^*}$ which was constructed such that:

$$\left(\max_{i \in S^{j^*}} r_i - \min_{i \in S^{j^*}} r_i \right) \leq \gamma$$

holds. Observing that $|T \triangle T'| = 2|T \setminus T'|$ completes the proof. $\square$

With this lemma, we are ready to prove the main technical lemma for our lower bound:

Lemma 3.3. *For any δ bounded away from 1 by a constant, let M be an (ϵ, δ)-useful mechanism with respect to class C. Given as input an unknown private database D_T for some $T \subseteq S^{j^*}$ with $|T| = d/2$, with constant probability $1 - \delta$, there is a procedure to reconstruct a new database D_{T^*} such that $|T \triangle T^*| \leq \frac{4\epsilon}{\gamma}$.*

Proof. Suppose that mechanism M is (ϵ, δ) useful with respect to C for some constant δ bounded away from 1. Then by definition, with constant probability, given input D_T, it outputs some database D' such that for all $q_i \in C$, $|q_i(D_T) - q_i(D')| \leq \epsilon$. For each $T' \subseteq S^{j^*}$ with $|T'| = d/2$ let:

$$v(T') = q_{T'}(D_{T'}) - q_{T'}(D')$$

and let $T^* = \operatorname{argmin}_{T'} v(T')$. Therefore, we have:

$$v(T^*) \leq v(T) = q_T(D_T) - q_T(D') \leq \epsilon \tag{1}$$

where the last inequality follows from the usefulness of the mechanism. We also have:

$$\begin{aligned}
v(T^*) &= q_{T^*}(D_{T^*}) - q_{T^*}(D') \\
&\geq q_{T^*}(D_{T^*}) - q_{T^*}(D_T) - \epsilon \\
&\geq \frac{\gamma}{2} \cdot |T \triangle T^*| - \epsilon
\end{aligned}$$

where the first inequality follows from the usefulness of the mechanism, and the second inequality follows from lemma 3.2. Combining this with equation 1, we get:

$$|T \triangle T^*| \leq \frac{4\epsilon}{\gamma}$$

$\square$

We are now ready to prove the lower bound:

Proof of Theorem. Let $T \subset S^{j^*}$ with $|T| = d/2$ be some arbitrarily selected subset. Let $D_T = \sum_{e_i \in T} e_i$ be the corresponding database. By lemma 3.3, given $M(D_T, \epsilon)$, with probability $1 - \delta$ there is a procedure P to reconstruct a database D_{T^*} such that $|T \triangle T^*| \leq 4\epsilon/\gamma$. Throughout the rest of the argument, we assume that this event occurs. Let $x \in T$ be an element selected from T uniformly at random, and let $y \in S \setminus T$ be an element selected from $S \setminus T$ uniformly at random. Let $T' = T \setminus x \cup \{y\}$. Observe that:

$$\Pr[x \in P(M(D_T, \epsilon))] \geq \frac{d/2 - 2\epsilon/\gamma}{d/2} = 1 - \frac{4\epsilon}{\gamma \cdot d}$$

$$\Pr[x \in P(M(D_{T'}, \epsilon))] \leq \frac{2\epsilon/\gamma}{d/2} = \frac{4\epsilon}{\gamma \cdot d}$$

Since $||D_T - D_{T'}||_1 \leq 2$, we have by the definition of α-differential privacy and fact 1:

$$
\begin{aligned}
e^\alpha &\geq \frac{\Pr[x \in P(M(D_T, \epsilon))]}{\Pr[x \in P(M(D_{T'}, \epsilon))]} \\
&\geq \frac{1 - \frac{4\epsilon}{\gamma \cdot d}}{\frac{4\epsilon}{\gamma \cdot d}} \\
&= \frac{\gamma \cdot d}{4\epsilon} - 1
\end{aligned}
$$

Solving for ϵ, we find that:

$$\epsilon \geq \Omega\left(\frac{\gamma \cdot d}{e^\alpha}\right)$$

Since this holds for all choices of γ, the claim follows from the fact that $d \geq \gamma \mathrm{FSD}_\gamma(C)$. $\qquad\square$

4 Upper Bound

We now show that (ignoring other important parameters), it is sufficient to add noise linear in the fat shattering dimension of C to simultaneously guarantee usefulness with respect to C and differential privacy.

We make use of a theorem of Bartlett and Long [BL95] (improving a bound of Alon et al. [ABDCBH97]) concerning the rate of convergence of uniform Glivenko-Cantelli classes with respect to their fat-shattering dimension.

