# Tight Bounds on Low-degree Spectral Concentration of Submodular and XOS Functions

Vitaly Feldman
*IBM Almaden Research Center*
*650 Harry rd.*
*San Jose, CA 95120*
*vitaly@post.harvard.edu*

Jan Vondrák
*IBM Almaden Research Center*
*650 Harry rd.*
*San Jose, CA 95120*
*jvondrak@gmail.com*

**Abstract**

Submodular and fractionally subadditive (or equivalently XOS) functions play a fundamental role in combinatorial optimization, algorithmic game theory and machine learning. Motivated by learnability of these classes of functions from random examples, we consider the question of how well such functions can be approximated by low-degree polynomials in $\ell_2$ norm over the uniform distribution. This question is equivalent to understanding the concentration of Fourier weight on low-degree coefficients, a central concept in Fourier analysis. Denoting the smallest degree sufficient to approximate $f$ in $\ell_2$ norm within $\epsilon$ by $\mathsf{deg}^{\ell_2}_{\epsilon}(f)$, we show that

- For any submodular function $f : \{0,1\}^n \to [0,1]$, $\mathsf{deg}^{\ell_2}_{\epsilon}(f) = O(\log(1/\epsilon)/\epsilon^{4/5})$ and there is a submodular function that requires degree $\Omega(1/\epsilon^{4/5})$.
- For any XOS function $f : \{0,1\}^n \to [0,1]$, $\mathsf{deg}^{\ell_2}_{\epsilon}(f) = O(1/\epsilon)$ and there exists an XOS function that requires degree $\Omega(1/\epsilon)$.

This improves on previous approaches that all showed an upper bound of $O(1/\epsilon^2)$ for submodular [CKKL12], [FKV13], [FV13] and XOS [FV13] functions. The best previous lower bound was $\Omega(1/\epsilon^{2/3})$ for monotone submodular functions [FKV13]. Our techniques reveal new structural properties of submodular and XOS functions and the upper bounds lead to nearly optimal PAC learning algorithms for these classes of functions.

**Keywords**

submodular function, XOS function, spectral concentration

## I. Introduction

Analysis of the discrete Fourier transform of functions over the hypercube has a wide range of notable applications in theoretical computer science. It is also the object of significant research interest in its own right [O'D14]. While most of this research has been devoted to Boolean-valued functions, many works analyze general real-valued functions (e.g. [Tal94], [DFKO06]). Recently, the analysis of real-valued functions over the hypercube has also attracted significant attention due to applications in learning theory, property testing, differential privacy, algorithmic game theory and quantum complexity [GHIM09], [BH12], [GHRU11], [SV11], [CKKL12], [BDF+12], [BCIW12], [RY13], [FKV13], [FV13], [FK14], [BRY14], [AA14], [BB14]. Most of the Fourier-analytic techniques apply to real-valued functions as well but many new questions arise when one considers the richer structure of real-valued functions.

Our focus is on *structural properties* of two fundamental classes of real-valued functions: submodular and fractionally subadditive. Submodularity, a discrete analog of convexity, has played an essential role in combinatorial optimization [Edm70], [Lov83], [Que95], [Fra97], [FFI01] and, more recently, in algorithmic game theory and machine learning [GKS05], [BLN06], [DS06], [KGGK06], [KSG08], [Von08]. In algorithmic game theory, submodular functions have found application as *valuation functions* with the property of diminishing returns [BLN06], [DS06], [Von08]. Along with submodular functions, fractionally subadditive functions have been studied in the algorithmic game theory context [BLN06] (see Sec. II for the definition). Feige showed that these functions have an additional characterization as a maximum of non-negative linear functions or XOS [Fei06]. Here we also show that the Rademacher complexity of a set of vectors that plays a fundamental role in statistical learning gives yet another equivalent way to define this class of functions. For comparison, we also discuss the class of self-bounding functions that contains both submodular and XOS functions and shares a number of properties with those classes such as dimension-free concentration of measure [BLM00]. Informally, a function $f : \{0,1\}^n \to \mathbb{R}$ is self-bounding if for every $x \in \{0,1\}^n$, $f(x)$ upper bounds the sum of all the $n$ marginal decreases in the value of the function at $x$. We define these classes and their relationships in Section II.

The primary property we consider is how well these functions can be approximated by *low-degree polynomials*, where the approximation is measured in $\ell_2$ norm over the uniform distribution $\mathcal{U}$ defined as $\|f-g\|_2 = \sqrt{\mathbf{E}_{\mathcal{U}}[(f(x)-g(x))^2]}$. By

| Class of functions | lower bound | upper bound |
|---|---|---|
| linear | 1 | 1 |
| coverage | $\Omega(\log(1/\epsilon))$[FK14] | $O(\log(1/\epsilon))$ [FK14] |
| submodular | $\Omega(1/\epsilon^{4/5})$ | $O(1/\epsilon^{4/5} \cdot \log(1/\epsilon))$ |
| XOS | $\Omega(1/\epsilon)$ | $O(1/\epsilon)$ |
| self-bounding | $\Omega(1/\epsilon^2)$ | $O(1/\epsilon^2)$[FV13] |

Figure 1. Overview of low-degree approximations: bounds on $(\ell_2, \epsilon)$-approximate degree for a function with range $[0, 1]$.

the standard duality for the $\ell_2$ norm, approximability of $f$ by polynomials of degree $d$ is characterized by how much of $f$'s Fourier weight resides on coefficients of degree above $d$. Concentration of the Fourier spectrum on low-degree coefficients is one of the central and most well-studied properties in Fourier analysis and its applications. In particular, following the seminal work of Linial, Mansour and Nisan [LMN93], a large number of learning algorithms over the uniform (and other) distributions relies crucially on approximation by low-degree polynomials (e.g. [KKMS08], [KS08], [KKM13]).

Motivated by learning of submodular functions and its application in differential privacy in [GHRU11], Cheraghchi *et al.* [CKKL12] proved that every submodular function[1] can be $\epsilon$-approximated in $\ell_2$ norm by a polynomial of degree $O(1/\epsilon^2)$. Their proof is based on the analysis of the noise sensitivity of submodular functions, a standard tool from Fourier analysis for establishing low-degree spectral concentration. Subsequently, Feldman *et al.* proved the same upper bound of $O(1/\epsilon^2)$ using approximation of submodular functions by real-valued decision trees [FKV13]. They also gave a lower bound for learning that implies a lower bound of $\Omega(1/\epsilon^{2/3})$ on the degree necessary to $\epsilon$-approximate submodular functions.

Most recently, we considered the approximability of submodular and XOS functions by functions of few variables or *juntas* [FV13]. We showed that submodular functions are $\epsilon$-approximated in $\ell_2$ by functions depending on $O(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon})$ variables, while for XOS functions, a junta of size $2^{O(1/\epsilon^2)}$ suffices. In addition, we showed that submodular and XOS functions (in fact, all self-bounding functions) have constant total influence implying that they can be approximated by a polynomial of degree $O(1/\epsilon^2)$. These results have lead to substantially faster learning and testing algorithms for these classes of functions, most notably, a $2^{\tilde{O}(1/\epsilon^2)} \cdot n^2$ time PAC learning algorithm for submodular functions and a $2^{O(1/\epsilon^4)} \cdot n$ time PAC learning algorithm for XOS functions. Learning of submodular and XOS functions is also the main motivating application of this work.

### A. *Our Results*

In this work, we investigate the degree that is necessary to approximate XOS and submodular functions in detail. For a real-valued function $f : \{0,1\}^n \to \mathbb{R}$ let $\mathsf{deg}^{\ell_2}_\epsilon(f)$ denote the smallest $d$ such that there is a polynomial $p$ of degree $d$ for which $\|f - p\|_2 \le \epsilon$ and we refer to it as $(\ell_2, \epsilon)$-approximate degree of $f$. The three known upper bounds on $(\ell_2, \epsilon)$-approximate degree of submodular functions are all $O(1/\epsilon^2)$ [CKKL12], [FKV13], [FV13]. The bounds are derived via three different approaches suggesting that this might be the right answer. This bound also applies to XOS and self-bounding functions [FV13] and the known lower bound of $\Omega(1/\epsilon^{2/3})$ also applies to all of these classes of functions [FKV13]. Here we show that, in fact, the picture is substantially richer: each of these classes requires a different degree to approximate that corresponds to the increasing complexity of functions in these classes. We detail our bounds below and also summarize them in Figure 1.

- For any submodular function $f : \{0,1\}^n \to [0,1]$, $\mathsf{deg}^{\ell_2}_\epsilon(f) = O(\log(1/\epsilon)/\epsilon^{4/5})$. This is almost tight: we prove that even for very simple submodular functions of the form $f(x) = \min\{\frac{2}{k}\sum_{i=1}^k x_i, 1\}$ for some $k$, $\mathsf{deg}^{\ell_2}_\epsilon(f) = \Omega(1/\epsilon^{4/5})$.
- For any XOS function $f : \{0,1\}^n \to [0,1]$, $\mathsf{deg}^{\ell_2}_\epsilon(f) = O(1/\epsilon)$. We also show that degree $\Omega(1/\epsilon)$ is necessary for XOS functions.

For comparison we show that the bounds above do not hold for the more general class of self-bounded functions (and, consequently, for functions with constant total influence). Namely, we show that there exists a self-bounding function $f : \{0,1\}^n \to [0,1]$, such that $\mathsf{deg}^{\ell_2}_\epsilon(f) = \Omega(1/\epsilon^2)$. This matches the upper bound in [FV13]. As an additional point of comparison, coverage functions, a subclass of submodular functions, can be approximated by polynomials of exponentially smaller $O(\log(1/\epsilon))$ degree [FK14]. At the same time monotone functions and subadditive functions cannot be approximated by polynomials of dimension-free degree and require $\Omega(\sqrt{n})$ and $\Omega(n)$ degree, respectively, to approximate within a constant.

As a first application we show that the improved upper bound on $\mathsf{deg}^{\ell_2}_\epsilon$ of XOS functions leads to an upper bound of $2^{O(1/\epsilon)}$ on the size of junta sufficient to approximate an XOS function within $\ell_2$ error of $\epsilon$. This improves on the $2^{O(1/\epsilon^2)}$ upper bound and matches the lower bound of $2^{\Omega(1/\epsilon)}$ in [FV13].

[1]Here and below we normalize the function range to $[0, 1]$.

*Our techniques:* It is easy to verify that previous approaches to proving upper bounds on $\mathsf{deg}_\epsilon^{\ell_2}$ cannot lead to upper bounds stronger than $1/\epsilon^2$ even in the case of submodular functions. For example, a bound on the total sum of squared influences $\mathsf{Inf}^2(f)$ leads to $\mathsf{deg}_\epsilon^{\ell_2}(f) \leq \mathsf{Inf}^2(f)/\epsilon^2$. However, $\mathsf{Inf}^2(f) = 1$ even for the monotone submodular function $f(x) = x_1$.

The first step of both of our upper bounds is a spectral concentration bound based on the total *second-degree* influences. Namely, we consider the quantity $\sum_{i,j=1}^n \|\partial_{ij} f\|_2^2$, where $\partial_{ij} f$ is a second-degree discrete partial derivative of $f$. This quantity measures interactions between pairs of variables. It is particularly meaningful in the setting of submodular functions, where it measures the drop in marginal value of element $i$ due to the presence of $j$. That is, we always have $\partial_{ij} f \leq 0$ for submodular functions. We prove that for XOS functions the quantity $\sum_{i,j=1}^n \|\partial_{ij} f\|_2^2$ is at most a constant. This leads to an upper bound of $O(1/\epsilon)$ on $\mathsf{deg}_\epsilon^{\ell_2}(f)$, since $\sum_{i,j=1}^n \|\partial_{ij} f\|_2^2 \simeq 16 \cdot \sum_S |S|^2 \hat{f}^2(S)$. The proof of this bound is based on a careful analysis of contributions of the linear functions in the XOS representation. In particular, it also reveals that XOS functions satisfy a degree-two version of self-boundedness property: for all $x$, $\sum_{i,j:x_i=x_j=1} (\partial_{ij} f(x))^2 \leq 5(f(x))^2$ (for comparison, self-bounding monotone functions satisfy $\sum_{i:x_i=1} \partial_i f(x) \leq f(x)$).

The upper bound above is optimal for XOS functions. To prove this we give an embedding of monotone DNF formulas into XOS functions. We then use the high noise sensitivity of Talagrand's random DNF [MO02] to prove our lower bound on the low-degree spectral concentration.

For submodular functions, we use a different approach to obtain the stronger bound. We examine how the sum of second-degree influences $\sum_{i,j=1}^n \|\partial_{ij} f\|_2^2$ behaves when no individual influence is too large. The technical notion of "large" that we use is the following "threshold norm": $\|\partial_i f\|_T = \sup\{\alpha \geq 0 : \Pr[|\partial_i f(x)| \geq \alpha] \geq \alpha^3\}$. We prove that at most $O(\frac{1}{\epsilon} \log \frac{1}{\epsilon})$ partial derivatives can be large in the sense that $\|\partial_i f\|_T > \epsilon$. This result is a special case of almost-everywhere boundedness of almost all the partial derivatives of a submodular function that we show. Namely, the number of variables $i$ for which $\Pr[|\partial_i f(x)| \geq \alpha] \geq \delta$ is at most $O(\log(1/\delta)/\epsilon)$. To prove this result we rely on the "boosting lemma" of Goemans and Vondrak [GV06], also used in our recent work [FV13]. (We note that an equivalent statement also appeared in [KK07].) Finally, we prove that for submodular functions with partial derivatives bounded by $\|\partial_i f\|_T \leq \epsilon$, we have $\sum_{i,j=1}^n \|\partial_{ij} f\|_2^2 = \tilde{O}(\sqrt{\epsilon})$. This leads to the upper bound of $\tilde{O}(1/\epsilon^{4/5})$.

