Electronic Colloquium on Computational Complexity, Revision 2 of Report No. 90 (2011)

Submodular Functions are Noise Stable

Mahdi Cheraghchi
UT Austin
mahdi@cs.utexas.edu

Adam Klivans
UT Austin
klivans@cs.utexas.edu

Pravesh Kothari
UT Austin
kothari@cs.utexas.edu

Homin K. Lee*
UT Austin
homin@cs.utexas.edu

October 15, 2012

Abstract

We show that all non-negative submodular functions have high *noise-stability*. As a consequence, we obtain a polynomial-time learning algorithm for this class with respect to any product distribution on $\{-1,1\}^n$ (for any constant accuracy parameter ϵ). Our algorithm also succeeds in the agnostic setting. Previous work on learning submodular functions required either query access or strong assumptions about the types of submodular functions to be learned (and did not hold in the agnostic setting).

Additionally we give simple algorithms that efficiently release differentially private answers to all Boolean conjunctions and to all halfspaces with constant average error, subsuming and improving the recent work due to Gupta, Hardt, Roth and Ullman (STOC 2011).

1 Introduction

A function $f : 2^{[n]} \to \mathbb{R}$ is *submodular* if

$$\forall S, T \subseteq [n] : f(S \cup T) + f(S \cap T) \leq f(S) + f(T). \tag{1}$$

Submodular functions have been extensively studied in the context of combinatorial optimization [Edm71, NWF78, FNW78, Lov83] where the functions under consideration (such as the cut function of a graph) are submodular. An equivalent formulation of submodularity is that of decreasing marginal returns,

$$\forall S \subseteq T \subseteq [n], i \in [n] \setminus T : f(T \cup \{i\}) - f(T) \leq f(S \cup \{i\}) - f(S), \tag{2}$$

and thus submodular functions are also a topic of study in economics and the algorithmic game theory community [DNS06, MR07]. In most contexts, the submodular functions considered are non-negative [DNS06, FMV07, MR07, Von09, OV11, BH11, GHRU11], and we will be focusing on non-negative submodular functions as well.

*Supported by NSF grant 1019343 subaward CIF-B-108

ISSN 1433-8092

The main contribution of this paper is a proof that non-negative submodular functions are *noise stable.* Informally, a noise stable function f is one whose value on a random input x does not change much if x is subjected to a small, random perturbation. Noise stability is a fundamental topic in the analysis of Boolean functions with applications in hardness of approximation, learning theory, social choice, and pseudorandomness [KKL88, Hås01, BKS99, O'D04, KOS04, MOO10].

In order to define noise stability, we first define a noise operator that acts on $\{-1,1\}^n$.

Definition 1 (Noise operators)**.** For any product distribution $\Pi = \Pi_1 \times \Pi_2 \times \cdots \times \Pi_n$ over $\{-1,1\}^n$, $\rho \in [0,1]$, $x \in \{-1,1\}^n$, let the random variable y drawn from the distribution $N_\rho(x)$ over $\{-1,1\}^n$ have $y_i = x_i$ with probability ρ and be randomly drawn from Π_i with probability $1-\rho$. The *noise operator* T_ρ on $f : \{-1,1\}^n \to \mathbb{R}$ is defined by letting $T_\rho f : \{-1,1\}^n \to \mathbb{R}$ be the function given by $T_\rho f(x) = \mathbb{E}_{y \sim N_\rho(x)} f(y)$.

N.B.: For the uniform distribution $y \sim N_\rho(x)$ has $y_i = x_i$ with probability $1/2 + \rho/2$, and $y_i = -x_i$ with probability $1/2 - \rho/2$.

Now we can precisely define noise stability:

Definition 2 (Noise stability)**.** The *noise stability* of f at noise rate ρ is defined to be

$$\mathbb{S}_\rho(f) = \langle f, T_\rho f \rangle = \mathbb{E}_{x \sim \Pi}[f(x) T_\rho f(x)].$$

The precise statement of our main theorem is as follows (see Section 2 for definitions):

Theorem 3. Let $\Pi = \Pi_1 \times \Pi_2 \times \cdots \times \Pi_n$ be a product distribution over $\{-1,1\}^n$ with minimum and maximum probability $p_{\min} := \min_{i \in [n]}\{\Pr_{x_i \sim \Pi_i}[x_i = 1]\}$ and $p_{\max} := \max_{i \in [n]}\{\Pr_{x_i \sim \Pi_i}[x_i = 1]\}$ and let $f : \{-1,1\}^n \to \mathbb{R}^+$ be a submodular function. Then for all, $\rho \in [0,1]$,

$$\mathbb{S}_\rho(f) \geq (\rho - (1-\rho)(p_{\max} - p_{\min}))\|f\|_2^2.$$

N.B.: For the uniform distribution we get the bound $\mathbb{S}_\rho(f) \geq \rho \|f\|_2^2$.

