Testing Fourier dimensionality and sparsity

Parikshit Gopalan, Ryan O'Donnell, Rocco A. Servedio,
Amir Shpilka, and Karl Wimmer

parik@microsoft.com, odonnell@cs.cmu.edu, rocco@cs.columbia.edu,
shpilka@cs.technion.ac.il, kwimmer@andrew.cmu.edu

Abstract. We present a range of new results for testing properties of Boolean functions that are defined in terms of the Fourier spectrum. Broadly speaking, our results show that the property of a Boolean function having a concise Fourier representation is locally testable.

We first give an efficient algorithm for testing whether the Fourier spectrum of a Boolean function is supported in a low-dimensional subspace of $\mathbb{F}_2^n$ (equivalently, for testing whether f is a junta over a small number of parities). We next give an efficient algorithm for testing whether a Boolean function has a sparse Fourier spectrum (small number of nonzero coefficients). In both cases we also prove lower bounds showing that any testing algorithm — even an adaptive one — must have query complexity within a polynomial factor of our algorithms, which are nonadaptive. Finally, we give an "implicit learning" algorithm that lets us test *any* sub-property of Fourier concision.

Our technical contributions include new structural results about sparse Boolean functions and new analysis of the pairwise independent hashing of Fourier coefficients from [12].

1 Introduction

Recent years have witnessed broad research interest in the local testability of mathematical objects such as graphs, error-correcting codes, and Boolean functions. One of the goals of this study is to understand the minimal conditions required to make a property locally testable. For graphs and codes, works such as [1, 5, 3, 4] and [16, 17] have given fairly general characterizations of when a property is testable. For Boolean functions, however, testability is less well understood. On one hand, there are a fair number of testing algorithms for specific classes of functions such as $\mathbb{F}_2$-linear functions [10, 6], dictators [7, 21], low-degree $\mathbb{F}_2$-polynomials [2, 22], juntas [14, 9], and halfspaces [20]. But there is not much by way of general characterizations of what makes a property of Boolean functions testable. Perhaps the only example is the work of [11], showing that any class of functions sufficiently well-approximated by juntas is locally testable.

It is natural to think that general characterizations of testability for Boolean functions might come from analyzing the Fourier spectrum (see e.g. [13, Section 9.1]). For one thing, many of the known tests — for linearity, dictators, juntas, and halfspaces — involve a careful analysis of the Fourier spectrum. Further intuition comes from learning theory, where the class of functions that are learnable using many of the well-known algorithms [19, 18, 15] can be characterized in terms of the Fourier spectrum.

In this paper we make some progress toward this goal, by giving efficient algorithms for testing Boolean functions that have *low-dimensional* or *sparse* Fourier representations. These are two natural ways to formalize what it means for a Boolean function to have a "concise" Fourier representation; thus, roughly speaking our results show that the property of having a concise Fourier representation is efficiently testable. Further, as we explain below, Boolean functions with low-dimensional or sparse Fourier representations are closely related to linear functions, juntas, and low-degree polynomials whose testability has been intensively studied, and thus the testability of these classes is a natural question in its own right. Building on our testing algorithms, we are able to give an "implicit learner" (in the sense of [11]), which determines the "truth table" of a sparse Fourier spectrum without actually knowing the identities of the underlying Fourier characters. This lets us test *any* sub-property of having a concise Fourier representation. We view this as a step toward the goal of a more unified understanding of the testability of Boolean functions.

Our algorithms rely on new structural results on Boolean functions with sparse and close-to-sparse Fourier spectrums, which may find applications elsewhere. As one such application, we show that the well-known Kushilevitz-Mansour algorithm is in fact an exact proper learning algorithm for Boolean functions with sparse Fourier representations. As another application, we give polynomial-time unique-decoding algorithms for sparse functions and k-dimensional functions; due to space limitations these results will only appear in the full version of the paper.