As a warm-up to the upper bound for general submodular functions we also show a substantially simpler analysis for totally symmetric submodular functions (functions invariant under permutations of variables). In this case we avoid the logarithmic factor and get an $O(1/\epsilon^{4/5})$ upper bound. While the exponent of $\epsilon$ in our upper bound is quite unexpected it is actually optimal. In particular, using direct estimation of spectral concentration we show that the simple function $f(x) = \min\{\frac{2}{k} \sum_{i=1}^k x_i, 1\}$ requires degree $\Omega(1/\epsilon^{4/5})$ for $k = \Theta(1/\epsilon^{4/5})$. This function is monotone, totally symmetric, budget-additive and also can be viewed as a scaled rank function of a uniform matroid. Hence the lower bound applies to these subclasses of submodular functions as well. We remark that the weaker lower bound of $\Omega(1/\epsilon^{2/3})$ in [FKV13] was given for same function.

Finally, the lower bound of $\Omega(1/\epsilon^2)$ for self-bounding functions is based on an embedding of any Boolean function into a self-bounding Boolean function over $n + \log(n) + O(1)$ variables using the Hamming error-correcting code of distance 3.

*Learning:* The new structural results directly translate into improved learning algorithms using the techniques from [FKV13], [FV13]. For brevity we describe the improvements for learning from random examples in the PAC model with $\ell_2$ error. Similar improvements can be obtained for agnostic learning and learning with value queries (which allow the learner to ask for the value of the function at any point). For both XOS and submodular functions we give a new lower bound which shows that our learning algorithms are close to optimal.

**Theorem 1.** *There exists an algorithm $\mathcal{A}$ that given $\epsilon > 0$ and access to random uniform examples of a submodular XOS function $f : \{0,1\}^n \to [0,1]$, with probability at least $2/3$, outputs a function $h$, such that $\|f - h\|_2 \leq \epsilon$. Further, $\mathcal{A}$ runs in time $2^{\tilde{O}(1/\epsilon^{4/5})} \cdot n^2$ and uses $2^{\tilde{O}(1/\epsilon^{4/5})} \log n$ random examples.*

The best previous algorithm for this task runs in time $2^{\tilde{O}(1/\epsilon^2)} \cdot n^2$ and uses $2^{\tilde{O}(1/\epsilon^2)} \log n$ random examples [FV13]. We complement the new learning upper bound by a nearly tight information-theoretic lower bound of $2^{\Omega(1/\epsilon^{4/5})}$ examples (of value queries) for any PAC learning algorithm (see Thm. 33 for a formal statement).

The proof of the lower bound relies on the reduction in [FKV13]. The improved polynomial approximation and junta size for XOS functions lead to the following improved PAC learning algorithm.

**Theorem 2.** *There exists an algorithm $\mathcal{A}$ that given $\epsilon > 0$ and access to random uniform examples of an XOS function $f : \{0,1\}^n \to [0,1]$, with probability at least $2/3$, outputs a function $h$, such that $\|f - h\|_2 \leq \epsilon$. Further, $\mathcal{A}$ runs in time $r^{O(1/\epsilon)} \cdot n$ and uses $r^{O(1/\epsilon)} \log n$ random examples, where $r = \min\{n, 2^{1/\epsilon}\}$.*

The best previous upper bound was polynomial in $r^{O(1/\epsilon^2)}$ for $r = \min\{n, 2^{1/\epsilon^2}\}$ [FV13]. We prove that any PAC algorithm for XOS functions requires $2^{\Omega(1/\epsilon)}$ examples(see Thm. 40 for a formal statement). This upper bound is close to being tight when $n$ is subexponential in $2^{1/\epsilon}$. The lower bound is based on the embedding of monotone DNF into XOS functions that we used in the lower bound on $\deg_\epsilon^{\ell_2}$ together with the lower bound for learning monotone DNF of Blum *et al.* [BBL98]. Finally, using the Hamming code-based embedding we mentioned above we give a stronger lower bound of $2^{\Omega(1/\epsilon^2)}$ examples for any PAC learning algorithm for learning self-bounding functions.

*Organization:* Following the preliminaries we present the proofs of our main upper bounds: in Section III for XOS functions and in Section IV for submodular functions. Applications to approximation of XOS functions by juntas and learning algorithms appear in Section V. Details of lower bounds on spectral concentration and learning appear in Section VI. In the full version of this work [FV15] we prove the equivalence of Rademacher complexity and XOS functions.

### *B. Related Work*

Analysis of functions on the Boolean hypercube has a long history with strong ties to combinatorics, probability, learning theory, cryptography and complexity theory (see [O'D14]). One of the fundamental and most well-studied properties of Boolean functions is monotonicity. There is now a rich and detailed literature on the structure of the Fourier spectrum of monotone Boolean functions and their learnability over the uniform distribution [KLV94], [Tal94], [Tal96], [BT96], [BBL98], [MO02], [O'D03], [AM06], [DLM+08], [OW13], [DSFT+15]. Starting with the work of Goldreich *et al.* [GGL+00] numerous works have also investigated testing of monotone functions over the Boolean hypercube. Submodularity is closely related to monotonicity: indeed a function is submodular if and only if its partial derivatives are monotone non-increasing. In addition, XOS functions which are monotone share structural similarities with monotone DNF formulas (we make this explicit in Section VI-B). Hence our work is both inspired by the research on understanding of monotonicity over the Boolean hypercube and builds on techniques and results developed in that research. At the same time we are not aware of techniques closely related to those we use to prove our upper bounds for submodular and XOS functions having been used before.

We now review some recent work on learning of submodular, XOS and related classes of real-valued functions. Reconstruction of submodular functions up to some multiplicative factor (on every point) from value queries was first considered by Goemans *et al.* [GHIM09]. They show a polynomial-time algorithm for reconstructing monotone submodular functions with $\tilde{O}(\sqrt{n})$-factor approximation and prove a nearly matching lower-bound. This was extended to the class of all subadditive functions in [BDF+12] which studies small-size approximate representations of valuation functions (referred to as *sketches*).

Motivated by applications in economics, Balcan and Harvey initiated the study of learning submodular functions from random examples coming from an unknown distribution and introduced the PMAC learning model that requires a multiplicative approximation to the target function on most of the domain [BH12]. They give an $O(\sqrt{n})$-factor PMAC learning algorithm and show an information-theoretic $\Omega(\sqrt[3]{n})$-factor impossibility result for submodular functions. Subsequently, Balcan *et al.* gave a distribution-independent PMAC learning algorithm for XOS functions that achieves an $\tilde{O}(\sqrt{n})$-approximation and showed that this is essentially optimal [BCIW12].

Learning of submodular functions with additive rather than multiplicative guarantees over the uniform distribution was first considered by Gupta *et al.* who were motivated by applications in private data release [GHRU11]. They show that submodular functions can be $\epsilon$-approximated by a collection of $n^{O(1/\epsilon^2)}$ $\epsilon^2$-Lipschitz submodular functions. Concentration properties imply that each $\epsilon^2$-Lipschitz submodular function can be $\epsilon$-approximated by a constant. This leads to a learning algorithm running in time $n^{O(1/\epsilon^2)}$, which however requires value queries in order to build the collection. Using the upper bound of $O(1/\epsilon^2)$ on $\deg_\epsilon^{\ell_2}$ of submodular functions Cheraghchi *et al.* gave a $n^{O(1/\epsilon^2)}$ learning algorithm which uses only random examples and, in addition, works in the agnostic setting [CKKL12]. Feldman *et al.* show that the decomposition from [GHRU11] can be computed by a low-rank binary decision tree [FKV13]. They then show that this decision tree can then be pruned to obtain depth $O(1/\epsilon^2)$ decision tree that approximates a submodular function. This construction implies approximation by a $2^{O(1/\epsilon^2)}$-junta of degree $O(1/\epsilon^2)$. They also show how approximation by a junta can be used to obtain a $2^{O(1/\epsilon^4)}$ PAC learning algorithm for submodular functions. Feldman *et al.* extend the results on noise sensitivity of submodular functions in [CKKL12] to all self-bounding functions and show that they imply approximation within $\ell_1$ distance of $\epsilon$ by a polynomial of $O(\log(1/\epsilon)/\epsilon)$ degree and $2^{O(\log(1/\epsilon)/\epsilon)}$-junta [FKV14]. Note that approximation in $\ell_2$ norm we give here is stronger and our lower bound for self-bounding functions implies that any approach that works for all self-bounding functions cannot improve on the $O(1/\epsilon^2)$ bound on $\deg_\epsilon^{\ell_2}$.

Raskhodnikova and Yaroslavtsev consider learning and testing of submodular functions taking values in the range $\{0, 1, \ldots, k\}$ (referred to as *pseudo-Boolean*) [RY13]. The error of a hypothesis in their framework is the probability that the hypothesis disagrees with the unknown function. They show that pseudo-Boolean submodular functions can be expressed as $2k$-DNF and thus obtain a $\text{poly}(n) \cdot k^{O(k \log k/\epsilon)}$-time PAC learning algorithm using value queries. In a subsequent work, Blais *et* *al.* prove existence of a junta of size $(k\log(1/\epsilon))^{O(k)}$ and use it to give an algorithm for testing submodularity using $(k\log(1/\epsilon))^{\tilde{O}(k)}$ value queries [BOSY13].

## II. PRELIMINARIES

Let us define submodular, fractionally subadditive and subadditive functions. These classes are well known in combinatorial optimization and there has been a lot of recent interest in these functions in algorithmic game theory, due to their expressive power as *valuations* of self-interested agents.

**Definition 3.** *A set function $f : 2^N \to \mathbb{R}$ is*

- *monotone, if $f(A) \leq f(B)$ for all $A \subseteq B \subseteq N$.*
- *submodular, if $f(A \cup B) + f(A \cap B) \leq f(A) + f(B)$ for all $A, B \subseteq N$.*
- *subadditive, if $f(A \cup B) \leq f(A) + f(B)$ for all $A \subseteq B \subseteq N$.*
- *fractionally subadditive, if $f(A) \leq \sum \beta_i f(B_i)$ whenever $\beta_i \geq 0$ and $\sum_{i:a\in B_i} \beta_i \geq 1 \ \forall a \in A$.*

We identify functions on $\{0,1\}^n$ with set functions on $N = [n]$ in a natural way. By $\mathbf{0}$ and $\mathbf{1}$, we denote the all-zeroes and all-ones vectors in $\{0,1\}^n$ respectively. Submodular functions are not necessarily nonnegative, but in many applications (especially when considering multiplicative approximations), this is a natural assumption. All our approximations are shift-invariant and hence also apply to submodular functions with range $[-1/2, 1/2]$ (and can also be scaled in a straightforward way). Fractionally subadditive functions are nonnegative by definition (by considering $A = B_1, \beta_1 > 1$) and satisfy $f(\mathbf{0}) = 0$ (by considering $A = B_1 = \emptyset, \beta_1 = 0$). There is an equivalent definition known as "XOS" or maximum of non-negative linear functions [Fei06]: $f(x) = \max_{c\in C} \sum_{i=1}^n w_{ci} x_i$. Here, $w_{ci} \geq 0$ are nonnegative weights. This class includes all (nonnegative) monotone submodular functions such that $f(\mathbf{0}) = 0$ (but does not contain non-monotone functions). the full version of this work [FV15] we show that Rademacher complexity of a set of vectors, a powerful and well-studied tool in statistical learning theory [KP00], [BM02], gives an equivalent way to define XOS functions. We also show that the class of monotone self-bounding functions is stricly broader than than of XOS functions.

A broader class is that of *self-bounding functions*. Self-bounding functions were defined by Boucheron, Lugosi and Massart [BLM00] and further generalized by McDiarmid and Reed [MR06] as a unifying class of functions that enjoy strong concentration properties. Here, we define self-bounding functions in the special case of $\{0,1\}^n$ as follows. A function $f : \{0,1\}^n \to \mathbb{R}$ is called $a$-self-bounding, if $f$ is 1-Lipschitz and for all $x \in \{0,1\}^n$,

$$\sum_{i=1}^n (f(x) - f(x \oplus e_i))_+ \leq a f(x),$$

where $x \oplus e_i$ is $x$ with $i$-th bit flipped and $(\alpha)_+$ denotes $\max\{0, \alpha\}$. The 1-Lipschitz condition does not play a role in this paper, as we normalize functions to have values in the $[0,1]$ range. Self-bounding functions subsume fractionally subadditive functions, and 2-self-bounding functions subsume (possibly non-monotone) submodular functions. See [FV13] for a more detailed discussion of these classes of functions.

The $\ell_1$ and $\ell_2$-norms of $f : \{0,1\}^n \to \mathbb{R}$ are defined by $\|f\|_1 = \mathbf{E}_{x\sim\mathcal{U}}[|f(x)|]$ and $\|f\|_2 = (\mathbf{E}_{x\sim\mathcal{U}}[f(x)^2])^{1/2}$, respectively, where $\mathcal{U}$ is the uniform distribution.

**Definition 4** (Discrete derivatives)**.** *For $x \in \{0,1\}^n$, $b \in \{0,1\}$ and $i \in n$, let $x_{i\leftarrow b}$ denote the vector in $\{0,1\}^n$ that equals $x$ with $i$-th coordinate set to $b$. For a function $f : \{0,1\}^n \to \mathbb{R}$ and index $i \in [n]$ we define $\partial_i f(x) = f(x_{i\leftarrow 1}) - f(x_{i\leftarrow 0})$. We also define $\partial_{i,j} f(x) = \partial_i \partial_j f(x)$.*

A function is monotone (non-decreasing) if and only if for all $i \in [n]$ and $x \in \{0,1\}^n$, $\partial_i f(x) \geq 0$. For a submodular function, $\partial_{i,j} f(x) \leq 0$, by considering the submodularity condition for $x_{i\leftarrow 0,j\leftarrow 0}$, $x_{i\leftarrow 0,j\leftarrow 1}$, $x_{i\leftarrow 1,j\leftarrow 0}$, and $x_{i\leftarrow 1,j\leftarrow 1}$.