Given the high noise-stability of submodular functions, we can apply known results from computational learning theory to show that submodular functions are well-approximated by low-degree polynomials and can be learned agnostically. Our main learning result is as follows:

Corollary 4. Let $\mathcal{C}$ be the class of non-negative submodular functions with $\|f\|_2 := \sqrt{\mathbb{E}_x[f(x)^2]} = 1$ and let $\mathcal{D}$ be any distribution on $\{-1,1\}^n \times \mathbb{R}$ such that the marginal distribution over $\{-1,1\}^n$ is a product distribution. Then there is a statistical query algorithm that outputs a hypothesis h with probability $1-\delta$ such that

$$\mathbb{E}_{(x,y) \sim \mathcal{D}}[|h(x) - y|] \leq opt + \epsilon,$$

in time $\text{poly}(n^{O(1/\epsilon^2)}, \log(1/\delta))$,

Here *opt* is the L_1-error of the best fitting concept in the concept class. (See Section 4 for the precise definition.) Note that the above algorithm will succeed given only statistical query access [Kea98] to the underlying function to be learned.

1.1 Related Work

Recently the study of learning submodular functions was initiated in two very different contexts. Gupta et al. [GHRU11] gave an algorithm for learning bounded submodular functions that arose as a technical necessity for differentially privately releasing the class of disjunctions. Their learning algorithm requires value query access to the target function, but their algorithm works even when the value queries are answered with additive error (value queries that are answered with additive error at most τ are said to be τ-tolerant).

Theorem 5 ([GHRU11])**.** Let $\epsilon, \delta > 0$ and let Π be any product distribution over $[n]$. There is a learning algorithm that when given ($\epsilon/4$-tolerant) value query access to any submodular function $f : 2^{[n]} \to [0, 1]$, outputs a hypothesis h in time $n^{O(\log(1/\delta)/\epsilon^2)}$ such that,

$$\Pr_{S \sim \Pi}[|f(S) - h(S)| \leq \epsilon] \geq 1 - \delta.$$

The learning algorithm of Gupta et al. crucially relies on its query access to the submodular function in order to break the function down into Lipschitz continuous parts that are easier to learn. Compare this to Corollary 4, which has similar learning guarantees, but where the learner only has access to statistical queries and can learn in the agnostic model of learning. (See Section 4.)

The other recent work [BH11] on learning submodular functions was motivated by bundle pricing and used passive supervised learning as a model for learning consumer valuations of added options. In particular, they have a polynomial-time algorithm that can learn (using random examples only) *monotone*, non-negative, submodular functions within a *multiplicative* factor of $\sqrt{n}$ over *arbitrary* distributions. As our machinery breaks down over non-product distributions, none of our results hold in this setting. For product distributions, Balcan and Harvey gave the first $\text{poly}(n, 1/\epsilon)$-time algorithm that can learn (using random examples only) *monotone*, non-negative, *Lipschitz* submodular functions with minimum value m within a *multiplicative* factor of $O(\log(1/\epsilon)/m)$.

1.2 Applications to Differential Privacy

We discuss some applications to differential privacy in Section 5. In particular, we obtain a simple proof of Gupta et al. [GHRU11]'s recent result on releasing disjunctions with improved parameters.

2 Preliminaries

Throughout, we will identify sets $S \subseteq [n]$ with their indicator vectors $\mathbf{1}(S) \in \{\text{-}1, 1\}^n$ where $\mathbf{1}(S)_i = 1$ if $i \in S$ and $\mathbf{1}(S)_i = -1$ if $i \notin S$ (as opposed to the usual $(0, 1)$-indicator vectors). For any distribution over Π over $\{\text{-}1, 1\}^n$, we define the inner product on functions $f, g : \{\text{-}1, 1\}^n \to \mathbb{R}$ by $\langle f, g \rangle = \mathbb{E}_{x \sim \Pi}[f(x)g(x)]$ and the L_2-norm of a function of f as $\|f\|_2 = \sqrt{\langle f, f \rangle} = \sqrt{\mathbb{E}_{x \sim \Pi}[f(x)^2]}$.

Our result will depend on the minimum and maximum marginal marginal probabilities of the given product distribution. For this, we define:

Definition 6 (Maximum and Minimum Probabilities)**.** Let $\Pi = \Pi_1 \times \Pi_2 \times \cdots \times \Pi_n$ be a product distribution over $\{\text{-}1, 1\}^n$. We define $p_{\max} = \max_{i \in [n]} \Pr_{x_i \sim \Pi_i}[x_i = 1]$ and $p_{\min} = \min_{i \in [n]} \Pr_{x_i \sim \Pi_i}[x_i = 1]$.

3 Submodular Functions are Noise Stable

We will start by showing that submodular functions are noise stable under the uniform distribution as a warm-up as the notation is less cumbersome in this setting. In Section 3.2 we will prove Theorem 3 in the general setting of arbitrary product distributions.