1.1 The Fourier spectrum, dimensionality, and sparsity

We are concerned with testing various properties defined in terms of the *Fourier representation* of Boolean functions $f : \mathbb{F}_2^n \to \{-1, 1\}$. Input bits will be treated as $0, 1 \in \mathbb{F}_2$, the field with two elements; output bits will be treated as $-1, 1 \in \mathbb{R}$. Every Boolean function $f : \mathbb{F}_2^n \to \mathbb{R}$ has a unique representation as

$$f(x) = \sum_{\alpha \in \mathbb{F}_2^n} \hat{f}(\alpha)\chi_\alpha(x) \text{ where } \chi_\alpha(x) \stackrel{\text{def}}{=} (-1)^{\langle \alpha, x\rangle} = (-1)^{\sum_{i=1}^n \alpha_i x_i}. \tag{1}$$

The coefficients $\hat{f}(\alpha)$ are the *Fourier coefficients* of f, and the functions $\chi_\alpha(\cdot)$ are sometimes referred to as *linear functions* or *characters*. In addition to treating input strings x as lying in $\mathbb{F}_2^n$, we also index the characters by vectors $\alpha \in \mathbb{F}_2^n$. This is to emphasize the fact that we are concerned with the linear-algebraic structure. We write $\mathrm{Spec}(f)$ for the Fourier spectrum of f, i.e. the set $\{\alpha \in \mathbb{F}_2^n : \hat{f}(\alpha) \neq 0\}$.

Dimensionality and sparsity (and degree). A function $f : \mathbb{F}_2^n \to \{-1, 1\}$ is said to be *k-dimensional* if $\mathrm{Spec}(f)$ lies in a k-dimensional subspace of $\mathbb{F}_2^n$. An equivalent definition is that f is k-dimensional if it is a function of k characters $\chi_{\alpha_1}, \ldots, \chi_{\alpha_k}$, i.e. f is a junta over k parity functions (this is easily seen by picking $\{\alpha_i\}$ to be a basis for $\mathrm{Spec}(f)$). We write $\dim(f)$ to denote the smallest k for which f is k-dimensional. A function f is said to be *s-sparse* if $|\mathrm{Spec}(f)| \leq s$. We write $\mathrm{sp}(f)$ to denote $|\mathrm{Spec}(f)|$, i.e. the smallest s for which f is s-sparse.

We recall the notion of the $\mathbb{F}_2$-*degree* of a Boolean function, $\deg_2(f)$, which is the degree of the unique multilinear $\mathbb{F}_2$-polynomial representation for f when viewed as a function $\mathbb{F}_2^n \to \mathbb{F}_2$. (This should not be confused with the real-degree/Fourier-degree. For example, $\deg_2(\chi_\alpha) = 1$ for all $\alpha \neq 0$.) Let us note some relations between $\dim(f)$ and $\mathrm{sp}(f)$. For any Boolean function f, we have

$$\deg_2(f) \leq \log \mathrm{sp}(f) \leq \dim(f), \tag{2}$$

except that the first inequality fails when $\deg_2(f) = 1$. (Throughout this paper, $\log$ always means $\log_2$.) The first inequality above is not difficult (see e.g. [8, Lemma 3]) and the second one is essentially immediate. Either of the above inequalities can be quite loose; for the first inequality, the inner product function on n variables has $\deg_2(f) = 2$ but $\log \mathrm{sp}(f) = n$. For the second inequality, the addressing function with $\frac{1}{2}\log s$ addressing variables and $s^{1/2}$ addressee variables can be shown to be s-sparse but has $\dim(f) \geq s^{1/2}$. (It is trivially true that $\dim(f) \leq s$ for any s-sparse function.)