**Absolute error vs. error relative to norm:** In our results, we typically assume that the values of $f(x)$ are in a bounded interval $[0,1]$, and our goal is to learn $f$ with an additive error of $\epsilon$. Some prior work considered an error relative to the norm of $f$, for example at most $\epsilon\|f\|_2$ [CKKL12]. In fact, it is known that for a non-negative submodular, XOS or self-bounding function $f$, $\|f\|_2 = \Omega(\|f\|_\infty)$ [Fei06], [FMV07], [FKV14] and hence this does not make much difference. If we scale $f(x)$ by $\frac{1}{4\|f\|_2}$, we obtain a function with values in $[0,1]$ and learning the original function within an additive error of $\epsilon\|f\|_2$ is equivalent to learning the scaled function within an error of $\epsilon/4$.

**Fourier Analysis:** We rely on the standard Fourier transform representation of real-valued functions over $\{0,1\}^n$ as linear combinations of parity functions. For $S \subseteq [n]$, the parity function $\chi_S : \{0,1\}^n \to \{-1,1\}$ is defined by $\chi_S(x) = (-1)^{\sum_{i\in S} x_i}$. The Fourier expansion of $f$ is given by $f(x) = \sum_{S\subseteq[n]} \hat{f}(S)\chi_S(x)$. The *Fourier degree* of $f$ is the largest

$|S|$ such that $\hat{f}(S) \neq 0$. Note that Fourier degree of $f$ is exactly the polynomial degree of $f$ when viewed over $\{-1,1\}^n$ instead of $\{0,1\}^n$ and therefore it is also equal to the polynomial degree of $f$ over $\{0,1\}^n$.

For degree $d$, let $W^d(f) = \sum_{S \subseteq [n],\ |S|=d} (\hat{f}(S))^2$ and $W^{>d}(f) = \sum_{i>d} W^i(f)$. For any function $f$, Parseval's identity states that $\|f\|_2^2 = \sum_{S \subseteq [n]} (\hat{f}(S))^2$. This implies that the degree $d$ polynomial closest in $\ell_2$ distance to $f$ is precisely $p(x) = \sum_{S \subseteq [n],\ |S| \leq d} \hat{f}(S)\chi_S(x)$ and $\|f - p\|_2 = \sqrt{W^{>d}(f)}$. In other words, $\deg_\epsilon^{\ell_2}(f) = d$ if and only if $d$ is the smallest such that $W^{>d}(f) \leq \epsilon^2$.

Observe that: $\partial_i f(x) = -2 \sum_{S \ni i} \hat{f}(S)\chi_{S\setminus\{i\}}(x)$, and $\partial_{i,j} f(x) = 4 \sum_{S \ni i,j} \hat{f}(S)\chi_{S\setminus\{i,j\}}(x)$.

## III. Degree $O(1/\epsilon)$ approximation for XOS functions

In this section, we consider XOS functions $f : \{0,1\}^n \to \mathbb{R}_+,\ f(x) = \max_{c \in C} \sum_{i=1}^n w_{ci} x_i$, where, $w_{ci} \geq 0$ are nonnegative weights. We call each $c \in C$ a *clause* of the XOS function.

We recall that XOS functions, and more generally self-bounding functions, satisfy the following inequality for each $x \in \{0,1\}^n$: $\sum_{i=1}^n (f(x) - f(x \oplus e_i))_+ \leq f(x)$. In particular, for XOS functions (which are monotone), this can be written as

$$\sum_{i: x_i = 1} \partial_i f(x) \leq f(x). \tag{1}$$

This leads to a bound of the form $\sum_S |S| \hat{f}^2(S) = O(\|f\|_2^2)$, which implies that degree $O(1/\epsilon^2)$ is sufficient to approximate XOS functions within $\ell_2$-error $\epsilon$. Here, we aim to improve the degree bound from $O(1/\epsilon^2)$ to $O(1/\epsilon)$. For this purpose, we seek a "second-degree variant" of inequality (1), using the second-degree derivatives

$$\partial_{ij} f(x) = f(x_{i\leftarrow 1, j\leftarrow 1}) - f(x_{i\leftarrow 1, j\leftarrow 0}) - f(x_{i\leftarrow 0, j\leftarrow 1}) + f(x_{i\leftarrow 0, j\leftarrow 0}).$$

(For $i = j$, we define $\partial_{ii} f(x) = 0$.) Our plan is to use these expressions as follows.

**Lemma 5.** *For any function $f : \{0,1\}^n \to \mathbb{R}_+$ and any $1 \leq k \leq n$,*

$$\sum_{|S|>k} \hat{f}^2(S) \leq \frac{1}{16k^2} \sum_{i,j=1}^n \|\partial_{ij} f\|_2^2.$$

*Proof:* For every pair $i \neq j \in [n]$, we have $\|\partial_{ij} f\|_2^2 = 16 \sum_{S: i,j \in S} \hat{f}^2(S)$. Summing up over all choices of $i \neq j$, each set $S$ appears $|S|(|S|-1)$ times:

$$\sum_{i \neq j} \|\partial_{ij} f\|_2^2 = 16 \sum_{S \subseteq [n]} |S|(|S|-1) \hat{f}^2(S).$$

Therefore, we obtain $\sum_{i,j=1}^n \|\partial_{ij} f\|_2^2 \geq 16 \sum_{S \subseteq [n]} |S|(|S|-1) \hat{f}^2(S) \geq 16k^2 \sum_{|S|>k} \hat{f}^2(S)$. ∎

Our goal in the following is to bound the expression $\sum_{i,j=1}^n \|\partial_{ij} f\|_2^2$. First we prove the following.

**Lemma 6.** *For an XOS function $f : \{0,1\}^n \to \mathbb{R}_+$ and any $x \in \{0,1\}^n$,*

$$\sum_{i,j: x_i = x_j = 1} (\partial_{ij} f(x))^2 \leq 5(f(x))^2. \tag{2}$$

*Proof:* Let $S$ denote the set of coordinates such that $x_i = 1$. Let $c \in C$ be a clause that achieves the maximum, defining $f(x) = \sum_{j \in S} w_{cj}$ (if there are multiple such clauses, fix one arbitrarily). Fix any $i \in S$ and define $c'_i \in C$ to be a clause achieving the maximum that defines $f(x_{i\leftarrow 0}) = \sum_{j \in S\setminus\{i\}} w_{c'_i j}$. Fix another $j \in S$. We claim the following bounds:

$$-\min\{w_{ci} + w_{cj}, w_{c'_i j}\} \leq \partial_{ij} f(x) \leq \min\{w_{ci}, w_{cj}\}. \tag{3}$$

First, assume that $\partial_{ij} f(x) > 0$. Since $f$ is monotone, we have $\partial_{ij} f(x) \leq \min\{\partial_i f(x), \partial_j f(x)\}$. Since $c$ is the clause defining $f(x)$, $f(x)$ cannot decrease by more than $w_{ci}$ when flipping $x_i$ from 1 to 0. Similarly, $f(x)$ cannot decrease by more than $w_{cj}$ when flipping $x_j$ from 1 to 0. Therefore, $\partial_{ij} f(x) \leq \min\{w_{ci}, w_{cj}\}$.

Second, assume that $\partial_{ij} f(x) < 0$. Here we have $\partial_{ij} f(x) \geq -\min\{\partial_i f(x_{j\leftarrow 0}), \partial_j f(x_{i\leftarrow 0})\}$. Recall that after flipping $x_i$ from 1 to 0, $c'_i$ is a maximizing clause, and therefore $\partial_j f(x_{i\leftarrow 0}) = f(x_{i\leftarrow 0, j\leftarrow 1}) - f(x_{i\leftarrow 0, j\leftarrow 0}) \leq w_{c'_i j}$. To bound $\partial_i f(x_{j\leftarrow 0})$, we use the following (by monotonicity): $\partial_i f(x_{j\leftarrow 0}) = f(x_{i\leftarrow 1, j\leftarrow 0}) - f(x_{i\leftarrow 0, j\leftarrow 0}) \leq f(x_{i\leftarrow 1, j\leftarrow 1}) - f(x_{i\leftarrow 0, j\leftarrow 0}) \leq w_{ci} + w_{cj}$, using the fact that $c$ is a maximizing clause for $f(x)$. (We remark that although this seems like a weak bound, it could be actually tight.) This proves (3).

Next, we sum up over all pairs of coordinates $i,j \in S$. Note that $c \in C$ is fixed before choosing $i,j$, and we can assume for convenience that the coordinates are ordered so that $i \le j$ implies $w_{ci} \le w_{cj}$. We have

$$\begin{aligned}
\sum_{i,j\in S} (\partial_{ij} f(x))^2 &= \sum_{i,j\in S:\partial_{ij} f(x)>0} (\partial_{ij} f(x))^2 + 2 \sum_{i,j\in S: i>j, \partial_{ij} f(x)<0} (\partial_{ij} f(x))^2 \\
&\le \sum_{i,j\in S} w_{ci} w_{cj} + 2 \sum_{i,j\in S, i>j} (w_{ci} + w_{cj}) w_{c'_i j} \\
&\le \sum_{i,j\in S} w_{ci} w_{cj} + 4 \sum_{i,j\in S, i>j} w_{ci} w_{c'_i j} \\
&= \left(\sum_{i\in S} w_{ci}\right)^2 + 4 \sum_{i\in S} \left( w_{ci} \sum_{j\in S, j<i} w_{c'_i j} \right) \\
&\le (f(x))^2 + 4(f(x))^2 = 5(f(x))^2
\end{aligned}$$

since $\sum_{j\in S} w_{c'j} \le f(x)$ for every clause $c' \in C$. ■

**Lemma 7.** *For any XOS function $f : \{0,1\}^n \to \mathbb{R}_+$,*

$$\sum_{i,j=1}^{n} \|\partial_{ij} f\|_2^2 \le 20 \|f\|_2^2.$$

*Proof:* Since all norms here are over the uniform distribution, we have $\|\partial_{ij} f\|_2^2 = \frac{1}{2^n} \sum_{x\in\{0,1\}^n} (\partial_{ij} f(x))^2$. Note that Lemma 6 counts only the contributions from points such that $x_i = x_j = 1$. However, $\partial_{ij} f(x)$ does not depend on the values of $x_i$ and $x_j$. Therefore, we can write equivalently

$$\|\partial_{ij} f\|_2^2 = \frac{4}{2^n} \sum_{x\in\{0,1\}^n : x_i = x_j = 1} (\partial_{ij} f(x))^2.$$

Summing up over all $i,j$ and switching the sums, we get

$$\sum_{i,j=1}^{n} \|\partial_{ij} f\|_2^2 = \frac{4}{2^n} \sum_{i,j=1}^{n} \sum_{x\in\{0,1\}^n : x_i = x_j = 1} (\partial_{ij} f(x))^2 = \frac{4}{2^n} \sum_{x\in\{0,1\}^n} \sum_{i,j: x_i = x_j = 1} (\partial_{ij} f(x))^2.$$

Now, we can apply Lemma 6 to conclude that $\sum_{i,j=1}^{n} \|\partial_{ij} f\|_2^2 \le \frac{4}{2^n} \sum_{x\in\{0,1\}^n} 5(f(x))^2 = 20 \|f\|_2^2$. ■

We can conclude as follows.

**Corollary 8.** *For any XOS function $f : \{0,1\}^n \to [0,1]$, there is a polynomial $p$ of degree $O(1/\epsilon)$ such that $\|f - p\|_2 \le \epsilon$.*

*Proof:* By Lemma 7, we have $\sum_{i,j=1}^{n} \|\partial_{ij} f\|_2^2 \le 20$, since $\|f\|_2 \le 1$ here. Therefore, applying Lemma 5, $\sum_{|S|>k} \hat{f}^2(S) \le \frac{5}{4k^2}$. We choose $k = \sqrt{5}/(2\epsilon)$, which ensures that $\sum_{|S|>k} \hat{f}^2(S) \le \epsilon^2$ and therefore the polynomial consisting of all terms up to degree $k$ approximates $f$ within $\ell_2$-error $\epsilon$. ■

## IV. Degree $\tilde{O}(1/\epsilon^{4/5})$ approximation for submodular functions

In this section, we show that the $O(1/\epsilon)$ degree approximation for XOS functions can be improved to $\tilde{O}(1/\epsilon^{4/5})$ for submodular functions. Interestingly, $1/\epsilon^{4/5}$ turns out to be the right answer for submodular functions (ignoring logarithmic factors).

We build on the technique of bounding $\sum_{i,j} \|\partial_{ij} f\|_2^2$, which in the case of submodular functions seems particularly appropriate since we know that $\partial_{ij} f(x) \le 0$ for every $x \in \{0,1\}^n$, which simplifies certain expressions. However, Lemma 7 itself cannot be improved to a sub-constant bound — it is easy to see that $\sum_{i,j} \|\partial_{ij} f\|_2^2$ could be at least $\|f\|_2^2$ for a submodular function (e.g., $f(x) = 1 - (1-x_1)(1-x_2)$). However, as we show below this can happen only when some variables have a very large influence. Our goal is to handle such variables separately and prove that under the assumption of low influences, the quantity $\sum_{i,j} \|\partial_{ij} f\|_2^2$ cannot be large.

Once we can control the influences of individual variables (for now imagine that we can control $\|\partial_i f\|_2$), we use the following way of bounding the sum of second partial derivatives.

**Lemma 9.** *For any submodular function $f : \{0,1\}^n \to \mathbb{R}$, any coordinate $i$ and a subset of coordinates $A$,*

$$\sum_{j \in A} \|\partial_{ij} f\|_1 \leq 2\sqrt{|A|}\ \|\partial_i f\|_2.$$

Note the improvement from $2|A|\|\partial_i f\|_1$ (which is trivial) to $2\sqrt{|A|}\|\partial_i f\|_2$ on the right-hand-side.