3.1 Uniform Distribution

For the rest of Section 3.1 we will assume that the distribution over inputs is uniform.

Let the Fourier expansion of f be given by $\sum_{S\subseteq[n]} \hat{f}(S)\chi_S$, it can be shown that

$$T_\rho f(x) = \sum_{S\subseteq[n]} \rho^{|S|}\hat{f}(S)\chi_S(x), \qquad \text{and thus} \qquad \mathbb{S}_\rho(f) = \sum_{S\subseteq[n]} \rho^{|S|}\hat{f}(S)^2.$$

The following lemma is our key observation.

Lemma 7. Let $f : \{\text{-}1,1\}^n \to \mathbb{R}$ be a submodular function. Then for all $x \in \{\text{-}1,1\}^n$, $\rho \in [0,1]$,

$$T_\rho f(x) \geq \rho f(x) + ((1-\rho)/2)(f(-1^n) + f(1^n)).$$

For $f : \{\text{-}1,1\}^n \to \mathbb{R}^+$, $T_\rho f(x) \geq \rho f(x)$.

Proof. We will be viewing the domain of f as $2^{[n]}$, and the input $x \in \{\text{-}1,1\}^n$ as $X \in 2^{[n]}$ such that $\mathbf{1}(X) = x$. For a fixed $x \in \{\text{-}1,1\}^n$, let $\pi : [n] \to [n]$ be a permutation such that $x_{\pi(1)} \geq \cdots \geq x_{\pi(n)}$, and then define $X_j = \{\pi(1), \ldots, \pi(j)\}$. (N.B.: $X_0 = \emptyset$ and $X_n = [n]$.) Finally, we define $x_{\pi(0)} = 1$ and $x_{\pi(n+1)} = -1$. Note that there is only one value $j \in \{0, \ldots, n\}$ for which $x_{\pi(j)} \neq x_{\pi(j+1)}$.

$$\begin{aligned}
\mathbb{E}_{Y\sim N_\rho(X)} f(Y) &= f(X_0) + \mathbb{E}_{Y\sim N_\rho(X)} \sum_{j=1}^{n} f(Y \cap X_j) - f(Y \cap X_{j-1}) \\
&\geq f(X_0) + \mathbb{E}_{Y\sim N_\rho(X)} \sum_{j=1}^{n} f((Y \cap \{\pi(j)\}) \cup X_{j-1}) - f(X_{j-1}) \qquad (3) \\
&= f(X_0) + \sum_{j=1}^{n} \frac{1+\rho x_{\pi(j)}}{2}(f(X_j) - f(X_{j-1})) \qquad (4) \\
&\geq \sum_{j=0}^{n} \frac{\rho}{2}(x_{\pi(j)} - x_{\pi(j+1)})f(X_j) + \frac{1-\rho}{2}(f(X_0) + f(X_n)) \\
&= \rho f(X) + \frac{1-\rho}{2}(f(X_0) + f(X_n)).
\end{aligned}$$

The inequality in (3) can be shown either using the decreasing marginal returns characterization of the submodularity in (2) of f (with $S = Y \cap X_{j-1}$ and $T = X_{j-1}$, and $i = \pi_j$, unless Y does not contain π_j in which case the inequality becomes trivial), or the union-intersection characterization in (1) (by taking $S = Y \cap X_j$ and $T = X_{j-1}$ and noting that $X_{j-1} \subset X_j$ and $S \cup T = (Y \cap X_j) \cup X_{j-1} = (Y \cap \{\pi_j\}) \cup X_{j-1}$). The equality in (4) comes from moving the expectation inside and observing that each summand is non-zero only if $\pi(j) \in Y$. This happens with probability $(1+\rho)/2$ when $x_{\pi(j)} = 1$ and with probability $(1-\rho)/2$ when $x_{\pi(j)} = -1$. ∎

Remark. The proof technique is not new (for instance it was used by Madiman and Tetali [MT10] to show a large class of Shannon-type inequalities for the joint entropy function). In fact, it can be viewed as a special case of the "Threshold Lemma" [Von09]. However, to the best of our knowledge the statement of Lemma 7 has never been expressed using the language of noise operators.

Corollary 8. Let $f : \{\text{-}1, 1\}^n \to \mathbb{R}^+$ be a submodular function. Then for all $\rho \in [0, 1]$,

$$\mathbb{S}_\rho(f) \geq \rho \|f\|_2^2.$$

3.2 Product Distributions

For the rest of Section 3.2 we will assume that the distribution is a product distribution $\Pi = \Pi_1 \times \Pi_2 \times \cdots \times \Pi_n$ on $\{\text{-}1, 1\}^n$ such that $\Pr_{x_i \sim \Pi_i}[x_i = 1] = p_i$. Then, $p_{\min} = \min_{1 \leq i \leq n} p_i$ and $p_{\max} = \max_{1 \leq i \leq n} p_i$.