We may rephrase these bounds as containments between classes of functions:

$$\{k\text{-dimensional}\} \subseteq \{2^k\text{-sparse}\} \subseteq \{\mathbb{F}_2 - \text{degree-}k\}\} \tag{3}$$

where the right containment is proper for $k > 1$ and the left is proper for k larger than some small constant such as 6. Alon et al. [2] gave essentially matching upper and lower bounds for testing the class of $\mathbb{F}_2$-degree-k functions, showing that $2^{\Theta(k)}$ nonadaptive queries are necessary and sufficient. We show that $2^{\Theta(k)}$ queries are also necessary and sufficient for testing each of the first two classes as well; in fact, by our implicit learning result, we can test *any* sub-class of k-dimensional functions using $2^{O(k)}$ queries.[1]

1.2 Our results and techniques

Testing Low-Dimensionality. We give nearly matching upper and lower bounds for testing whether a function is k-dimensional:

Theorem 1. [Testing k-dimensionality – informal] *There is a nonadaptive $O(k2^{2k}/\epsilon)$-query algorithm for ϵ-testing whether f is k-dimensional. Moreover, any algorithm (adaptive, even) for 0.49-testing this property must make $\Omega(2^{k/2})$ queries.*

We outline the basic idea behind our dimensionality test. Given $h \in \mathbb{F}_2^n$, we say that $f : \mathbb{F}_2^n \to \mathbb{R}$ is h-*invariant* if it satisfies $f(x + h) = f(x)$ for all $x \in \mathbb{F}_2^n$. We define the subspace $\mathrm{Inv}(f) = \{h : f \text{ is } h\text{-invariant}\}$. If f is truly k-dimensional, then $\mathrm{Inv}(f)$ has codimension k; we use this as the characterization of k-dimensional functions. We estimate the size of $\mathrm{Inv}(f)$ by randomly sampling vectors h and testing if they belong to $\mathrm{Inv}(f)$. We reject if the fraction of such h is much smaller than 2^{-k}. The crux of our soundness analysis is to show that if a function passes the test with good probability, most of its Fourier spectrum is concentrated on a k-dimensional subspace. From this we conclude that it must in fact be close to a k-dimensional function. Because of space constraints, this algorithm is omitted from this version of the paper.

Testing Sparsity. We next give an algorithm for testing whether a function is s-sparse. Its query complexity is $\mathrm{poly}(s)$, which is optimal up to the degree of the polynomial:

Theorem 2. [Testing s-sparsity – informal] *There is a nonadaptive $\mathrm{poly}(s, 1/\epsilon)$-query algorithm for ϵ-testing whether f is s-sparse. Moreover, any algorithm (adaptive, even) for 0.49-testing this property must make $\Omega(\sqrt{s})$ queries.*

The high-level idea behind our tester is that of "hashing" the Fourier coefficients, following [12]. We choose a random subspace H of $\mathbb{F}_2^n$ with codimension $O(s^2)$. This

[1] We remind the reader that efficient testability does not translate downward: if C_1 is a class of functions that is efficiently testable and $C_2 \subsetneq C_1$, the class C_2 need not be efficiently testable.

partitions all the Fourier coefficients into the cosets (affine subspaces) defined by H. If f is s-sparse, then each vector in $\mathrm{Spec}(f)$ is likely to land in a distinct coset. We define the "projection" of f to a coset $r + H$ to be the real-valued function given by zeroing out all Fourier coefficients not in $r + H$. Given query access to f, one can obtain approximate query access to a projection of f by a certain averaging. Now if each vector in $\mathrm{Spec}(f)$ is hashed to a different coset, then each projection function will have sparsity either 1 or 0, so we can try to test that at most s of the projection functions have sparsity 1, and the rest have sparsity 0.

A similar argument to the one used for k-dimensionality shows that if f passes this test, most of its Fourier mass lies on a few coefficients. However, unlike in the low-dimensionality test, this is not *a priori* enough to conclude that f is close to a sparse Boolean function. The obvious way to get a Boolean function close to f would be to truncate the Fourier spectrum to its s largest coefficients and then take the sign, but taking the sign could destroy the sparsity and give a function which is not at all sparse.