*Proof:* Since $f$ is submodular, we have $\partial_{ij} f(x) \leq 0$, and

$$\sum_{j \in A} \|\partial_{ij} f\|_1 = \sum_{j \in A} \mathbf{E}_{x \sim \mathcal{U}}[-\partial_{ij} f(x)] = 2 \sum_{j \in A} \mathbf{E}_{x \sim \mathcal{U}}[(-1)^{x_j} \partial_i f(x)] = 2 \cdot \mathbf{E}_{x \sim \mathcal{U}}[\partial_i f(x) g(x)] \leq 2\|\partial_i f\|_2 \|g\|_2$$

where $g(x) = \sum_{j \in A} (-1)^{x_j}$ and we used the Cauchy-Schwartz inequality at the end. It is easy to check that $\|g\|_2 = \sqrt{|A|}$ which proves the lemma. ■

First, let us sketch how this argument leads to an $O(1/\epsilon^{4/5})$ bound in the case of *totally symmetric* submodular functions, to illustrate some of the ideas employed in the general case.

*Totally symmetric submodular functions:* Let us assume that $f : \{0,1\}^n \to [0,1]$ is totally symmetric in the sense that $f(x)$ depends only on $\sum_{i=1}^n x_i$. Note that such a function is simply a concave function of $\sum_{i=1}^n x_i$. First, we observe that the influences of individual variables in this case cannot be too large.

**Lemma 10.** *For any totally symmetric submodular function $f : \{0,1\}^n \to [0,1]$ and $x \in \{0,1\}^n$ such that $\frac{n}{3} \leq \sum_{i=1}^n x_i \leq \frac{2n}{3}$, we have $|\partial_i f(x)| \leq \frac{3}{n}$ for all $i \in [n]$.*

*Proof:* Assume that $\partial_i f(x) > \frac{3}{n}$ (the opposite case is similar). Since the function is totally symmetric, we actually have $\partial_j f(x) > \frac{3}{n}$ for every $j \in [n]$. Also, $\sum_{i=1}^n x_i \geq \frac{n}{3}$. By submodularity, $f(x) - f(0) \geq \sum_{j : x_j = 1} \partial_j f(x) > \frac{3}{n} \sum_{j=1}^n x_j \geq 1$. This contradicts the fact that the values of $f(x)$ are in $[0,1]$. ■

To simplify the analysis, let us assume that in fact, $|\partial_i f(x)| = O(\frac{1}{n})$ for all $i \in [n]$ and all $x \in \{0,1\}^n$. This can be accomplished by modifying the function in the regions where $\sum_{i=1}^n x_i < \frac{n}{3}$ or $> \frac{2n}{3}$ in such a way that $\partial_i f(x)$ is constant in each region. For example, if $t$ is maximum such that $\partial_i f(x') > \frac{3}{n}$ for $\sum_{i=1}^n x'_i = t$, let $f(x') = F$ for this point $x'$ (and we know that $t = \sum_{i=1}^n x'_i < \frac{n}{3}$). We can set $f(x) = F - \frac{3}{n}(t - \sum_{i=1}^n x_i)$ whenever $\sum_{i=1}^n < t$. Similarly, we adjust the function for $\sum_{i=1}^n x_i > \frac{2n}{3}$. These are sets of small measure (under the uniform distribution) and so any approximation of the modified function also works well for the original function. In the following, we assume that $|\partial_i f(x)| = O(\frac{1}{n})$ everywhere. Now we can show the following bound.

**Lemma 11.** *If $|\partial_i f(x)| = O(\frac{1}{n})$ for all $i \in [n]$ and $x \in \{0,1\}^n$, then $\sum_{i,j=1}^n \|\partial_{ij} f\|_2^2 = O\left(\frac{1}{\sqrt{n}}\right)$.*

*Proof:* Note that the assumption on partial derivatives also implies that $|\partial_{ij} f(x)| = O(\frac{1}{n})$. We estimate $\sum_{i,j=1}^n \|\partial_{ij} f\|_2^2$ as follows:

$$\sum_{i,j=1}^n \|\partial_{ij} f\|_2^2 = O\left(\frac{1}{n}\right) \sum_{i,j=1}^n \|\partial_{ij} f\|_1 = O\left(\frac{1}{\sqrt{n}}\right) \sum_{i=1}^n \|\partial_i f\|_2$$

using Lemma 9 with $A = [n]$. Since we assume that $|\partial_i f(x)| = O(\frac{1}{n})$, it follows that $\|\partial_i f\|_2 = O(\frac{1}{n})$ for all $i \in [n]$ which proves the lemma. ■

Now we can apply the method of bounding the Fourier tail above a certain level using Lemma 5:

$$\sum_{|S| > k} \hat{f}^2(S) \leq \frac{16}{k^2} \sum_{i,j=1}^n \|\partial_{ij} f\|_2^2 = O\left(\frac{1}{k^2 \sqrt{n}}\right).$$

We choose $k = 1/(\epsilon n^{1/4})$ in order to make the Fourier tail bounded by $O(\epsilon^2)$ as it should be. Finally, note that if $n \leq 1/\epsilon^{4/5}$, we can take trivially a polynomial of degree $n$. Therefore, the non-trivial case is when $n > 1/\epsilon^{4/5}$ and then we have $k = 1/(\epsilon n^{1/4}) \leq 1/\epsilon^{4/5}$. This proves that degree $O(1/\epsilon^{4/5})$ is sufficient for totally symmetric submodular functions.

*General submodular functions:* Let us turn now to the case of general submodular functions. The main complication here is that some variables can have large influences and we need to handle those separately. The main technical lemma here is that there cannot be too many variables of large influence, measured in a suitable way. The most technical part of the proof is to prove that there cannot be too many variables of large influence, and the influences decay relatively fast as we consider more variables. We also have to define "influence" in a suitable way. We denote by $\mu_p$ a product distribution on $\{0,1\}^n$ such that $\Pr_{x \sim \mu_p}[x_i = 1] = p$ for each $i \in [n]$. We prove the following.

**Lemma 12.** *Let $f : \{0,1\}^n \to [0,1]$ be a submodular function, $0 < \delta, \epsilon < 1$, and let*

$$J(\epsilon, \delta) = \{i \in [n] : \Pr_{x \sim \mu_{1/2}}[|\partial_i f(x)| \geq \epsilon] \geq \delta\}.$$

*Then $|J(\epsilon, \delta)| = O(\frac{1}{\epsilon} \log \frac{1}{\delta})$.*

We prove this using the "boosting lemma" of [GV06] (which was already used for the purpose of approximating submodular functions by juntas in [FV13]).

**Boosting Lemma.** *Let $\mathcal{F} \subseteq \{0,1\}^X$ be down-monotone (if $x \in \mathcal{F}$ and $y \leq x$ coordinate-wise, then $y \in \mathcal{F}$). For $p \in (0,1)$, define $\sigma_p = \Pr_{x \sim \mu_p}[x \in \mathcal{F}]$. Then $\sigma_p = (1-p)^{\phi(p)}$ where $\phi(p)$ is a non-decreasing function for $p \in (0,1)$.*

*Proof of Lemma 12:* Let

- $J^+(\epsilon, \delta) = \{i \in [n] : \Pr_{x \sim \mu_{1/2}}[\partial_i f(x) \geq \epsilon] \geq \delta/2\}$.
- $J^-(\epsilon, \delta) = \{i \in [n] : \Pr_{x \sim \mu_{1/2}}[\partial_i f(x) \leq -\epsilon] \geq \delta/2\}$.

We have $J(\epsilon, \delta) \subseteq J^+(\epsilon, \delta) \cup J^-(\epsilon, \delta)$. Hence it is enough to bound $|J^+(\epsilon, \delta)|$; the same bound on $|J^-(\epsilon, \delta)|$ follows by considering the function $\bar{f}(x) = f(\mathbf{1} - x)$.

For each $j \in [n]$, define

$$\mathcal{F}_j^+ = \{x \in \{0,1\}^n : \partial_j f(x) \geq \epsilon\}.$$

By submodularity, this set is down-monotone. By assumption, we have $\Pr_{x \sim \mu_{1/2}}[x \in \mathcal{F}_j^+] \geq \delta/2$ for $j \in J^+(\epsilon, \delta)$. Using the terminology of the boosting lemma, we have $\sigma_{1/2} = (1/2)^{\phi(1/2)}$ where $\phi(1/2) \leq \log_2(2/\delta)$. We define $q = 1 - (1/2)^{1/\log_2(2/\delta)} \leq 1/2$. By the boosting lemma [GV06], we have

$$\Pr_{x \sim \mu_q}[x \in \mathcal{F}_j^+] = (1-q)^{\phi(q)} \geq (1-q)^{\log_2(2/\delta)} = \frac{1}{2}$$

for each $j \in J^+(\epsilon, \delta)$. We also have $\Pr_{x \sim \mu_q}[x_j = 1] = q$. Note that $x_j = 1$ and $x \in \mathcal{F}_j^+$ are independent events, since $x \in \mathcal{F}_j^+$ depends only on $\partial_j f(x)$ and this is independent of $x_j$. Therefore,

$$\Pr_{x \sim \mu_q}[x_j = 1 \ \& \ x \in \mathcal{F}_j^+] \geq \frac{q}{2}$$

for each $j \in J^+(\epsilon, \delta)$. Let $L(x) = \{j : x_j = 1 \ \& \ x \in \mathcal{F}_j^+\}$. We have

$$\mathbf{E}_{x \sim \mu_q}[|L(x)|] \geq \mathbf{E}_{x \sim \mu_q}[|\{j \in J^+(\epsilon, \delta) : x_j = 1 \ \& \ x \in \mathcal{F}_j^+\}|] \geq \frac{q}{2}|J^+(\epsilon, \delta)|.$$

On the other hand, denoting by $\mathbf{1}_S$ the indicator vector of $S$, for each $j \in L(x)$, we have $\partial_j f(\mathbf{1}_{L(x)}) \geq \epsilon$ and therefore

$$\epsilon |L(x)| \leq \sum_{j \in L(x)} \partial_j f(\mathbf{1}_{L(x)}) \leq f(\mathbf{1}_{L(x)}) \leq 1$$

where we used submodularity in the second inequality. This means that $|L(x)| \leq 1/\epsilon$ with probability 1. Therefore, we have $|J^+(\epsilon, \delta)| \leq \frac{2}{\epsilon q}$. Recall that $q = 1 - (1/2)^{1/\log_2(2/\delta)} \geq \frac{1}{2\log_2(2/\delta)}$ (using $\delta < 1$) which means $|J^+(\epsilon, \delta)| \leq \frac{4}{\epsilon} \log_2 \frac{2}{\delta}$. ∎

We use Lemma 12 for two purposes. First, it allows us to take out a small set of variables $L$ whose derivatives can be large with large probability. Conditioned on these variables, we get an "almost $\epsilon$-Lipschitz" function, for which using Lemma 12 again allows us to prove an improved bound on $\sum_{i,j \notin L} \|\partial_{ij} f\|_2^2$.

We introduce the following notation (the "threshold norm"). In the following, all probabilities and expectations are over the uniform distribution ($x \sim \mathcal{U}$).

**Definition 13.** *For a function $f : \{0,1\}^n \to \mathbb{R}$, we define*

$$\|f\|_T = \sup\{\alpha : \Pr_x[|f(x)| \geq \alpha] \geq \alpha^3\}.$$

We remark that $\|f\|_T$ is not really a norm — it is not linear under scalar multiplication. In fact $\|f\|_T$ is never more than 1. The choice of $\alpha^3$ is somewhat arbitrary here. The notation $\|f\|_T$ is convenient for our proof but in general we do not attribute any significance to it. Lemma 12 (with $\delta = \epsilon^3$) implies the following.

**Corollary 14.** *For a submodular function $f : \{0,1\}^n \to [0,1]$, the number of coordinates with $\|\partial_i f\|_T \geq \epsilon$ is at most $O(\frac{1}{\epsilon} \log \frac{1}{\epsilon})$.*

We also have the following useful property (which we apply to $h = \partial_i f$).

**Lemma 15.** *For any $h : \{0,1\}^n \to [-1,1]$, $\|h\|_2 \leq \sqrt{2}\|h\|_T$.*

*Proof:* Suppose that $\|h\|_T = \eta$ and note that $0 \leq \eta \leq 1$. For every $\alpha > \eta$, we have by definition $\Pr[\|h(x)\| \geq \alpha] < \alpha^3]$. Consequently

$$\|h\|_2^2 = \mathbf{E}[(h(x))^2] \leq \alpha^2 \cdot \Pr[|h(x)| \leq \alpha] + 1 \cdot \Pr[|h(x)| > \alpha] \leq \alpha^2 + \alpha^3.$$

Since this holds for every $\alpha > \eta$, we also have $\|h\|_2^2 \leq \eta^2 + \eta^3 \leq 2\eta^2$. ■

The following is our main bound on the quantity $\sum_{i,j} \|\partial_{ij} f\|_2^2$.

**Lemma 16.** *Let $f : \{0,1\}^n \to [0,1]$ be a submodular function such that $\|\partial_i f\|_T \leq \alpha$ for all $i \in S$. Then*

$$\sum_{i,j \in S} \|\partial_{ij} f\|_2^2 = O\left(\sqrt{\alpha} \log^{3/2} \frac{1}{\alpha}\right).$$

*Proof:* First, note that coordinates $i \in S$ such that $\|\partial_i f\|_T = 0$ do not contribute anything to the sum $\sum_{i,j \in S} \|\partial_{ij} f\|_2^2$. This is because if $\|\partial_i f\|_T = 0$ then $\partial_i f(x) = 0$ for all $x \in \{0,1\}^n$ and hence also $\partial_{ij} f(x) = 0$ for every other coordinate $j \in S$. Therefore, we can assume that $\|\partial_i f\|_T > 0$ for all $i \in S$.