Lemma 9. Let $\Pi = \Pi_1 \times \Pi_2 \times \cdots \times \Pi_n$ be a product distribution over $\{\text{-}1, 1\}^n$ with maximum and minimum probabilities $p_{\max}$ and $p_{\min}$ and let $f : \{\text{-}1, 1\}^n \to \mathbb{R}^+$ be a submodular function. Then for all $x \in \{\text{-}1, 1\}^n$, $\rho \in [0, 1]$,

$$T_\rho f(x) \geq (\rho - (1 - \rho)(p_{\max} - p_{\min})) \cdot f(x).$$

Remark 1. Notice that since $0 \leq (p_{\max} - p_{\min}) \leq 1$, the above bound is always non-trivial for all $\rho \geq \frac{1}{2}$.

Proof. As in the proof of Lemma 7, we will be viewing the domain of f as $2^{[n]}$, and the input $x \in \{\text{-}1, 1\}^n$ as $X \in 2^{[n]}$ such that $\mathbf{1}(X) = x$. For $i \in [n]$, let $p_i = \Pr_{x_i \sim \Pi_i}[x_i = 1]$, $p_{\max} = \max_{i \in [n]} p_i$ and $p_{\min} = \min_{i \in [n]} \Pr_{x_i \sim \Pi_i}[x_i = 1]$.
In the proof of Lemma 7, we proceed by choosing a facilitating permutation on $[n]$. Choosing the right permutation here is a little more tricky and depends not just on x but also on the parameters of the marginal distributions on coordinates, that is, $p_1, p_2, \ldots, p_n$. For a fixed $x \in \{\text{-}1, 1\}^n$, let $\pi : [n] \to [n]$ be a permutation such that $x_{\pi(i)} \geq x_{\pi(i+1)}$ and whenever $x_{\pi(i)} = x_{\pi(i+1)}$, $p_{\pi(i)} \geq p_{\pi(i+1)}$ for every $i \in [n-1]$. In other words, the permutation sorts $[n]$ in descending order according to values x_i for each $i \in [n]$, breaking ties by using the the descending order in marginal probabilities p_i for each $i \in [n]$. This choice will be crucial in further analysis.

As in the proof of Lemma 7, define $X_j = \{\pi(1), \ldots, \pi(j)\}$. In addition, let $X_0 = \emptyset$ and $X_n = [n]$.

$$\mathbb{E}_{Y \sim N_\rho(X)} f(Y) = f(X_0) + \mathbb{E}_{Y \sim N_\rho(X)} \sum_{j=1}^{n} f(Y \cap X_j) - f(Y \cap X_{j-1}).$$

Now using submodularity of f (similar to the proof of Lemma 7), we get:

$$\begin{aligned} \mathbb{E}_{Y \sim N_\rho(X)} f(Y) &\geq f(X_0) + \mathbb{E}_{Y \sim N_\rho(X)} \sum_{j=1}^{n} f((Y \cap \{\pi(j)\}) \cup X_{j-1}) - f(X_{j-1}) \qquad (5) \\ &= f(X_0) + \sum_{j=1}^{n} [\Pr[\pi(j) \in Y]] \cdot (f(X_j) - f(X_{j-1})). \end{aligned}$$

Observe that $\Pr[\pi(j) \in Y] = \rho + (1-\rho)p_{\pi(j)}$ if $\pi(j) \in X$ ($x_{\pi(j)} = 1$) and $(1-\rho)p_{\pi(j)}$ when $\pi(j) \notin X$ ($x_{\pi(j)} = -1$). In either case, $\Pr[j \in Y] = \frac{\rho}{2} \cdot x_{\pi(j)} + \frac{\rho}{2} + (1-\rho)p_{\pi(j)}$. Thus,

$$\begin{aligned}
\mathbb{E}_{Y \sim N_\rho(X)} &= f(X_0) + \sum_{j=1}^{n} \left[\frac{\rho}{2} \cdot x_{\pi(j)} + \frac{\rho}{2} + (1-\rho)p_{\pi(j)}\right] (f(X_j) - f(X_{j-1})) \\
&= \sum_{j=1}^{n-1} \left[\frac{\rho}{2}(x_{\pi(j)} - x_{\pi(j+1)}) + (1-\rho)(p_{\pi(j)} - p_{\pi(j+1)})\right] f(X_j) \\
&+ \left(1 - \frac{\rho}{2} - \frac{\rho}{2}x_{\pi(1)} - (1-\rho)p_{\pi(1)}\right) f(X_0) \\
&+ \left(\frac{\rho}{2} + \frac{\rho}{2}x_{\pi(n)} + (1-\rho)p_{\pi(n)}\right) f(X_n). \qquad (6)
\end{aligned}$$