Theorem 4.1 ([BL95] Theorem 9). *Let C be a class of functions from some domain X into $[0, 1]$. Then for all distributions $\mathbb{P}$ over X and for all $\eta, \delta \geq 0$:*

$$\Pr\left[\sup_{f \in C} \left|\frac{1}{m}\sum_{i=1}^m f(x_i) - \mathrm{E}_{x \sim \mathbb{P}}[f(x)]\right| \geq \eta\right] \leq \delta$$

where $\{x_i\}_{i=1}^m$ are m independent draws from $\mathbb{P}$ and

$$m = O\left(\frac{1}{\eta^2}\left(d_{\eta/5} \ln^2 \frac{1}{\eta} + \ln \frac{1}{\delta}\right)\right)$$

where $d_{\eta/5} = FSD_{\eta/5}(C)$.

We use this theorem to prove the following useful corollary:

Corollary 4.2. *Let C be a class of linear functions from $\mathbb{R}_+^n$ to $\mathbb{R}$. For any database $D \in \mathbb{R}_+^n$, there is a database $D' \in \mathbb{N}^n$ with*

$$||D'||_1 = O\left(\frac{d_{\eta/5}}{\eta^2} \cdot \log^2\left(\frac{1}{\eta}\right)\right)$$

such that for each $q \in C$,

$$\left| q(D) - \frac{||D||_1}{||D'||_1} q(D') \right| \leq \eta ||D||_1$$

where $d_{\eta/5} = FSD_{\eta/5}(C)$.

Proof. Let $B = \{e_i\}_{i=1}^n$ denote the set of n standard basis vectors over $\mathbb{R}^n$. Let $\mathbb{P}_D$ be the probability distribution over B that places probability $D_i/||D||_1$ on e_i. Note that for any $q \in C$:

$$\mathrm{E}_{e_i \sim \mathbb{P}_D}[q(e_i)] = \sum_{i=1}^n \frac{D_i}{||D||_1} q(e_i) = \frac{1}{||D||_1} \sum_{i=1}^n q(D_i e_i) = \frac{q(D)}{||D||_1}$$

Let $x_1, \ldots, x_m$ be $m = O\left(\frac{1}{\eta^2} \left(d_{\eta/5} \ln^2 \frac{1}{\eta} + \ln 2 \right) \right)$ independent draws from $\mathbb{P}_D$, and let $D' = \sum_{i=1}^m x_i$. Then:

$$q(D') = \sum_{i=1}^n q(D_i' e_i) = \sum_{i=1}^m q(x_i)$$

By lemma 4.1, we have:

$$\Pr\left[\left| \frac{q(D')}{m} - \frac{q(D)}{||D||_1} \right| \geq \eta \right] = \Pr\left[\left| \frac{1}{m} \sum_{i=1}^m q(x_i) - \mathrm{E}_{e_i \sim \mathbb{P}_D}[q(e_i)] \right| \geq \eta \right]$$

$$\leq \frac{1}{2}$$

In particular, there exists some $D' \in \mathbb{N}^n$ with $||D'||_1 = m$ that satisfies $\left| \frac{q(D')}{||D'||_1} - \frac{q(D)}{||D||_1} \right| \leq \eta$. Multiplying through by $||D||_1$ gives the desired bound. $\square$

Armed with Corollary 4.2, we may no proceed to instantiate the exponential mechanism over a sparse domain, analogously to the instantiation of the exponential mechanism in [BLR08].

Definition 4.3 (The Exponential Mechanism [MT07]). Let $\mathcal{D}$ be some domain, and let $s : \mathbb{R}_+^n \times \mathcal{D} \to \mathbb{R}$ be some *quality score* mapping database/domain-element pairs to some real value. Let

$$\Delta_s \geq \max_{r \in \mathcal{D}} \sup_{D_1, D_2 \in \mathbb{R}_n^+ : ||D_1 - D_2||_1 \leq 1} |s(D_1, r) - s(D_2, r)|$$

be an upper bound on the ℓ_1 sensitivity of s. The exponential mechanism defined with respect to domain $\mathcal{D}$ and score s is the probability distribution (parameterized by the private database D) which outputs each $r \in \mathcal{D}$ with probability proportional to:

$$r \sim \exp\left(\frac{s(D, r) \cdot \alpha}{2\Delta_s} \right)$$

Theorem 4.4 (McSherry and Talwar [MT07]). *The exponential mechanism preserves α-differential privacy.*