We circumvent this obstacle by using some new structural theorems about sparse Boolean functions. We show that if most of the Fourier mass of a function f lies on its largest s coefficients, then these coefficients are close to being "$\lceil \log s \rceil$–granular," i.e. close to integer multiples of $1/2^{\lceil \log s \rceil}$. We then prove that truncating the Fourier expansion to these coefficients and rounding them to nearby granular values gives a sparse *Boolean*-valued function (Theorem 6). Thus our sparsity test and its analysis depart significantly from the tests for juntas [14] and from our test for low-dimensionality.

Testing subclasses of k-dimensional functions. Finally, we show that a broad range of subclasses of k-dimensional functions are also testable with $2^{O(k)}$ queries. Recall that k-dimensional functions are all functions $f(x) = g(\chi_{\alpha_1}(x), \ldots, \chi_{\alpha_k}(x))$ where g is any k-variable Boolean function. We say that a class $\mathcal{C}$ is an *induced subclass of k-dimensional functions* if there is some collection $\mathcal{C}'$ of k-variable Boolean functions such that $\mathcal{C}$ is the class of all functions $f = g(\chi_{\alpha_1}, \ldots, \chi_{\alpha_k})$ where g is any function in $\mathcal{C}'$ and $\chi_{\alpha_1}, \ldots, \chi_{\alpha_k}$ are any linear functions from $\mathbb{F}_2^n$ to $\mathbb{F}_2$ as before. For example, let $\mathcal{C}$ be the class of all k-sparse polynomial threshold functions over $\{-1, 1\}^n$; i.e., each function in $\mathcal{C}$ is the sign of a *real* polynomial with at most k nonzero terms. This is an induced subclass of k-dimensional functions, corresponding to the collection $\mathcal{C}' = \{$ all linear threshold functions over k Boolean variables$\}$.

We show that any induced subclass of k-dimensional functions can be tested:

Theorem 3. [Testing induced subclasses of k-dimensional functions – informal] *Let $\mathcal{C}$ be any induced subclass of k-dimensional functions. There is a nonadaptive $\mathrm{poly}(2^k, 1/\epsilon)$-query algorithm for ϵ-testing $\mathcal{C}$.*

We note that the upper bound of Theorem 3 is essentially best possible in general, by the $2^{\Omega(k)}$ lower bound for testing the whole class of k-dimensional functions.

Our algorithm for Theorem 3 extends the approach of Theorem 2 with ideas from the "testing by implicit learning" work of [11]. Briefly, by hashing the Fourier coefficients we are able to construct a matrix of size $2^k \times 2^k$ whose entries are the values taken by the characters χ_α in the spectrum of f. This matrix, together with a vector of the corresponding values of f, serves as a data set for "implicit learning" (we say the learning is "implicit" since we do not actually know the names of the relevant characters). Our test inspects sub-matrices of this matrix and tries to find one which, together with the vector of f-values, matches the truth table of some k-variable function $g \in \mathcal{C}'$.

Organization of the paper. We give standard preliminaries and an explanation of our techniques for hashing the Fourier spectrum in Section 2. Section 3 gives our new structural theorems about sparse Boolean functions, and Section 4 uses these theorems to give our test for s-sparse functions. Because of space constraints, our results for testing k-dimensional functions, for unique-decoding, for testing induced subclasses of k-dimensional functions, and our lower bounds are given in the full version.