We now break the analysis into two cases for simplicity.
Suppose $X \notin \{\emptyset, [n]\}$; i.e., $x \notin \{(-1)^n, 1^n\}$. Then, there is exactly one $j^* \in [n-1]$ such that $x_{\pi(j^*)} \neq x_{\pi(j^*+1)}$ and, $X_{j^*} = X$. Also observe that whenever $x_{\pi(j)} = x_{\pi(j+1)}$, then $p_{\pi(j)} \geq p_{\pi(j+1)}$ by our choice of π. Hence, for all the terms in the summation in (6) indexed by $j \neq j^*$ have a non-negative coefficient and are thus non-negative as $f(x) \geq 0$ for every $x \in \{-1,1\}^n$. The coefficient of $X_{j^*} = X$ is $\frac{\rho}{2}(x_{\pi(j^*)} - x_{\pi(j^*+1)}) + (1-\rho)(p_{\pi(j^*)} - p_{\pi(j^*+1)}) = \rho + (1-\rho)(p_{\pi(j^*)} - p_{\pi(j^*+1)}) \geq \rho + (1-\rho)(p_{\max} - p_{\min})$ which yields the required result in this case.

When $X = \emptyset$ or $[n]$, all the terms in the summation in (6) are non-negative by our choice of π and non-negativity of f and the coefficient of $f(\emptyset)$ or $f([n])$ can be easily verified to be at least $\rho - (1-\rho)(p_{\max} - p_{\min})$. ■

As with Fourier analysis over the uniform distribution, it can be easily verified that

$$\mathbb{S}_\rho(f) = \langle f, T_\rho f \rangle = \sum_{S \subseteq [n]} \rho^{|S|} \hat{f}(S)^2$$

over any product distribution Π, where the Fourier coefficients are now defined with respect to the Gram-Schmidt orthonormalization of the χ basis with respect to the the Π-norm [Bah61, FJS91]. Thus, once again we get a lower-bound on the noise-stability of submodular functions as an immediate consequence of Lemma 9.

Theorem 10 (Theorem 3 Restated)**.** Let $\Pi = \Pi_1 \times \Pi_2 \times \cdots \times \Pi_n$ be a product distribution over $\{-1,1\}^n$ with minimum and maximum probabilities $p_{\min}$ and $p_{\max}$ and let $f : \{-1,1\}^n \to \mathbb{R}^+$ be a submodular function. Then for all $\rho \in [0,1]$,

$$\mathbb{S}_\rho(f) \geq (\rho - (1-\rho)(p_{\max} - p_{\min}))\|f\|_2^2.$$

Remark 2. As before, this theorem is interesting only when $(p_{\max} - p_{\min}) \leq \rho \leq 1$, as otherwise the right hand side of the inequality above is non-positive. However, $\rho \in [\frac{1}{2}, 1]$ always yields a non-trivial bound on $\mathbb{S}_\rho(f)$.

4 Learning

In the agnostic learning framework [KSS94], the learner receives labelled examples (x, y) drawn from a fixed distribution over example-label pairs.

Definition 11 (Agnostic Learning)**.** Let $\mathcal{D}$ be any distribution on $\{\text{-}1,1\}^n \times \mathbb{R}$ such that the marginal distribution over $\{\text{-}1,1\}^n$ is a product distribution Π. Define

$$opt = \min_{f \in \mathcal{C}} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[|f(x) - y|\right].$$

That is, *opt* is the error of the best fitting L_1-approximation in $\mathcal{C}$ with respect to $\mathcal{D}$.

We say that an algorithm A agnostically learns a concept class $\mathcal{C}$ over Π if the following holds for any $\mathcal{D}$ with marginal Π: if A is given random examples drawn from $\mathcal{D}$, then with high probability A outputs a hypothesis h such that $\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[|h(x) - y|\right] \leq opt + \epsilon$.

The following lemma, considered folklore (see [KOS04]), shows that noise stable functions are well-approximated by low-degree polynomials.

Lemma 12. Let $\Pi = \Pi_1 \times \Pi_2 \times \cdots \times \Pi_n$ be a product distribution over $\{\text{-}1,1\}^n$, $\rho \in [0,1]$ be any fixed constant and let $f : \{\text{-}1,1\}^n \to \mathbb{R}$ be a function such that $\|f\|_2 = 1$ and $\mathbb{S}_\rho(f) \geq 1 - 2\gamma$. Then there exists a multilinear polynomial $p : \{\text{-}1,1\}^n \to \mathbb{R}$ of degree $2/(1-\rho)$ such that

$$\mathbb{E}_{x\sim\Pi}\left[(f-p)^2\right] < \left(\frac{2}{1-e^{-2}}\right)\gamma = O(\gamma).$$

The "L_1 Polynomial Regression Algorithm" due to Kalai et al. [KKMS08] shows that one can *agnostically* learn low-degree polynomials.