Let us partition the coordinates as follows. For each $\ell \geq 0$, let

$$B_k = \{i \in S : \|\partial_i f\|_T > 2^{-k}\alpha\}.$$

Note that by assumption, $B_0 = \emptyset$, the sets $B_k$ form a chain and each $i \in S$ belongs to $B_k$ for large enough $k$. By Corollary 14, the sets $B_k$ are bounded in size:

$$|B_k| = O\left(\frac{1}{2^{-k}\alpha} \log \frac{1}{2^{-k}\alpha}\right) = O\left(\frac{2^k}{\alpha} \log \frac{2^k}{\alpha}\right).$$

We define $A_k = B_{k+1} \setminus B_k$; clearly, each coordinate belongs to exactly one set $A_k, k \geq 0$, and we have $\|\partial_i f\|_T \leq 2^{-k}\alpha$ for each $i \in A_k$.

We estimate $\sum_{i,j \in S} \|\partial_{ij} f\|_2^2$ as follows. We can write

$$|\partial_{ij} f(x)| = |\partial_i f(x_{j \leftarrow 1}) - \partial_i f(x_{j \leftarrow 0})| \leq |\partial_i f(x_{j \leftarrow 1})| + |\partial_i f(x_{j \leftarrow 0})|.$$

Therefore,

$$\|\partial_{ij} f\|_2^2 = \mathbf{E}_x[|\partial_{ij} f(x)|^2] \leq \mathbf{E}_x[(|\partial_i f(x_{j \leftarrow 1})| + |\partial_i f(x_{j \leftarrow 0})|) \cdot |\partial_{ij} f(x)|].$$

Assuming that $i \in A_k$, we know that $\|\partial_i f\|_T \leq 2^{-k}\alpha$, and hence $\Pr_x[|\partial_i f(x)| \geq 2^{1-k}\alpha] \leq 2^{-3k}\alpha^3$. Therefore, we also have $\Pr_x[|\partial_i f(x_{i \leftarrow 1})| + |\partial_i f(x_{i \leftarrow 0})| \geq 2^{2-k}\alpha] \leq 2^{1-3k}\alpha^3$. Hence for $i \in A_k$ we can estimate

$$\|\partial_{ij} f\|_2^2 \leq \mathbf{E}_x[(|\partial_i f(x_{j \leftarrow 1})| + |\partial_i f(x_{j \leftarrow 0})|) \cdot |\partial_{ij} f(x)|] \leq 2^{2-k}\alpha \cdot \mathbf{E}_x[|\partial_{ij} f(x)|] + 2^{2-3k}\alpha^3$$

using the trivial bound that $|\partial_{ij} f(x)| \leq 2$ in the case where $|\partial_i f(x_{i \leftarrow 1})| + |\partial_i f(x_{i \leftarrow 0})|$ is large.

Overall, we obtain

$$\begin{aligned}
\sum_{i,j \in S} \|\partial_{ij} f\|_2^2 &\leq 2 \sum_{0 \leq \ell \leq k} \sum_{i \in A_k} \sum_{j \in A_\ell} \|\partial_{ij} f\|_2^2 \\
&\leq 2 \sum_{0 \leq \ell \leq k} \sum_{i \in A_k} \sum_{j \in A_\ell} \left(2^{2-k}\alpha \cdot \mathbf{E}_x[|\partial_{ij} f(x)|] + 2^{2-3k}\alpha^3\right) \\
&= \sum_{k \geq 0} 2^{3-k}\alpha \sum_{i \in A_k} \sum_{\ell=0}^{k} \sum_{j \in A_\ell} \|\partial_{ij} f\|_1 + \sum_{0 \leq \ell \leq k} 2^{3-3k}\alpha^3 |A_k| \cdot |A_\ell|.
\end{aligned}$$

Here we use the bounds $|A_k| \leq |B_{k+1}| = O(\frac{2^k}{\alpha} \log \frac{2^k}{\alpha})$ to estimate the second term. We get (up to constant factors) $\sum_{0 \leq \ell \leq k} 2^{-3k}\alpha \cdot 2^{k+\ell}(k + \log \frac{1}{\alpha})(\ell + \log \frac{1}{\alpha}) \leq \sum_{k \geq 0} k^3 2^{-k}\alpha \log^2 \frac{1}{\alpha} = O(\alpha \log^2 \frac{1}{\alpha})$. Hence, we get

$$\sum_{i,j \in S} \|\partial_{ij} f\|_2^2 \leq \sum_{k \geq 0} 2^{3-k}\alpha \sum_{i \in A_k} \sum_{\ell=0}^{k} \sum_{j \in A_\ell} \|\partial_{ij} f\|_1 + O\left(\alpha \log^2 \frac{1}{\alpha}\right). \tag{4}$$

Here we use Lemma 9 to estimate $\sum_{\ell=0}^{k} \sum_{j\in A_\ell} \|\partial_{ij} f\|_1$. Recall that the $A_\ell$'s are disjoint and $\bigcup_{\ell=0}^{k} A_\ell = B_{k+1}$. By Lemma 9,

$$\sum_{\ell=0}^{k} \sum_{j\in A_\ell} \|\partial_{ij} f\|_1 = \sum_{j\in B_{k+1}} \|\partial_{ij} f\|_1 \le 2\sqrt{|B_{k+1}|}\|\partial_i f\|_2 = O\left(\frac{2^{k/2}}{\sqrt{\alpha}} \log^{1/2} \frac{2^k}{\alpha}\right) \cdot \|\partial_i f\|_2.$$

Assuming $i \in A_k$, Lemma 15 says $\|\partial_i f\|_2 \le \sqrt{2}\|\partial_i f\|_T \le \sqrt{2}\frac{\alpha}{2^k}$. Also, $|A_k| = O(\frac{2^k}{\alpha} \log \frac{2^k}{\alpha})$, so we get

$$\sum_{i\in A_k} \sum_{\ell=0}^{k} \sum_{j\in A_\ell} \|\partial_{ij} f\|_1 = O\left(\frac{2^{k/2}}{\sqrt{\alpha}} \log^{1/2} \frac{2^k}{\alpha}\right) \cdot \sum_{i\in A_k} \|\partial_i f\|_2 = O\left(\frac{2^{k/2}}{\sqrt{\alpha}} \log^{3/2} \frac{2^k}{\alpha}\right).$$

Continuing the computation from equation (4), we have

$$\begin{aligned}\sum_{i,j\in S} \|\partial_{ij} f\|_2^2 &\le \sum_{k\ge 0} 2^{3-k}\alpha \sum_{i\in A_k} \sum_{\ell=0}^{k} \sum_{j\in A_\ell} \|\partial_{ij} f\|_1 + O\left(\alpha \log^2 \frac{1}{\alpha}\right) \\ &= O\left(\sum_{k\ge 0} \frac{\sqrt{\alpha}}{2^{k/2}} \log^{3/2} \frac{2^k}{\alpha}\right) + O\left(\alpha \log^2 \frac{1}{\alpha}\right) \\ &= O\left(\sqrt{\alpha} \log^{3/2} \frac{1}{\alpha}\right).\end{aligned}$$

■

**Theorem 17.** *For any submodular function $f : \{0,1\}^n \to [0,1]$, there is a polynomial of degree $O(\frac{1}{\epsilon^{4/5}} \log \frac{1}{\epsilon})$ such that $\|f - p\|_2 \le \epsilon$.*

*Proof:* Let $\alpha = \epsilon^{4/5}$. Let $L$ be the set of variables $i \in [n]$ such that $\|\partial_i f\|_T > \alpha$. By Corollary 14, the number of such variables is $|L| = O(\frac{1}{\alpha} \log \frac{1}{\alpha}) = O(\frac{1}{\epsilon^{4/5}} \log \frac{1}{\epsilon})$. By Lemma 16 (for $S = [n] \setminus L$), we have $\sum_{i,j\notin L} \|\partial_{ij} f\|_2^2 = O(\sqrt{\alpha} \log^{3/2} \frac{1}{\alpha}) = O(\epsilon^{2/5} \log^{3/2} \frac{1}{\epsilon})$. On the other hand (recalling that $\|\partial_{ii} f\|_2 = 0$ and $\|\partial_{ij} f\|_2^2 = 16 \sum_{S\supseteq\{i,j\}} \hat{f}^2(S)$ for $i \ne j$),

$$\sum_{i,j\notin L} \|\partial_{ij} f\|_2^2 = 16 \sum_{S:|S\setminus L|\ge 2} |S\setminus L|(|S\setminus L|-1)\hat{f}^2(S) \ge 16 \sum_{S:|S\setminus L|>k} k^2 \hat{f}^2(S).$$

We set $k = \frac{1}{\epsilon^{4/5}} \log \frac{1}{\epsilon}$ and obtain

$$\sum_{S:|S\setminus L|>k} \hat{f}^2(S) \le \frac{1}{16k^2} \sum_{i,j\notin L} \|\partial_{ij} f\|_2^2 = \frac{\epsilon^{8/5}}{\log^2 \frac{1}{\epsilon}} \cdot O\left(\epsilon^{2/5} \log^{3/2} \frac{1}{\epsilon}\right) = O\left(\epsilon^2 \log^{-1/2} \frac{1}{\epsilon}\right).$$

For $\epsilon > 0$ sufficiently small, this is less than $\epsilon^2$. Therefore, the polynomial

$$p(x) = \sum_{S:|S\setminus L|\le k} \hat{f}^2(S)\chi_S(x)$$

satisfies

$$\|f - p\|_2^2 = \sum_{S:|S\setminus L|>k} \hat{f}^2(S) < \epsilon^2$$

and has degree $|L| + k = O(\frac{1}{\epsilon^{4/5}} \log \frac{1}{\epsilon})$. ■

## V. APPLICATIONS

### A. Approximation of XOS functions by juntas

We use several notions of *influence* of a variable on a real-valued function which are based on the standard notion of influence for Boolean functions (e.g. [BOL85], [KKL88]).

**Definition 18** (Influences). *For a real-valued $f : \{0,1\}^n \to \mathbb{R}$, $i \in [n]$, and $\kappa \ge 0$ we define the $\ell_\kappa^\kappa$-*influence *of variable $i$ as $\mathsf{Inf}_i^\kappa(f) = \|\frac{1}{2}\partial_i f\|_\kappa^\kappa = \mathbf{E}[|\frac{1}{2}\partial_i f|^\kappa]$. We define $\mathsf{Inf}^\kappa(f) = \sum_{i\in[n]} \mathsf{Inf}_i^\kappa(f)$.*

The most commonly used notion of influence for real-valued functions is the $\ell_2^2$-influence which satisfies

$$\mathsf{Inf}_i^2(f) = \left\| \frac{1}{2}\partial_i f \right\|_2^2 = \sum_{S \ni i} \hat{f}^2(S) \ .$$

From here, the total $\ell_2^2$-influence is equal to $\mathsf{Inf}^2(f) = \sum_S |S|\hat{f}^2(S)$. We use the following generalization of Friedgut's theorem [Fri98] from [FV13].

**Theorem 19** ([FV13]). *Let $f : \{0,1\}^n \to \mathbb{R}$ be any function, $\epsilon \in (0,1)$ and $\kappa \in (1,2)$. For $d$ such that $\sum_{|S|>d} \hat{f}(S)^2 \leq \epsilon^2/2$, let*

$$I = \{i \in [n] \mid \mathsf{Inf}_i^\kappa(f) \geq \alpha\} \text{ for}$$
$$\alpha = \left((\kappa-1)^{d-1} \cdot \epsilon^2/(2 \cdot \mathsf{Inf}^\kappa(f))\right)^{\kappa/(2-\kappa)} \ .$$

*Then for the set $\mathcal{I}_d = \{S \subseteq I \mid |S| \leq d\}$ we have $\sum_{S \notin \mathcal{I}_d} \hat{f}(S)^2 \leq \epsilon^2$.*

Finally, to apply this generalization we need a bound on the total influence of any XOS function:

**Lemma 20.** *Let $f : \{0,1\}^n \to \mathbb{R}_+$ be an XOS function. Then $\mathsf{Inf}^1(f) \leq \|f\|_1$. In particular, for an XOS function $f : \{0,1\}^n \to [0,1]$, for all $\kappa \geq 1$, $\mathsf{Inf}^\kappa(f) \leq \mathsf{Inf}^1(f) \leq 1$.*

Combining these results with the degree bounds from Corollary 8 gives the following bound:

**Corollary 21.** *Let $f : \{0,1\}^n \to [0,1]$ be an XOS function and $\epsilon > 0$. There exists a $2^{O(1/\epsilon)}$-junta $p$ of Fourier degree $O(1/\epsilon)$, such that $\|f - p\|_2 \leq \epsilon$. In particular, the spectral $\ell_1$-norm of $p$ is $\|\hat{p}\|_1 = \sum_{S \subseteq [n]} |\hat{p}(S)| = 2^{O(1/\epsilon^2)}$.*

*Proof:* By Corollary 8 we can use $d = O(1/\epsilon)$ in Theorem 19 and we choose $\kappa = 4/3$. Let $\alpha = \left((1/3)^{d-1} \cdot \epsilon^2/(2 \cdot \mathsf{Inf}^{4/3}(f))\right)^2$ be the lower bound on the influence of variables in the junta given in Theorem 19. By Lemma 20, $\mathsf{Inf}^{4/3}(f) \leq 1$. Note that $g = \sum_{S \in \mathcal{I}_d} \hat{f}(S)\chi_S$ is a function of Fourier degree $d$ that depends only on variables in $I$. Further, $\|f - g\|_2^2 \leq \epsilon^2$ and the set $I$ has size at most

$$|I| \leq \mathsf{Inf}^{4/3}(f)/\alpha \leq 3^{2(d-1)} \cdot (2/\epsilon^2)^2 = 2^{O(1/\epsilon)}.$$

■

### *B. Applications to Learning*

*1) Preliminaries: Models of Learning:* We consider two models of learning based on the PAC model [Val84] which assumes that the learner has access to random examples of an unknown function from a known class of functions. Here we only consider learning over the uniform distribution over $\{0,1\}^n$ and hence simplify the definitions for this setting.