2 Preliminaries

Throughout the paper we view Boolean functions as mappings from $\mathbb{F}_2^n$ to $\{-1,1\}$. We will also consider functions which map from $\mathbb{F}_2^n$ to $\mathbb{R}$. Such functions have a unique Fourier expansion as in Equation (1). For $\mathcal{A}$ a collection of vectors $\alpha \in \mathbb{F}_2^n$, we write $\mathrm{wt}(\mathcal{A})$ to denote the "Fourier weight" $\mathrm{wt}(\mathcal{A}) = \sum_{\alpha \in \mathcal{A}} \hat{f}(\alpha)^2$ on the elements of A. This notation suppresses the dependence on f, but it will always be clear from context. We frequently use Parseval's identity: $\mathrm{wt}(\mathbb{F}_2^n) = \sum_{\alpha \in \mathbb{F}_2^n} \hat{f}(\alpha)^2 = \|f\|_2^2 \stackrel{\text{def}}{=} \mathbf{E}_{x \in \mathbb{F}_2^n}[f(x)^2]$. Here and elsewhere, an expectation or probability over "$x \in X$" refers to the uniform distribution on X.

As defined in the previous section, the sparsity of f is $\mathrm{sp}(f) = |\mathrm{Spec}(f)|$. We may concisely restate the definition of dimension as $\dim(f) = \dim(\mathrm{span}(\mathrm{Spec}(f)))$.

Given two Boolean functions f and g, we say that f and g are *ϵ-close* if $\mathbf{Pr}_{x \in \mathbb{F}_2^n}[f(x) \neq g(x)] \leq \epsilon$ and say they are *ϵ-far* if $\mathbf{Pr}_{x \in \mathbb{F}_2^n}[f(x) \neq g(x)] \geq \epsilon$. We use the standard definition of property testing:

Definition 1. *Let $\mathcal{C}$ be a class of functions mapping $\mathbb{F}_2^n$ to $\{-1,1\}$. A* property tester for *$\mathcal{C}$ is an oracle algorithm $\mathcal{A}$ which is given a distance parameter $\epsilon > 0$ and oracle access to a function $f : \mathbb{F}_2^n \to \{-1,1\}$ and satisfies the following conditions:*

1. *if $f \in \mathcal{C}$ then $\mathcal{A}$ outputs "accept" with probability at least $2/3$;*
2. *if f is ϵ-far from every $g \in \mathcal{C}$ then $\mathcal{A}$ outputs "accept" with probability at most $1/3$.*

We also say that $\mathcal{A}$ ϵ-tests $\mathcal{C}$. The main interest is in the number of queries *the testing algorithm makes.*

All of our testing upper and lower bounds allow "two-sided error" as described above. Our lower bounds are for adaptive query algorithms and our upper bounds are via nonadaptive query algorithms.

2.1 Projections of the Fourier spectrum

The idea of "isolating" or "hashing" Fourier coefficients by projection, as done in [12] in a learning-theoretic context, plays an important role in our tests.

Definition 2. *Given a subspace $H \leq \mathbb{F}_2^n$ and a coset $r + H$, define the* projection operator P_{r+H} *on functions $f : \mathbb{F}_2^n \to \mathbb{R}$ as follows:*

$$\widehat{\mathrm{P}_{r+H} f}(\alpha) \stackrel{\textit{def}}{=} \begin{cases} \widehat{f}(\alpha) & \textit{if } \alpha \in r+H, \\ 0 & \textit{otherwise.} \end{cases}$$

In other words, we have $\mathrm{P}_{r+H} f = A_{r+H} * f$, *where* $A_{r+H} \stackrel{def}{=} \sum_{\alpha \in r+H} \chi_\alpha$ *and* $*$ *is the convolution operator:* $f * g(x) = \mathbf{E}_y[f(x+y) \cdot g(y)]$.

Clearly $A_{r+H} = \chi_r \cdot \sum_{h \in H} \chi_h$, and it is a simple and well-known fact that $\sum_{h \in H} \chi_h = |H| \cdot \mathbf{1}_{H^\perp}$. Thus we conclude the following (see also Lemma 1 of [12]):

Fact 4 $\quad \mathrm{P}_{r+H} f(x) = \mathbf{E}_{y \in H^\perp}[\chi_r(y) f(x+y)]$.