Theorem 13 ([KKMS08])**.** Suppose $\mathbb{E}_{x\sim\mathcal{D}_X}[(f-p)^2] < \epsilon^2$ for some degree d polynomial p, some distribution $\mathcal{D}$ on $X \times \mathbb{R}$ where the marginal $\mathcal{D}_X$ is a product distribution on $\{\text{-}1,1\}^n$, and any f in the concept class $\mathcal{C}$. Then, with probability $1-\delta$, the L_1 Polynomial Regression Algorithm outputs a hypothesis h such that $\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[|h(x) - y|\right] \leq opt + \epsilon$ in time $\text{poly}(n^d/\epsilon, \log(1/\delta))$.

Now by combining the regression algorithm with Lemma 12 and the noise stability results (Corollary 8 and Theorem 10 with an appropriately chosen parameter $\rho > \frac{1}{2}$ to make the stability bound close to 1), we obtain the following result.

Corollary 14. Let $\mathcal{C}$ be the class of non-negative submodular functions with $\|f\|_2 = 1$ and let $\mathcal{D}$ be any distribution on $\{\text{-}1,1\}^n \times \mathbb{R}$ such that the marginal distribution over $\{\text{-}1,1\}^n$ is a product distribution. Then for all $f \in \mathcal{C}$, the L_1 Polynomial Regression Algorithm outputs a hypothesis h with probability $1-\delta$ such that

$$\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[|h(x) - y|\right] \leq opt + \epsilon,$$

given random examples in time $\text{poly}(n^{O(1/\epsilon^2)}/\epsilon, \log(1/\delta))$.

We note that the L_1 Polynomial Regression Algorithm can be implemented as a statistical query algorithm [Kal11]. (N.B.: The access offered to the learning algorithm by the statistical query model is much weaker than that offered by random examples or the tolerant value query model. The tolerant value query model allows arbitrary value queries that get answered with some noise, whereas the statistical query model requires that the queries to be of the form $g : \{\text{-}1,1\}^n \times \mathbb{R} \to \mathbb{R}$ where g is computable by a $\text{poly}(n, 1/\epsilon)$-size circuit, and the answer is $\mathbb{E}_{(x,y)\sim\mathcal{D}}[g(x,y)]$ with some noise.)

Corollary 15 (Corollary 4 Restated)**.** Let $\mathcal{C}$ be the class of non-negative submodular functions with $\|f\|_2 = 1$ and let $\mathcal{D}$ be any distribution on $\{\text{-}1, 1\}^n \times \mathbb{R}$ such that the marginal distribution over $\{\text{-}1, 1\}^n$ is a product distribution. Then for all $f \in \mathcal{C}$, there is a statistical query algorithm that outputs a hypothesis h with probability $1 - \delta$ such that

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[|h(x) - y|] \leq opt + \epsilon,$$

in time $\text{poly}(n^{O(1/\epsilon^2)}, \log(1/\delta))$.

5 Private Query Release and Low-Degree Polynomials

In this section, we make a simple observation connecting approximability by low-degree polynomials with private query release.

The study of differential privacy is motivated by a scenario in which a database is stored on a trusted server, and the goal is to extract global statistics on the stored data without affecting privacy of the individual records. For instance, suppose that the database stores a set of different binary features on a population of n individuals (e.g., outcomes of a vote) and we wish to query the average number of individuals for whom one or more feature is positive. Roughly speaking, a query outcome is differentially private if it is not sensitive to any individual record of the database, in the sense that modifying any single record does not affect the probability distribution of the outcome. In this section we briefly review the notions in differential privacy that we use. However, for a more detailed discussion we refer the reader to [GHRU11] that is most relevant to our discussion.

In the context of differential privacy, we will call $D \subset X$ a *database* and two databases $D, D' \subset X$ are *adjacent* if one can be obtained from the other by adding a single item.

Definition 16 (Differential privacy [DMNS06])**.** An algorithm $A : X^* \to R$ is *ϵ-differentially private* if for all $Q \subset R$ and every pair of adjacent databases D, D', we have $\Pr[A(D) \in Q] \leq e^\epsilon \Pr[A(D') \in Q]$.

A *counting query* over a database D is just the average value of a query over each entry in the database.

Definition 17 (Counting Query Function)**.** Let $c : X \to \mathbb{R}$ be a real-valued query function. For a fixed $r \in X$, let $\mathbf{q}_r(c) := c(r)$. For a class of queries $\mathcal{C}$ and a fixed database $D \subset X$, the *counting query function* $\mathbf{CQ}_D : \mathcal{C} \to \mathbb{R}$ is the function defined by $\mathbf{CQ}_D(c) := \frac{1}{n}\sum_{r\in D} \mathbf{q}_r(c) = \frac{1}{n}\sum_{r\in D} c(r)$, where $n := |D|$ is the size of the database.

A *counting query releasing* algorithm's objective is to release a data structure H whose answers on queries $c \in \mathcal{C}$ are close to those of the counting query over the original database D.