**Definition 22** (PAC learning with $\ell_2$-error). *Let $\mathcal{F}$ be a class of real-valued functions on $\{0,1\}^n$. An algorithm $\mathcal{A}$ PAC learns $\mathcal{F}$ with $\ell_2$ error over $\mathcal{U}$, if given $\epsilon > 0$, for every target function $f \in \mathcal{F}$, given access to random independent samples from $\mathcal{U}$ labeled by $f$, with probability at least $2/3$, $\mathcal{A}$ returns a hypothesis $h$ such that $\|f - h\|_2 \leq \epsilon$.*

**Definition 23** (Agnostic learning with $\ell_2$-error). *Let $\mathcal{F}$ be a class of real-valued functions on $\{0,1\}^n$. For any distribution $\mathcal{P}$ over $\{0,1\}^n \times [0,1]$, let opt$(\mathcal{P},\mathcal{F})$ be defined as:*

$$opt(\mathcal{P},\mathcal{F}) = \inf_{f \in \mathcal{F}} \sqrt{\mathbf{E}_{(x,\ell)\sim\mathcal{P}}[(\ell - f(x))^2]}.$$

*An algorithm $\mathcal{A}$, is said to agnostically learn $\mathcal{F}$ with $\ell_2$* excess error *over $\mathcal{U}$ if for every $\epsilon > 0$ and any distribution $\mathcal{P}$ on $\{0,1\}^n \times [0,1]$ such that the marginal of $\mathcal{P}$ on $\{0,1\}^n$ is $\mathcal{U}$, given access to random independent examples drawn from $P$, with probability at least $\frac{2}{3}$, $\mathcal{A}$ outputs a hypothesis $h$ such that*

$$\sqrt{\mathbf{E}_{(x,\ell)\sim\mathcal{P}}[(h(x) - \ell)^2]} \leq opt(\mathcal{P},\mathcal{F}) + \epsilon.$$

We remark that one can also define optimality with respect to labels from a different range. For simplicity we use the $[0,1]$ range since that is also the range of the functions we consider.

For both PAC and agnostic learning we will rely on the fact that polynomials of degree $d$ over $n$ variables can be learned agnostically in time polynomial in $(e \cdot n/d)^d$. For the uniform distribution this follows from the agnostic properties of the low-degree algorithm by Linial *et al.* [LMN93] observed by Kearns *et al.* [KSS94].

**Theorem 24.** *Let $\mathcal{H}_d$ be a class of all degree $d$ polynomials over $n$ variables of $\ell_2$-norm at most 1. Then $\mathcal{H}_d$ can be learned agnostically over $\mathcal{U}$ with excess $\ell_2$ error of $\epsilon$ in time polynomial in $t$ and $1/\epsilon$, where $t = \sum_{i=0}^{d} \binom{n}{i} = O((e \cdot n/d)^d)$.*

We remark that this result also holds over arbitrary distributions and follows from the standard uniform convergence bounds for linear models with squared loss (e.g. [KST08]).

*2) PAC and Agnostic Learning of Submodular and XOS Functions:* The algorithms for PAC learning of submodular and XOS functions in [FV13] are based on two steps:

1) Identify a set of influential variables $J$ such that there exists a submodular (or XOS accordingly) function $h$ that depends only on variables in $J$ and is close to $f$.
2) Use regression over all parity functions of degree at most $d$ on variables in $J$ to find the polynomial that best fits sampled examples.

For XOS functions the first step involves simply choosing variables with large enough Fourier coefficients of degree 1. The analysis of both of these steps in [FV13] is in $\ell_2$ norm and therefore we can directly plug in our new bounds to obtain Theorem 2.

In the case of submodular functions in [FV13] the algorithm that finds the set of influential variables only ensures that there exists a function that depends on variables in $J$ and is close in $\ell_1$ distance to $f$. We therefore provide an analogous result for $\ell_2$. As in [FV13] our algorithm selects all variables that have a large degree-1 or 2 Fourier coefficient (with different values of thresholds). The set of variables it returns is larger but analysis is substantially simpler than that in [FV13].

Before proceeding we will need a few simple definitions. For a real-valued $f$ over $\{0,1\}^n$ and $\epsilon \in [0,1]$ let $s_f(\epsilon)$ denote the smallest $s$ such that there exists an $s$-junta $g$ for which $\|f-g\|_2 \leq \epsilon$. For a set of indices $J \subseteq [n]$ we say that a function is a $J$-junta if it depends only on variables in $J$. For a function $f$ and a set of indices $I$, we define the *projection* of $f$ to $I$ to be the function over $\{0,1\}^n$ whose value depends only on the variables in $I$ and its value at $x_I$ is the expectation of $f$ over all the possible values of variables outside of $I$, namely $f_I(x) = \mathbf{E}_{y\sim\mathcal{U}}[f(x_I, y_{\bar{I}})]$. Observe that an equivalent representation of $f_I$ is as follows:

$$f_I(x) = \sum_{S \subseteq I} \hat{f}(S)\chi_S(x).$$

We will also need the following bound on the number of variables with large degree-1 or degree-2 Fourier coefficient from [FV13] and the property of degree-2 Fourier coefficient of submodular functions from [FKV13].

**Lemma 25** ([FV13])**.** *Let $f : \{0,1\}^n \to [0,1]$ be a submodular function and $\alpha, \beta > 0$. Let*

$$I = \left\{ i \;\middle|\; |\hat{f}(\{i\})| \geq \alpha \right\} \bigcup \left\{ i \;\middle|\; \exists j, |\hat{f}(\{i,j\})| \geq \beta \right\} .$$

*Then $|I| \leq \frac{2}{\min\{\alpha,\beta\}}$.*

**Lemma 26** ([FKV13])**.** *Let $f : \{0,1\}^n \to [0,1]$ be a submodular function and $i, j \in [n]$, $i \neq j$.*

$$|\hat{f}(\{i,j\})| \geq \frac{1}{2} \sum_{S \ni i,j} (\hat{f}(S))^2.$$

We now state the guarantees of our algorithm for finding relevant variables.

**Theorem 27.** *Let $f : \{0,1\}^n \to [0,1]$ be a submodular function. There exists an algorithm, that given any $\epsilon > 0$ and access to random and uniform examples of $f$, with probability at least $5/6$, finds a set of variables $I$ of size at most $32 \cdot (s_f(\epsilon/2))^2/\epsilon^2$ such that there exists a submodular $I$-junta $h$ satisfying $\|f-h\|_2 \leq \epsilon$. The algorithm runs in time $O(n^2 \log(n) \cdot (s_f(\epsilon/2))^4/\epsilon^4)$ and uses $O(\log(n) \cdot (s_f(\epsilon/2))^4/\epsilon^4)$ examples.*

*Proof:* Denote $s = s_f(\epsilon/2)$ and let $J$ be the set of variables such that there exists a $J$-junta $g$ such that $\|f-g\|_2 \leq \epsilon/2$. We can assume without loss of generality that $g = f_J$ since $f_J$ is a submodular $J$-junta and it is the $J$-junta closest to $f$ in $\ell_2$ distance. Let

$$I' = \left\{ i \;\middle|\; |\hat{f}(\{i\})| \geq \frac{\epsilon}{4 \cdot \sqrt{s}} \right\} \bigcup \left\{ i \;\middle|\; \exists j, |\hat{f}(\{i,j\})| \geq \frac{\epsilon^2}{8 \cdot s^2} \right\} .$$

We claim that $\|f_J - f_{I' \cap J}\|_2 \leq \epsilon/2$. Note that by triangle inequality this would imply that $\|f - f_{I' \cap J}\|_2 \leq \epsilon$ meaning that it would suffice to find the variables in $I'$.

Using Lemma 26 and the definition of $I'$, we prove the claim as follows:

$$\begin{aligned}
\|f_J - f_{I'\cap J}\|_2^2 &= \sum_{S\subseteq J,\ S\not\subseteq I'} \hat{f}(S)^2 \\
&= \sum_{i\in J\setminus I'} \hat{f}(\{i\})^2 + \sum_{S\subseteq J,\ S\not\subseteq I',\ |S|\geq 2} \hat{f}(S)^2 \\
&\leq |J\setminus I'|\cdot \frac{\epsilon^2}{16\cdot s} + \sum_{i,j\in J,\ \{i,j\}\not\subseteq I',\ i>j}\ \sum_{S\subseteq J,\ S\ni i,j} \hat{f}(S)^2 \\
&\leq \frac{\epsilon^2}{16} + \sum_{i,j\in J,\ \{i,j\}\not\subseteq I',\ i>j} 2\cdot|\hat{f}(\{i,j\})| \\
&\leq \frac{\epsilon^2}{16} + \frac{|J|^2}{2}\cdot 2\cdot \frac{\epsilon^2}{8\cdot s^2} \leq \frac{\epsilon^2}{4}.
\end{aligned}$$

All we need now is to find a small set of indices $I \supseteq I'$. We simply estimate degree-1 and 2 Fourier coefficients of $f$ to accuracy $\epsilon^2/(32\cdot s^2)$ with confidence at least $5/6$ using random examples. Let $\tilde{f}(S)$ for $S\subseteq[n]$ of size 1 or 2 denote the obtained estimates. We define

$$I = \left\{ i \;\middle|\; |\tilde{f}(\{i\})| \geq \frac{3\epsilon}{16\cdot\sqrt{s}} \right\} \bigcup \left\{ i \;\middle|\; \exists j, |\tilde{f}(\{i,j\})| \geq \frac{3\epsilon^2}{32\cdot s^2} \right\} .$$

If estimates are within the desired accuracy, then clearly, $I\supseteq I'$. At the same time $I\subseteq I''$, where

$$I'' = \left\{ i \;\middle|\; |\hat{f}(\{i\})| \geq \frac{\epsilon}{8\cdot\sqrt{s}} \right\} \bigcup \left\{ i \;\middle|\; \exists j, |\hat{f}(\{i,j\})| \geq \frac{\epsilon^2}{16\cdot s^2} \right\} .$$

By Lem. 25, $|I''| \leq 32\cdot s^2/\epsilon^2$.

Finally, to bound the running time we observe that, by the standard application of Chernoff bound with the union bound, $O(\log(n)\cdot s^4/\epsilon^4)$ random examples are sufficient to obtain the desired estimates with confidence of $5/6$. The estimation of the coefficients can be done in $O(n^2\log(n)\cdot s^4/\epsilon^4)$ time. ∎

We can now use the result from [FV13] that for every submodular function $f:\{0,1\}^n\to[0,1]$, $s_f(\epsilon/2) = O(\log(1/\epsilon)/\epsilon^2)$ to obtain the following corollary.

**Corollary 28.** *Let $f:\{0,1\}^n\to[0,1]$ be a submodular function. There exists an algorithm, that given any $\epsilon>0$ and access to random and uniform examples of $f$, with probability at least $5/6$, finds a set of variables $I$ of size $\tilde{O}(1/\epsilon^6)$ such that there exists a submodular $I$-junta $h$ satisfying $\|f-h\|_2\leq\epsilon$. The algorithm runs in time $\tilde{O}(n^2/\epsilon^{12})$ and uses $\tilde{O}(\log(n)/\epsilon^{12})$ examples.*

We use Corollary 28 with the least squares regression over polynomials of degree $O(\log(1/\epsilon)/\epsilon^{4/5})$ on the influential variables to obtain the learning algorithm claimed in Theorem 1.

Finally, for completeness we also state the corollaries for agnostic learning of XOS and submodular functions:

**Theorem 29.** *Let $\mathcal{C}_s$ be the class of all submodular functions with range in $[0,1]$. There exists an algorithm that learns $\mathcal{C}_s$ agnostically with excess $\ell_2$-error $\epsilon$ and runs in time $n^{O(\log(1/\epsilon)/\epsilon^{4/5})}$.*

**Theorem 30.** *Let $\mathcal{C}_x$ be the class of all XOS functions with range in $[0,1]$. There exists an algorithm that learns $\mathcal{C}_x$ agnostically with excess $\ell_2$-error $\epsilon$ and runs in time $n^{O(1/\epsilon)}$.*

## VI. Lower Bounds

In this section we prove tight lower bounds on low-degree spectral concentration and learning of monotone submodular, XOS and self-bounding functions.

### A. Monotone Submodular Functions

We start by showing that for any $\epsilon>0$ there exists an explicit monotone submodular function over $\Theta(\epsilon^{-4/5})$ variables that requires degree $\Theta(\epsilon^{-4/5})$ to $\ell_2$-approximate within $\epsilon$. The "hockey-stick" function of $k$ (out of $n$) variables is defined as follows: $\mathsf{hs}_k(x) = \min\{1, 2\cdot w_k(x)/k\}$, where $w_k(x) = \sum_{i=1}^k x_i$ is the Hamming weight of the first $k$ bits of $x$. In [FKV13] it was shown that this function has a Fourier coefficient of degree $k$ whose value is at least $\Omega(k^{-3/2})$. This immediately implies a lower bound of $\Omega(\epsilon^{-2/3})$ on $\mathsf{deg}_\epsilon^{\ell_2}(\mathsf{hs}_k)$ for $k=\Theta(\epsilon^{-2/3})$. We now give a more careful analysis of the low-degree spectral concentration of $\mathsf{hs}_k$ that leads to the nearly tight lower bound.

The hockey-stick function is closely related to the well-studied Boolean majority function for which tight spectral concentration bounds are known [O'D14]. Specifically, it is easy to see that for every $i$,

$$\partial_i \mathsf{hs}_k(x) = 2(1 - \mathsf{maj}_k(x))/k, \tag{5}$$

where $\mathsf{maj}_k(x) = 1$ if $w_k(x) \geq k/2$ and 0 otherwise. This correspondence allows us to easily obtain a lower bound on the low-degree spectral concentration of $\mathsf{hs}_k(x)$ .