We now show that for any coset $r+H$, we can approximately determine both $\mathrm{P}_{r+H} f(x)$ and $\|\mathrm{P}_{r+H} f\|_2^2$.

Proposition 1. *For any* $x \in \mathbb{F}_2^n$, *the value* $\mathrm{P}_{r+H} f(x)$ *can be estimated to within* $\pm\tau$ *with confidence* $1-\delta$ *using* $O(\log(1/\delta)/\tau^2)$ *queries to* f.

Proof. Empirically estimate the right-hand side in Fact 4. Since the quantity inside the expectation is bounded in $[-1, 1]$, the result follows from a Chernoff bound. □

Recall that $\mathrm{wt}(r+H) = \sum_{\alpha \in r+H} \hat{f}(\alpha)^2 = \|\mathrm{P}_{r+H} f\|_2^2$. We have:

Fact 5 $\quad \mathrm{wt}(r+H) = \mathbf{E}_{x \in \mathbb{F}_2^n, z \in H^\perp}[\chi_r(z) f(x) f(x+z)]$.

Proof. Using Parseval and Fact 4, we have

$$\mathrm{wt}(r+H) = \mathop{\mathbf{E}}_{w \in \mathbb{F}_2^n}[(\mathrm{P}_{r+H} f(w))^2] = \mathop{\mathbf{E}}_{w \in \mathbb{F}_2^n, y_1, y_2 \in H^\perp}[\chi_r(y_1) f(w+y_1) \chi_r(y_2) f(w+y_2)],$$

which reduces to the desired equality upon writing $x = w + y_1$, $z = y_1 + y_2$. □

Proposition 2. *The value* $\mathrm{wt}(r+H)$ *can be estimated to within* $\pm\tau$ *with confidence* $1-\delta$ *using* $O(\log(1/\delta)/\tau^2)$ *queries to* f.

Proof. Empirically estimate the right-hand side in Fact 5. Since the quantity inside the expectation is bounded in $[-1, 1]$, the result follows from a Chernoff bound. □

2.2 Hashing to a random coset structure

In this section we present our technique for pairwise independently hashing the Fourier characters.

Definition 3. *For* $t \in \mathbb{N}$, *we define a* random t-dimensional coset structure $(H, \mathcal{C})$ *as follows: We choose vectors* $\beta_1, \ldots, \beta_t \in \mathbb{F}_2^n$ *independently and uniformly at random and set* $H = \mathrm{span}\{\beta_1, \ldots, \beta_t\}^\perp$. *For each* $b \in \mathbb{F}_2^t$ *we define the "bucket"*

$$C(b) \stackrel{def}{=} \{\alpha \in \mathbb{F}_2^n : \langle \alpha, \beta_i \rangle = b_i \text{ for all } i\}.$$

We take $\mathcal{C}$ *to be the set of* $C(b)$*'s, which has cardinality* 2^t.

Remark 1. Given such a random coset structure, if the β_i's are linearly independent then the buckets $C(b)$ are precisely the cosets in $\mathbb{F}_2^n/H$, and the coset-projection function $\mathrm{P}_{C(b)} f$ is defined according to Definition 2. In the (usually unlikely) case that the β_i's are linearly *dependent*, some of the $C(b)$'s will be cosets in $\mathbb{F}_2^n/H$ and some of them will be empty. For the empty buckets $C(b)$ we define $\mathrm{P}_{C(b)} f$ to be identically 0. It is algorithmically easy to distinguish empty buckets from genuine coset buckets.