Definition 18 (Counting query release [GHRU11])**.** Let $\mathcal{C}$ be a class of queries c from $X \to \mathbb{R}$, and let Π be a distribution on $\mathcal{C}$. We say that an algorithm A (α, β)-releases $\mathcal{C}$ over a database D of size n, if for $H = A(D)$,

$$\Pr_{c\sim\Pi}[|\mathbf{CQ}_D(c) - H(c)| \leq \alpha] \geq 1 - \beta.$$

The following proposition is implicit in [GHRU11] using results of [BDMN05] and [KLN+08].

Proposition 19. For a given concept class $\mathcal{C}$ with distribution Π, if there is a query learning algorithm for the concept class $\{\mathbf{CQ}_D : D \subset X\}$ using q τ-tolerant value queries[1] that outputs a hypothesis H s.t. $\Pr_{c\in\Pi}[|\mathbf{CQ}_D(c) - H(c)| \leq \alpha] \geq 1 - \beta$, then there is an ϵ-differentially private algorithm that (α, β)-releases $\mathcal{C}$ for any database of size $|D| \geq q(\log q + \log(1/\beta))/\epsilon\tau$.

For instance, Gupta et al. [GHRU11] show that for $\mathcal{C}$, the class of disjunctions, the class $\{\mathbf{CQ}_D : D \subset X\}$ is a submodular function. Thus, their tolerant value query learning algorithm for submodular functions leads to a private counting query release algorithm for disjunctions (equivalently conjunctions).

We make the following observation. For a given concept class $\mathcal{C}$ with distribution Π, if for every $r \in X$, $\mathbf{q}_r$ is well-approximated by a low-degree polynomial with respect to Π, then $\mathbf{CQ}_D$ is also well-approximated by a low-degree polynomial with respect to Π. As statistical queries are strictly weaker than tolerant value queries, the L_1 Polynomial Regression Algorithm satisfies the requirements of Proposition 19, and we have a private counting query release algorithm for $\mathcal{C}$. We note that it is easy to see that a $O(\log(1/\alpha))$-degree polynomial can L_1-approximate $\mathbf{q}_r$ to within α, when $\mathcal{C}$ is the class of disjunctions, and Π is the uniform distribution. (If $|r| = O(\log(1/\alpha))$, a $O(\log(1/\alpha))$-degree polynomial can interpolate the function exactly. Otherwise, the constant 1 function is within α of $\mathbf{q}_r$.) Thus, we are able to retrieve the result of Gupta et al. [GHRU11] on releasing disjunctions easily with an improved running-time of $|X|^{O(\log(1/\alpha))}$ as opposed $|X|^{O(1/\alpha^2)}$.

Theorem 20. There is an ϵ-differentially private algorithm (α, β)-releases the class of all Boolean disjunctions in time $|X|^{O(\log(1/\alpha))}$ for any database of size $|D| \geq |X|^{O(\log(1/\alpha))}/\epsilon$.

Theorem 20 can easily be generalized to hold for all product distributions as well as k-DNF formulas for any constant k.

As a second example, we consider the case where $\mathcal{C}$ is the class of linear threshold functions, or halfspaces, $c(r_1, \ldots, r_d) = \text{sgn}(w_1 r_1 + \cdots + w_d r_d)$ where $r_i, w_i \in \mathbb{R}$. The distribution Π over halfspaces is specified by picking the weights w_i according to a spherical Gaussian over $\mathbb{R}^d$. (We could also work over the uniform distribution on the sphere, but for simplicity we work with Gaussians). In this example the function $q_r(c) = c(r) = \text{sgn}(\sum_{i=1}^{d} w_i r_i)$ can be viewed as a halfspace whose inputs (the w_i's) are chosen according to a spherical Gaussian. It is well-known that halfspaces have high Gaussian noise stability and therefore can be approximated (within error α in L_1 norm) by a polynomial of degree $O(1/\alpha^4)$ [KOS08].

Theorem 21. There is an ϵ-differentially private algorithm that (α, β)-releases the class of all halfspaces in time $|X|^{O(1/\alpha^4)}$ for any database of size $|D| \geq |X|^{O(1/\alpha^4)}/\epsilon$.

Acknowledgements

We would like to thank Aaron Roth for explaining [GHRU11] to us. We would also like to thank Ashwinkumar B. V. for pointing out an error in the previous version of this paper.

[1] A query is τ-tolerant if its output differs from the actual value by an error of absolute value at most τ.

References

[Bah61] Raghu Bahadur. A representation of the joint distribution of responses to n dichotomous items. In *Studies in Item Analysis and Prediction*, pages 158–168. Stanford University Press, 1961.

[BDMN05] Avrim Blum, Cynthia Dwork, Frank Mcsherry, and Kobbi Nissim. Practical privacy: The sulq framework. In *Proceedings of the Twenty-fourth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*, pages 128–138. ACM Press, 2005.

[BH11] Maria-Florina Balcan and Nicholas J. A. Harvey. Learning submodular functions. In *Proc. 43rd Annual ACM Symposium on Theory of Computing (STOC)*. ACM Press, 2011.