**Lemma 31.** *For any $k \leq n$ and $d \leq k/2$, $W^{>d}(\mathsf{hs}_k) = \Omega(d^{-3/2}/k)$. In particular, for some constant $c_1$, $k = c_1\epsilon^{-4/5}$ and $d = \lfloor k/2 \rfloor$ gives $W^{>d}(\mathsf{hs}_k) \geq \epsilon^2$.*

*Proof:* We first observe that by the properties of partial derivatives given in Sec. II and eq.(5), for every $S \subseteq [k]$ such that $|S| \geq 2$ and $i \in S$,

$$\widehat{\mathsf{hs}_k}(S) = -\widehat{\partial_i \mathsf{hs}_k}(S \setminus i)/2 = \frac{\widehat{\mathsf{maj}_k}(S \setminus i)}{k}.$$

For $2 \leq j \leq k$,

$$\begin{aligned} W^j(\mathsf{hs}_k) &= \sum_{S \subseteq [k],\ |S|=j} \widehat{\mathsf{hs}_k}(S)^2 = \sum_{S \subseteq [k],\ |S|=j,\ i \in S} \frac{\widehat{\mathsf{maj}_k}(S \setminus i)^2}{k^2} \\ &= \frac{k-j+1}{j} \sum_{S \subseteq [k],\ |S|=j-1} \frac{\widehat{\mathsf{maj}_k}(S)^2}{k^2} = \frac{k-j+1}{k^2 j} \cdot W^{j-1}(\mathsf{maj}_k) \end{aligned}$$

We now use the estimate $W^{j-1}(\mathsf{maj}_k) \geq c(j-1)^{-3/2}$ for some constant $c > 0$ [O'D14]. This estimate implies that

$$\begin{aligned} W^{>d}(\mathsf{hs}_k) &\geq \sum_{2k/3+1 \geq j > d} W^j(\mathsf{hs}_k) = \sum_{2k/3+1 \geq j > d} \frac{k-j+1}{k^2 j} \cdot W^{j-1}(\mathsf{maj}_k) \\ &\geq c \sum_{2k/3 \geq j \geq d} \frac{k-j}{k^2} j^{-5/2} \geq \frac{c}{3k} \sum_{2k/3 \geq j \geq d} j^{-5/2} \geq \frac{c}{3k} \int_d^{2k/3} t^{-5/2} dt \\ &\geq \frac{c}{3k} \cdot \frac{2}{3} \left( d^{-3/2} - (2k/3)^{-3/2} \right) \geq \frac{2c}{27} \cdot \frac{d^{-3/2}}{k}, \end{aligned}$$

where in the last inequality we used the condition that $d \leq k/2$ and hence $d^{-3/2} - (2k/3)^{-3/2} \geq d^{-3/2}/3$. ■

We now show that any algorithm that PAC learns monotone submodular functions with $\ell_2$ error of $\epsilon$ must use $2^{\Omega(\epsilon^{-4/5})}$ examples. This result is based on a reduction from learning the class all Boolean functions on $k$ variables with error $1/4$ to the problem of learning submodular functions on $2t = k + \lceil \log k \rceil + O(1)$ variables with $\ell_2$ error of $\Theta(\frac{1}{t^{5/4}})$. Any algorithm that learns the class of all Boolean functions on $k$ variables to accuracy $1/4$ requires at least $2^{\Omega(k)}$ bits of information about the target function and, in particular, at least that many random examples or other Boolean-valued queries are necessary. The reduction is identical to the reduction in [FKV13] which proved an analogous result for learning with $\ell_1$ error of $\Theta(\frac{1}{t^{3/2}})$. Therefore the analysis of the reduction follows closely that from [FKV13] and is included in the full version of this work [FV15].

**Lemma 32.** *For $k > 0$, let $t > 0$ be the smallest such that $\binom{2t}{t} \geq 2^k$ (and thus $4 \cdot 2^k > \binom{2t}{t} \geq 2^k$). For every Boolean function $h : \{0,1\}^k \to \{0,1\}$ there exists a monotone submodular function $f : \{0,1\}^{2t} \to [0,1]$ such that:*

1) *$f$ can be computed at any point $x \in \{0,1\}^{2t}$ in at most a single query to $h$ and in time $O(k)$; given a single random and uniform example of $h$, a random and uniform example of $f$ can be produced in time $O(k)$.*
2) *Let $\alpha = \frac{2^k \cdot \sqrt{t}}{2^{2t}} = \Theta(1)$. For any $\beta > 0$, given a function $\tilde{f} : \{0,1\}^{2t} \to \mathbb{R}$ such that $\|f - \tilde{f}\|_2 \leq \frac{\sqrt{\alpha\beta}}{4 \cdot t^{5/4}}$, one can obtain a Boolean function $\tilde{h} : \{0,1\}^k \to \{0,1\}$ such that $\mathbf{Pr}_{\mathcal{U}}[\tilde{h}(x) \neq h(x)] \leq \beta$ and $\tilde{h}$ can be computed at any point $x \in \{0,1\}^k$, with a single query to $\tilde{f}$ in time $O(k)$.*

By choosing $\beta = 1/4$ in Lemma 32 we obtain the following result:

**Theorem 33.** *Any algorithm that PAC learns all monotone submodular functions with range $[0,1]$ to $\ell_2$ error of $\epsilon > 0$ requires $2^{\Omega(\epsilon^{-4/5})}$ random examples of (or value queries to) the target function.*

## B. XOS functions

The lower bounds for XOS functions are based on a simple mapping from monotone DNF (MDNF) formulas to XOS functions. We say that a function $h$ is $s$-term $t$-MDNF if $h(x) = \bigvee_{j\in[s]} T_j(x)$, where each $T_j \subseteq [k]$, $|T_j| \leq t$ and $T_j(x) = \bigwedge_{i\in T_j} x_i$.

**Lemma 34.** *For every $s$-term $t$-MDNF $h : \{0,1\}^k \to \{0,1\}$, let $f : \{0,1\}^k \to [0,1]$ be given by $f(x) = 1 - \frac{1-h(x)}{t}$, if $x \neq \mathbf{0}$ and $f(x) = 0$ otherwise. Then $f$ is an XOS function of size $s + k$.*

*Proof:* Let $h(x) = \bigvee_{j\in[s]} T_j(x)$ be an $s$-term $t$-MDNF representation of $h$. Then it is easy to verify that

$$f(x) = \max\left\{ \max \frac{\sum_{i\in T_j} x_i}{|T_j|},\ \max_{i\in[k]} \frac{t-1}{t} x_i \right\}.$$

■

An immediate corollary of Lemma 34 is that for any $\beta > 0$, a function $g$ such that $\|f - g\|_2 \leq \sqrt{\beta}/(2t)$ gives a function $\tilde{h}$ such that $\mathbf{Pr}_{\mathcal{U}}[\tilde{h}(x) \neq h(x)] \leq \beta + 2^{-k}$.

To obtain our lower bounds, we rely on known results for MDNFs obtained by choosing random conjunctions of size $\Theta(\sqrt{k})$. Such MDNFs were first analyzed by Talagrand [Tal96]. For our spectral concentration lower bound we will use the fact that Talagrand's DNFs are noise sensitive [MO02] together with a reverse connection between noise sensitivity and low-degree spectral concentration.

We first recall the definition and basic properties of the noise sensitivity.

**Definition 35** (Noise sensitivity)**.** *For $\alpha \in [0,1], x \in \{0,1\}^n$, we define a distribution $N_\alpha(x)$ over $y \in \{0,1\}^n$ by letting $y_i = x_i$ with probability $1-\alpha$ and $y_i = 1 - x_i$ with probability $\alpha$, independently for each $i$. For a Boolean function $h$, the noise sensitivity of $h$ with noise rate $\alpha$ is defined as*

$$\mathsf{NS}_\alpha(h) = \mathbf{Pr}_{x\sim\mathcal{U},\ y\sim N_\alpha(x)}[h(x) \neq h(y)].$$

*Noise sensitivity satisfies (e.g. [O'D14]):*

$$\mathsf{NS}_\alpha(h) = \frac{1}{2}\sum_{i=0}^{k}(1 - (1-2\alpha)^i) \cdot W^i(h). \tag{6}$$

The following theorem was proved in [MO02], following Talagrand's analysis [Tal96].

**Theorem 36** ([MO02])**.** *For every $k$, there exists a $\sqrt{k}$-MDNF $h$ such that $\mathsf{NS}_{1/\sqrt{k}}(h) = \Omega(1)$.*

This result implies that such functions have a large Fourier mass above level $\Omega(\sqrt{k})$.

**Corollary 37.** *For every $k$, there exists a $\sqrt{k}$-MDNF $h$ such that for $d = \Omega(\sqrt{k})$, $W^{>d}(h) = \Omega(1)$.*

*Proof:* Equation (6) implies that for every $d$,

$$\begin{aligned}\mathsf{NS}_\alpha(h) &= \frac{1}{2}\sum_{i=0}^{k}(1-(1-2\alpha)^k)\cdot W^k(h) \leq \frac{1}{2}\sum_{i=0}^{d}(1-(1-2\alpha)^d)\cdot W^i(h) + \frac{1}{2}W^{>d}(h) \\ &\leq \frac{1}{2}\left((1-(1-2\alpha)^d)\|h\|_2^2 + W^{>d}(h)\right) < \frac{1}{2}\left(2\alpha d\cdot\|h\|_2^2 + W^{>d}(h)\right) \\ &= \alpha d\cdot\|h\|_2^2 + W^{>d}(h)/2 \leq \alpha d + W^{>d}(h)/2.\end{aligned}$$

By Theorem 36, there exists a $\sqrt{k}$-MDNF $h$ such that for some constant $c > 0$, $\mathsf{NS}_{1/\sqrt{k}}(h) \geq c$. Let $d = c\sqrt{k}/2$ we obtain that

$$W^{>d}(h) \geq 2\left(\mathsf{NS}_{1/\sqrt{k}}(h) - \frac{d}{\sqrt{k}}\right) \geq c.$$

■

From here we obtain a lower bound on low-degree spectral concentration of XOS functions using Lemma 34.

**Theorem 38.** *For every $\epsilon > 0$ there exists $k = \Theta(1/\epsilon^2)$ and an XOS function $f : \{0,1\}^k \to [0,1]$ such that $\mathsf{deg}^{\ell_2}_{\epsilon}(f) = \Omega(1/\epsilon)$.*

*Proof:* For $k > 0$, let $h$ be the $\sqrt{k}$-MDNF $h$ such that for $d = \Omega(\sqrt{k})$, $W^{>d}(h) = \Omega(1)$. Let $f$ be the XOS function obtained from $h$ using Lemma 34. Then, by the linearity of Fourier coefficients and the fact that $f$ differs from $1 - \frac{1-h(x)}{t}$ only on a single point, we obtain that

$$W^{>d}(f) \geq W^{>d}(h)/d^2 - 2^{-k} = \Omega(1/k).$$

This means that for some $k = \Theta(1/\epsilon^2)$ and $d = \Omega(1/\epsilon)$ we have $W^{>d}(f) \geq \epsilon^2$. ∎

Our lower bound for PAC learning of XOS functions is based on the following lower bound for learning MDNF by Blum *et al.* [BBL98].

**Theorem 39** ([BBL98])**.** *For any sufficiently large $k$ and $q \geq k$, any algorithm that PAC learns $t$-MDNF for $t = \log(3qk)$ over the uniform distribution and uses at most $q$ random examples (or value queries) will have error of at least $1/2 - O(\log(qk)/\sqrt{k})$.*

We note that Theorem 39 implies a slightly weaker (by a logarithmic factor in the degree) version of Corollary 37 since low-degree spectral concentration implies learning (in fact, as shown in [DSFT+15] this argument also implies a lower bound on $\ell_1$-approximation by polynomials). We now prove a lower bound for PAC learning XOS functions which we state for the $\ell_1$ error (which implies the same lower bound for $\ell_2$ error).

**Theorem 40.** *Any algorithm that PAC learns all XOS functions from $\{0,1\}^n$ to $[0,1]$ with $\ell_1$ error of $\epsilon > 0$ requires $2^{\Omega(1/\epsilon)}$ random examples of (or value queries to) the target function.*

*Proof:* We reduce learning of $t$-MDNF over $k$ variables (for $t$ and $k$ to be chosen later) to learning of XOS using Lemma 34, namely we replace each example $(x, f(x))$ with $\left(x, 1 - \frac{1-f(x)}{t}\right)$ and then replace the hypothesis $h(x)$ with $h'$ such that $h'(x) = 1$ whenever $h(x) \geq 1 - 1/(2t)$. By Lemma 34, any algorithm that achieves $\ell_1$ error of $\frac{1/4}{2t} - 2^{-k}$ gives a Boolean hypothesis for the MDNF problem with error of less than $1/4$.

By Theorem 39, there exists a constant $c > 0$ such that for $q = 2^{c\sqrt{k}}$ and $t = \log(3qk)$, the error of any PAC learning algorithm for $t$-MDNF that uses at most $q$ random examples (or value queries) is at least $1/4$. Note that

$$\frac{1/4}{2t} - 2^{-k} = \frac{1}{8\log(3qk)} - 2^{-k} = \frac{1}{8(\log(3k) + c\sqrt{k})} - 2^{-k},$$

and therefore there exists a constant $c_1 > 0$ such that for every $\epsilon > 0$ and $k = c_1/\epsilon^2$, $\frac{1/4}{2t} - 2^{-k} \geq \epsilon$. Applying the guarantees of Theorem 39, we get that the number of random examples (or value queries) used to learn with $\ell_1$ error of $\epsilon$ must be larger than $q = 2^{c\sqrt{k}} = 2^{\Omega(1/\epsilon)}$.

∎

### C. Self-bounding functions

We now show that upper bounds on low-degree spectral concentration that we proved for XOS and submodular functions cannot be extended to the whole class of self-bounding functions. Our construction is based on the classical Hamming code which we briefly describe here for completeness. For an integer $r$ a Hamming code is a linear mapping (over $\mathbf{GF}(2)$) $c : \{0,1\}^{2^r-r-1} \to \{0,1\}^r$ such that for any two distinct $v, w \in \{0,1\}^{2^r-r-1}$, the Hamming distance between $v \circ c(v)$ and $w \circ c(w)$ is at least 3, where we use "$\circ$" to denote the concatenation of strings. We now show that for $k = 2^r - r - 1$ a Hamming code gives a way to embed any Boolean function into a self-bounding function which we describe below.