We let $m = O\left(\frac{d_{\eta/5}}{\eta^2} \cdot \log^2 \left(\frac{1}{\eta} \right) \right)\}$, and define the domain of our instantiation of the exponential mechanism to be:

$$\mathcal{D} = \{ D' \in \mathbb{N}^n : ||D'||_1 = m \}$$

We note that $|\mathcal{D}| = n^m$. Finally, we sample each $D' \in \mathcal{D}$ with probability proportional to:

$$D' \sim \exp\left(-\frac{\sup_{q \in C} \left| q(D) - \frac{||D||_1}{||D'||_1} \cdot q(D') \right| \alpha}{4} \right) \tag{2}$$

and output the database $D_{\text{out}} \equiv \frac{||D||_1}{||D'||_1} \cdot D'^2$. Observe that for any two databases D_1, D_2 such that $||D_1 - D_2||_1 \leq 1$ we have:

$$\sup_{q \in C} |q(D_1) - \frac{||D_1||_1}{||D'||_1} \cdot q(D')| - \sup_{q \in C} |q(D_2) - \frac{||D_2||_1}{||D'||_1} \cdot q(D')| \quad \leq \quad ||D_1 - D_2||_1 + \frac{|||D_1||_1 - ||D_2||_1|}{m}$$

$$\leq \quad 1 + \frac{1}{m}$$

Therefore, the distribution defined in equation 2 is a valid instantiation of the exponential mechanism, and by [MT07] preserves α-differential privacy. It remains to show that the above instantiation of the exponential mechanism yields a useful mechanism with low error:

Theorem 4.5. *For any constant δ and any query class C, there is a value of η such that the above instantiation of the exponential mechanism is (ϵ, δ)-useful with respect to C and α-differentially private for:*

$$\epsilon \leq \tilde{O}\left(||D||_1^{2/3} \cdot \left(\frac{FSD_{2\epsilon/(5||D||_1)}(C) \log n}{\alpha} \right)^{1/3} \right)$$

Such a bound gives us *relatively useful* mechanisms with respect to classes C for databases that have size linear in the fat shattering dimension of C, or only logarithmic in $|C|$ for finite classes C. This is in contrast to the bounds of [HT10] that require databases to be of size polynomial in $|C|$ before giving relatively-useful mechanisms.

Corollary 4.6. *For any constant δ and any query class C, there is an (η, δ)-relatively useful mechanism that preserves α-differential privacy for any database of size at least:*

$$||D||_1 \geq \tilde{\Omega}\left(\frac{FSD_{2\eta/5}(C) \log n}{\alpha \eta^3} \right)$$

Proof of Theorem. Recall that the domain $\mathcal{D}$ of our instantiation of the exponential mechanism consists of all databases $D' \in \mathbb{N}^n$ with $||D'||_1 = m$ with $m = O\left(\frac{d_{\eta/5}}{\eta^2} \cdot \log^2\left(\frac{1}{\eta}\right) \right)$} In particular, by corollary 4.2, there exists a $D^* \in \mathcal{D}$ such that:

$$\left| q(D) - \frac{||D^*||_1}{||D'||_1} q(D^*) \right| \leq \eta ||D||_1$$

Let $\eta = 2\epsilon/||D||_1$. By the definition of our mechanism, such a D^* is output with probability proportional to at least:

$$D^* \sim \exp\left(-\frac{\epsilon \alpha}{8}\right)$$

Similarly, any $D^B \in \mathcal{D}$ such that $\left| q(D) - \frac{||D^B||_1}{||D'||_1} q(D^*) \right| \geq 2\eta ||D||_1$ is output with probability proportional to at most:

$$D^B \sim \exp\left(-\frac{\epsilon \alpha}{4}\right)$$

Let $\mathcal{D}_B$ denote the set of all such D^B. Because $|\mathcal{D}| = n^m$, we have that:

$$\frac{\Pr[D' = D^*]}{D' \in \Pr[\mathcal{D}^B]} \geq \frac{\exp(-\frac{\epsilon \alpha}{8})}{n^m \cdot \exp(-\frac{\epsilon/\alpha}{4})} = n^{-m} \cdot \exp\left(-\frac{\epsilon \alpha}{4}\right)$$

Setting $n^{-m} \cdot \exp\left(-\frac{\epsilon \alpha}{4}\right) \geq 1 - \delta$ and solving, we find that $\Pr[D' = D^*] \geq 1/\delta$ for:

$$\epsilon \leq O\left(\frac{m \log n}{\alpha} \right) = \tilde{O}\left(\sqrt[1/3]{\frac{FSD_{2\epsilon/(5||D||_1)}(C) ||D||_1^2 \cdot \log n}{\alpha}} \right) =$$

$$\tilde{O}\left(||D||_1^{2/3} \cdot \left(\frac{\text{FSD}_{2\epsilon/(5||D||_1)}(C) \log n}{\alpha}\right)^{1/3}\right)$$

$\square$

We remark that the above mechanism is the analogue of the general release mechanism of [BLR08], and answers linear queries in the *offline* setting, when all queries C are known to the mechanism in advance. This is not necessary, however. In the same way as above, corollary 4.2 can also be used to generalize the Median Mechanism of Roth and Roughgarden [RR10], to achieve roughly the same bounds, but in the *online* setting, in which queries arrive online, and the mechanism must privately answer queries as they arrive, without knowledge of future queries. This results in the following theorem:

Theorem 4.7. *For every sequence of adaptively chosen queries $q_1, q_2, \ldots$ arriving online, chosen from some (possibly infinite) set C, there exists a mechanism which is (ϵ, δ) useful with respect to C and preserves (α, τ)-differential privacy*[3], *where τ is a negligible function of n, and:*

$$\epsilon \leq \tilde{O}\left(||D||_1^{2/3} \cdot \left(\frac{FSD_{2\epsilon/(5||D||_1)}^3(C) \log n}{\alpha}\right)^{1/3}\right)$$

Remark 4.8. Notice that for finite classes of linear queries, we may replace the fat shattering dimension in the bounds of both theorems 4.5 and 4.7 with $\log |C|$ if we so choose.

5 Conclusion

In this paper, we have generalized the techniques used by Blum Ligett and Roth, [BLR08] and Roth and Roughgarden [RR10] from the class of predicate queries to the more general class of *linear* queries. This gives the first mechanism for answering every linear query from some class C with noise that is bounded by a parameter other than the cardinality of C; in particular, we have given the first mechanism for answering all of the linear queries in certain *infinite* classes of queries beyond predicate queries. We have shown that the relevant parameter is the Fat-Shattering dimension of the class, which is a generalization of VC-dimension to non-boolean valued queries. In particular (ignoring other parameters), it is necessary and sufficient to add noise proportional to the fat shattering dimension of C. Our results show, among other things, that the sample complexity needed to privately agnostically learn real valued functions is polynomially related to the sample complexity needed to non-privately agnostically learn real valued functions.

At a high level, the same technique can be applied for any class of queries, all of the answers to which can be summarized by some 'small' object. It is then sufficient to instantiate the exponential mechanism only over this much smaller set of objects (rather than the set of all databases) to obtain a useful mechanism. In the case of linear queries, we have shown that the answers to many queries can be summarized by integer valued databases with small ℓ_1 norm. An interesting future direction is to determine what types of nonlinear (but low sensitivity) queries have similar small summarizes from which useful mechanisms can be derived.

6 Acknowledgements

---

[3]A mechanism satisfies (α, τ) differential privacy if it satisfies α-differential privacy, except with probability τ, where the probability is taken only over the internal coins of the mechanism.

References

[ABDCBH97]   N. Alon, S. Ben David, N. Cesa Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

[BL95]   P.L. Bartlett and P.M. Long. More theorems about scale-sensitive dimensions and learning. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 392–401. ACM, 1995.

[BLR08]   A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 609–618. ACM New York, NY, USA, 2008.

[BLW94]   P.L. Bartlett, P.M. Long, and R.C. Williamson. Fat-shattering and the learnability of real-valued functions. In *Proceedings of the seventh annual conference on Computational learning theory*, pages 299–310. ACM, 1994.

[DMNS06]   C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference TCC*, volume 3876 of *Lecture Notes in Computer Science*, page 265. Springer, 2006.

[DMT07]   C. Dwork, F. McSherry, and K. Talwar. The price of privacy and the limits of LP decoding. In *Proceedings of the thirty-ninth annual ACM Symposium on Theory of Computing*, page 94. ACM, 2007.

[DN03]   I. Dinur and K. Nissim. Revealing information while preserving privacy. In *22nd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 202–210, 2003.

[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 *Proceedings of the 41st annual ACM symposium on Symposium on theory of computing*, pages 381–390. ACM New York, NY, USA, 2009.

[DY08]   C. Dwork and S. Yekhanin. New efficient attacks on statistical disclosure control mechanisms. *Advances in Cryptology–CRYPTO 2008*, pages 469–480, 2008.

[HT10]   M. Hardt and K. Talwar. On the Geometry of Differential Privacy. In *The 42nd ACM Symposium on the Theory of Computing, 2010. STOC'10*, 2010.

[KRSU10]   S. Kasiviswanathan, M. Rudelson, A. Smith, and J. Ullman. The Price of Privately Releasing Contingency Tables and the Spectra of Random Matrices with Correlated Rows. In *The 42nd ACM Symposium on the Theory of Computing, 2010. STOC'10*, 2010.

[KS94]   M.J. Kearns and R.E. Schapire. Efficient distribution-free learning of probabilistic concepts*. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.

[MT07]   F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual Symposium on Foundations of Computer Science*, 2007.

[NRS07]   K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of Computing*, 2007.

[RR10]   A. Roth and T. Roughgarden. Interactive Privacy via the Median Mechanism. In *The 42nd ACM Symposium on the Theory of Computing, 2010. STOC'10*, 2010.

[UV10]   J. Ullman and S. Vadhan. PCPs and the Hardness of Generating Synthetic Data . *Manuscript*, 2010.