[BKS99] Itai Benjamini, Gil Kalai, and Oded Schramm. Noise sensitivity of boolean functions and applications to percolation. *Publications Mathématiques de l'I.H.E.S.*, 90:5–43, 1999.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proc. 3rd Theory of Cryptography Conference (TCC)*, Lecture Notes in Computer Science. Springer-Verlag, 2006.

[DNS06] Shahar Dobzinski, Noam Nisan, and Michael Schapira. Truthful randomized mechanisms for combinatorial auctions. In *Proc. 38th Annual ACM Symposium on Theory of Computing (STOC)*, pages 644–652. ACM Press, 2006.

[Edm71] Jack Edmonds. Matroids and the greedy algorithm. *Mathematical Programming*, 1(1):127–136, 1971.

[FJS91] Merrick Furst, Jeffrey C. Jackson, and Sean W. Smith. Improved learning of ac^0 functions. In *Proc. of the 4th Annual Conference on Computational Learning Theory (COLT)*, Lecture Notes in Computer Science, pages 317–325. Springer-Verlag, 1991.

[FMV07] Uriel Feige, Vahab S. Mirrokni, and Jan Vondrák. Maximizing non-monotone submodular functions. In *Proc. 48th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 461–471. IEEE Computer Society Press, 2007.

[FNW78] Marshall L. Fischer, George L. Nemhauser, and Laurence A. Wolsey. An analysis of approximations for maximizing submodular set functions II. *Mathematical Programming Studies*, 8:73–87, 1978.

[GHRU11] Anupam Gupta, Moritz Hardt, Aaron Roth, and Jonathan Ullman. Privately releasing conjunctions and the statistical query barrier. In *Proc. 43rd Annual ACM Symposium on Theory of Computing (STOC)*. ACM Press, 2011.

[Hås01] Johan Håstad. Some optimal inapproximability results. *Journal of the ACM*, 48(4):798–859, 2001. Prelim. ver. in *Proc. of STOC'97*.

[Kal11] Adam Kalai. Personal communication, 2011.

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

[KKL88] Jeff Kahn, Gil Kalai, and Nathan Linial. The influence of variables on Boolean functions. In *Proc. 29th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 68–80. IEEE Computer Society Press, 1988.

[KKMS08] Adam Kalai, Adam Klivans, Yishay Mansour, and Rocco A. Servedio. Learning intersections and thresholds of halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008. Prelim. ver. in *Proc. of FOCS'05*.

[KLN+08] Shiva Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? In *Proc. 49th IEEE Symposium on Foundations of Computer Science (FOCS)*, 2008.

[KOS04] Adam Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning intersections and thresholds of halfspaces. *Journal of Computer and System Sciences*, 68(4):808–840, 2004. Prelim. ver. in *Proc. of FOCS'02*.

[KOS08] Adam Klivans, Ryan O'Donnell, and Rocco A. Servedio. Learning geometric concepts via Gaussian surface area. In *Proc. 49th IEEE Symposium on Foundations of Computer Science (FOCS)*. IEEE Computer Society Press, 2008.

[KSS94] Michael J. Kearns, Robert E. Schapire, and Linda M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994. Prelim. ver. in *Proc. of COLT'92*.

[Lov83] László Lovász. Submodular functions and convexity. In A. Bachem et al., editor, *Mathematical Programming: The State of the Art*, pages 235–257. Springer-Verlag, 1983.

[MOO10] Elchanan Mossel, Ryan O'Donnell, and Krzysztof Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. *Annals of Mathematics*, 171(1):295–341, 2010. Prelim. ver. in *Proc. of FOCS'05*.

[MR07] Elchanan Mossel and Sebastien Roch. On the submodularity of influence in social networks. In *Proc. 39th Annual ACM Symposium on Theory of Computing (STOC)*, pages 128–134. ACM Press, 2007.

[MT10] Mokshay Madiman and Prasad Tetali. Information inequalities for joint distributions, with interpretations and applications. *IEEE Transactions on Information Theory*, 56(6):2699–2713, 2010.

[NWF78] G. L. Nemhauser, L. A. Wolsey, and M. L. Fischer. An analysis of approximations for maximizing submodular set functions I. *Mathematical Programming*, 14(1):265–294, 1978.

[O'D04] Ryan O'Donnell. Hardness amplification within NP. *Journal of Computer and System Sciences*, 69(1):68–94, 2004. Prelim. ver. in *Proc. of STOC'02*.

[OV11] Shayan Oveis Gharan and Jan Vondrák. Submodular maximization by simulated annealing. In *Proc. of the 22nd Annual ACM/SIGACT-SIAM Symposium on Discrete Algorithms (SODA)*, 2011.

[Von09] Jan Vondrák. Symmetry and approximability of submodular maximization problems. In *Proc. 50th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 651–670. IEEE Computer Society Press, 2009.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de