**Lemma 41.** *For an integer $r$, $k = 2^r - r - 1$ and any Boolean function $h : \{0,1\}^k \to \{0,1\}$ let $f : \{0,1\}^{k+r} \to [0,1]$ be given by $f(x \circ z) = h(x)$, if $z = c(x)$ and $f(x \circ z) = 1$, otherwise. Then $f$ is a self-bounding function.*

*Proof:* Let $x \circ z$ be a point in $\{0,1\}^{k+r}$. If $f(x \circ z) = 0$ then $f$ cannot be lower on any point that differs from $x \circ z$ in one coordinate, and therefore the self-bounding condition holds at $x \circ z$. If $f(x \circ z) = 1$ then there exists at most one point $y \in \{0,1\}^{k+r}$ that differs from $x \circ z$ in a single coordinate and $f(y) = 0$. This follows from the fact that, by definition of $f$, if $f(y) = 0$ then $y = x' \circ c(x')$ for some $x' \in \{0,1\}^k$. By the properties of $c$, any two points of this form are at Hamming distance at least 3 and therefore two distinct points cannot be at Hamming distance 1 to $x \circ z$. This means that

$$\sum_{i \in [k+r]} |f(x \circ z) - f((x \circ z) \oplus e_i)| \leq 1 = f(x \circ z).$$

∎

A spectral concentration bound can be obtained by analyzing the embedding of a $\{0,1\}$-parity function $h = \sum_{i \in S} x_i \mod 2$. To avoid the direct calculation which requires using additional properties of the Hamming code we will derive the lower-bound via lower bounds for learning below.

**Theorem 42.** *Any algorithm that PAC learns all self-bounding functions from $\{0,1\}^n$ to $[0,1]$ with $\ell_2$ error of $\epsilon > 0$ requires $2^{\Omega(1/\epsilon^2)}$ random examples of (or value queries to) the target function.*

*Proof:* We reduce learning of all Boolean functions on $k = 2^r - r - 1$ (for $r$ to be chosen later) variables over the uniform distribution to learning of self-bounding functions using Lemma 41. Namely, given a random and uniform example $(x, \ell)$ of some unknown Boolean target function $h$ we output a random example $(x \circ z, \ell')$ of the function $f$ that is equal to the embedding of $h$ given by Lemma 41. This is done by choosing $z$ uniformly from $\{0,1\}^r$ and having $\ell' = \ell$ if $z = c(x)$ and $\ell' = 1$ otherwise (a value query can be answered similarly using a single value query to $h$). Given a hypothesis $\tilde{f}$ we define $\tilde{h}(x) = 1$ if $\tilde{f}(x \circ c(x)) \geq 1/2$ and $\tilde{h}(x) = 0$ otherwise. Observe that,

$$\begin{aligned}\mathbf{Pr}_{\mathcal{U}_k}[\tilde{h}(y) \neq h(y)] &\leq \mathbf{Pr}_{x\circ z \sim \mathcal{U}_{k+r}}[|\tilde{f}(x \circ z) - f(x \circ z)| \geq 1/2 \mid c(x) = z] \\ &\leq 4 \cdot \mathbf{E}_{x\circ z \sim \mathcal{U}_{k+r}}[|\tilde{f}(x \circ z) - f(x \circ z)|^2 \mid c(x) = z] \\ &\leq 4 \cdot \frac{\mathbf{E}_{x\circ z \sim \mathcal{U}_{k+r}}[|\tilde{f}(x \circ z) - f(x \circ z)|^2}{2^{-r}} = 2^{r+2} \cdot \|\tilde{f} - f\|_2^2.\end{aligned}$$

We now let $r = \lfloor \log(1/\epsilon^2) \rfloor + 4$. This choice ensures that if $\tilde{f}$ has $\ell_2$ error of less than $\epsilon$ then $\tilde{h}$ has error of less than $1/4$. Learning all Boolean functions to error of at most $1/4$ requires $2^{\Omega(k)} = 2^{\Omega(2^r)} = 2^{\Omega(1/\epsilon^2)}$ random examples (or value queries) and therefore we obtain our claim. ■

We now observe that there exists some constant $c$ such that $\mathsf{deg}_\epsilon^{\ell_2}(f) \leq c/\epsilon^2$. Otherwise, for any constant $c_0$, using Theorem 24 we could obtain an algorithm that learns self-bounding functions using $2^{c_1/\epsilon^2}$ random examples contradicting Theorem 24.

**Theorem 43.** *For every $\epsilon > 0$ there exists $k = O(1/\epsilon^2)$ and a self-bounding function $f : \{0,1\}^k \to [0,1]$ such that $\mathsf{deg}_\epsilon^{\ell_2}(f) = \Omega(1/\epsilon^2)$.*


## Acknowledgements

We would like to thank Pravesh Kothari for useful discussions and his help with the proof of Lemma 31.


## References


[AA14] Scott Aaronson and Andris Ambainis. The need for structure in quantum speedups. *Theory of Computing*, 10:133–166, 2014.

[AM06] Kazuyuki Amano and Akira Maruoka. On learning monotone boolean functions under the uniform distribution. *Theor. Comput. Sci.*, 350(1):3–12, 2006.

[BB14] Arturs Backurs and Mohammad Bavarian. On the sum of L1 influences. In *CCC*, pages 132–143, 2014.

[BBL98] A. Blum, C. Burch, and J. Langford. On learning monotone boolean functions. In *FOCS*, pages 408–415, 1998.

[BCIW12] M.F. Balcan, F. Constantin, S. Iwata, and L. Wang. Learning valuation functions. *COLT*, 23:4.1–4.24, 2012.

[BDF+12] A. Badanidiyuru, S. Dobzinski, Hu Fu, R. Kleinberg, N. Nisan, and T. Roughgarden. Sketching valuation functions. In *SODA*, pages 1025–1035, 2012.

[BH12] M.F. Balcan and N. Harvey. Submodular functions: Learnability, structure, and optimization. *CoRR*, abs/1008.2159, 2012. Earlier version in STOC 2011.

[BLM00] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Struct. Algorithms*, 16(3):277–292, 2000.

[BLN06] D. J. Lehmann B. Lehmann and N. Nisan. Combinatorial auctions with decreasing marginal utilities. *Games and Economic Behavior*, 55:1884–1899, 2006.

[BM02] P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2002.

[BOL85] M. Ben-Or and N. Linial. Collective coin flipping, robust voting schemes and minima of banzhaf values. In *FOCS*, pages 408–416, 1985.

[BOSY13] E. Blais, K. Onak, R. Servedio, and G. Yaroslavtsev. Concise representations of discrete submodular functions, 2013. Personal communication.

[BRY14] Piotr Berman, Sofya Raskhodnikova, and Grigory Yaroslavtsev. $L_p$-testing. In *STOC*, pages 164–173, 2014.

[BT96] N. Bshouty and C. Tamon. On the Fourier spectrum of monotone functions. *Journal of the ACM*, 43(4):747–770, 1996.

[CKKL12] M. Cheraghchi, A. Klivans, P. Kothari, and H. Lee. Submodular functions are noise stable. In *SODA*, pages 1586–1592, 2012.

[DFKO06] I. Dinur, E. Friedgut, G. Kindler, and R. O'Donnell. On the Fourier tails of bounded functions over the discrete cube. In *STOC*, pages 437–446, 2006.

[DLM+08] Dana Dachman-Soled, Homin K. Lee, Tal Malkin, Rocco A. Servedio, Andrew Wan, and Hoeteck Wee. Optimal cryptographic hardness of learning monotone functions. In *ICALP Track A: Algorithms, Automata, Complexity, and Games*, pages 36–47, 2008.

[DS06] S. Dobzinski and M. Schapira. An improved approximation algorithm for combinatorial auctions with submodular bidders. In *SODA*, pages 1064–1073, 2006.

[DSFT+15] Dana Dachman-Soled, Vitaly Feldman, Li-Yang Tan, Andrew Wan, and Karl Wimmer. Approximate resilience, monotonicity, and the complexity of agnostic learning. In *SODA*, 2015.

[Edm70] Jack Edmonds. Matroids, submodular functions and certain polyhedra. *Combinatorial Structures and Their Applications*, pages 69–87, 1970.

[Fei06] Uriel Feige. On maximizing welfare when utility functions are subadditive. In *ACM STOC*, pages 41–50, 2006.

[FFI01] L. Fleischer, S. Fujishige, and S. Iwata. A combinatorial, strongly polynomial-time algorithm for minimizing submodular functions. *JACM*, 48(4):761–777, 2001.

[FK14] V. Feldman and P. Kothari. Learning coverage functions and private release of marginals. In *COLT*, pages 679–702, 2014.

[FKV13] V. Feldman, P. Kothari, and J. Vondrák. Representation, approximation and learning of submodular functions using low-rank decision trees. In *COLT*, 2013.

[FKV14] V. Feldman, P. Kothari, and J. Vondrák. Nearly tight bounds on $\ell_1$ approximation of self-bounding functions. *CoRR*, abs/1404.4702, 2014.

[FMV07] U. Feige, V. Mirrokni, and J. Vondrák. Maximizing non-monotone submodular functions. In *FOCS*, pages 461–471, 2007.

[Fra97] András Frank. Matroids and submodular functions. *Annotated Biblographies in Combinatorial Optimization*, pages 65–80, 1997.

[Fri98] E. Friedgut. Boolean functions with low average sensitivity depend on few coordinates. *Combinatorica*, 18(1):27–35, 1998.

[FV13] V. Feldman and J. Vondrák. Optimal bounds on approximation of submodular and XOS functions by juntas. *CoRR*, abs/1307.3301, 2013. Earlier version in FOCS 2013.

[FV15] Vitaly Feldman and Jan Vondrák. Tight bounds on low-degree spectral concentration of submodular and XOS functions. *CoRR*, abs/1504.03391, 2015.

[GGL+00] Oded Goldreich, Shafi Goldwasser, Eric Lehman, Dana Ron, and Alex Samorodnitsky. Testing monotonicity. *Combinatorica*, 20(3):301–337, 2000.

[GHIM09] M. Goemans, N. Harvey, S. Iwata, and V. Mirrokni. Approximating submodular functions everywhere. In *SODA*, pages 535–544, 2009.

[GHRU11] A. Gupta, M. Hardt, A. Roth, and J. Ullman. Privately releasing conjunctions and the statistical query barrier. In *STOC*, pages 803–812, 2011.

[GKS05] C. Guestrin, A. Krause, and A. Singh. Near-optimal sensor placements in gaussian processes. In *ICML*, pages 265–272, 2005.

[GV06] M. Goemans and J. Vondrák. Covering minimum spanning trees of random subgraphs. *Random Struct. Algorithms*, 29(3):257–276, 2006.

[KGGK06] A. Krause, C. Guestrin, A. Gupta, and J. Kleinberg. Near-optimal sensor placements: maximizing information while minimizing communication cost. In *IPSN*, pages 2–10, 2006.

[KK07] J. Kahn and G. Kalai. Thresholds and expectation thresholds. *Combinatorics, Probability and Computing*, 16(3):492–502, 2007.

[KKL88] J. Kahn, G. Kalai, and N. Linial. The influence of variables on Boolean functions. In *FOCS*, pages 68–80, 1988.

[KKM13] Daniel M. Kane, Adam Klivans, and Raghu Meka. Learning halfspaces under log-concave densities: Polynomial approximations and moment matching. In *COLT*, pages 522–545, 2013.

[KKMS08] A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.

[KLV94] M. Kearns, M. Li, and L. Valiant. Learning boolean formulas. *Journal of the ACM*, 41(6):1298–1328, 1994.

[KP00] Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, volume 47 of *Progress in Probability*, pages 443–457. Birkhauser Boston, 2000.

[KS08] Adam R. Klivans and Rocco A. Servedio. Learning intersections of halfspaces with a margin. *J. Comput. Syst. Sci.*, 74(1):35–48, 2008.

[KSG08] A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *JMLR*, 9:235–284, 2008.

[KSS94] M. Kearns, R. Schapire, and L. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2-3):115–141, 1994.

[KST08] S. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*, pages 793–800, 2008.

[LMN93] N. Linial, Y. Mansour, and N. Nisan. Constant depth circuits, Fourier transform and learnability. *Journal of the ACM*, 40(3):607–620, 1993.

[Lov83] László Lovász. Submodular functions and convexity. *Mathematical Programmming: The State of the Art*, pages 235–257, 1983.

[MO02] Elchanan Mossel and Ryan O'Donnell. On the noise sensitivity of monotone functions. In *Mathematics and Computer Science II*, pages 481–495. Springer, 2002.

[MR06] C. McDiarmid and B. Reed. Concentration for self-bounding functions and an inequality of talagrand. *Random structures and algorithms*, 29:549–557, 2006.

[O'D03] R. O'Donnell. *Computational Applications of Noise Sensitivity*. PhD thesis, 2003.

[O'D14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[OW13] Ryan O'Donnell and Karl Wimmer. KKL, Kruskal-Katona, and monotone nets. *SIAM J. Comput.*, 42(6):2375–2399, 2013.

[Que95] Maurice Queyranne. A combinatorial algorithm for minimizing symmetric submodular functions. In *SODA*, pages 98–101, 1995.

[RY13] S. Raskhodnikova and G. Yaroslavtsev. Learning pseudo-boolean k-DNF and submodular functions. In *SODA*, 2013.

[SV11] C. Seshadhri and J. Vondrák. Is submodularity testable? In *Innovations in computer science*, pages 195–210, 2011.

[Tal94] M. Talagrand. On Russo's approximate zero-one law. *The Annals of Probability*, pages 1576–1587, 1994.

[Tal96] M. Talagrand. How much are increasing sets positively correlated? *Combinatorica*, 16(2):243–258, 1996.

[Val84] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[Von08] J. Vondrák. Optimal approximation for the submodular welfare problem in the value oracle model. In *STOC*, pages 67–74, 2008.