We now derive some simple but important facts about this random hashing process:

Proposition 3. *Let $(H, \mathcal{C})$ be a random t-dimensional coset structure. Define the indicator random variable $I_{\alpha \to b}$ for the event that $\alpha \in C(b)$.*

1. *For each $\alpha \in \mathbb{F}_2^n \setminus \{0\}$ and each b we have $\mathbf{Pr}[\alpha \in C(b)] = \mathbf{E}[I_{\alpha \to b}] = 2^{-t}$.*
2. *Let $\alpha, \alpha' \in \mathbb{F}_2^n$ be distinct. Then $\mathbf{Pr}[\alpha, \alpha'$ belong to the same bucket$] = 2^{-t}$.*
3. *Fix any set $S \subseteq \mathbb{F}_2^n$ with $|S| \leq s + 1$. If $t \geq 2\log s + \log(1/\delta)$ then except with probability at most δ, all vectors in S fall into different buckets.*
4. *For each b, the collection of random variables $(I_{\alpha \to b})_{\alpha \in \mathbb{F}_2^n}$ is pairwise independent.*

Proof. Part 1 is because for any $\alpha \neq 0$, each $\langle \alpha, \beta_i \rangle$ is an independent uniformly random bit. Part 2 is because each $\langle \alpha - \alpha', \beta_i \rangle$ is an independent uniformly random bit, and hence the probability that $\langle \alpha, \beta_i \rangle = \langle \alpha', \beta_i \rangle$ for all i is 2^{-t}. Part 3 follows from Part 2 and taking a union bound over the at most $\binom{s+1}{2} \leq s^2$ distinct pairs in S. For Part 4, assume first that $\alpha \neq \alpha'$ are both nonzero. Then from the fact that α and α' are linearly independent, it follows that $\mathbf{Pr}[\alpha, \alpha' \in C(b)] = 2^{-2t}$ as required. On the other hand, if one of $\alpha \neq \alpha'$ is zero, then $\mathbf{Pr}[\alpha, \alpha' \in C(b)] = \mathbf{Pr}[\alpha \in C(b)]\mathbf{Pr}[\alpha' \in C(b)]$ follows immediately by checking the two cases $b = 0$, $b \neq 0$. □

With Proposition 3 in mind, we give the following simple deviation bound for the sum of pairwise independent random variables:

Proposition 4. *Let $X = \sum_{i=1}^n X_i$, where the X_i's are pairwise independent random variables satisfying $0 \leq X_i \leq \tau$. Assume $\mu = \mathbf{E}[X] > 0$. Then for any $\epsilon > 0$, we have $\mathbf{Pr}[X \leq (1-\epsilon)\mu] \leq \frac{\tau}{\epsilon^2 \mu}$.*

Proof. By pairwise independence, we have $\mathbf{Var}[X] = \sum \mathbf{Var}[X_i] \leq \sum \mathbf{E}[X_i^2] \leq \sum \tau \mathbf{E}[X_i] = \tau\mu$. The result now follows from Chebyshev's inequality. □

Finally, it is slightly annoying that Part 1 of Proposition 3 fails for $\alpha = 0$ (because 0 is always hashed to $C(0)$). However we can easily handle this issue by renaming the buckets with a simple random permutation.

Definition 4. *In a random* permuted *t-dimensional coset structure, we additionally choose a random $z \in \mathbb{F}_2^t$ and rename $C(b)$ by $C(b+z)$.*

Proposition 5. *For a random permuted t-dimensional coset structure, Proposition 3 continues to hold, with Part 1 even holding for $\alpha = 0$.*

Proof. Use Proposition 3 and the fact that adding a random z permutes the buckets. □

3 Structural theorems about s-sparse functions

In this section we prove structural theorems about close-to-sparse Boolean functions. These theorems are crucial to the analysis of our test for s-sparsity; we also present a learning application in the full version.

Definition 5. *Let $B = \{\alpha_1, \cdots, \alpha_s\}$ denote the (subsets of $[n]$ with the) s largest Fourier coefficients of f, and let $S = \bar{B}$ be its complement. We say that f is μ-*close to s-sparse in ℓ_2 *if $\sum_{\alpha \in S} \hat{f}(\alpha)^2 \leq \mu^2$.*

Definition 6. *We say a rational number* has granularity $k \in \mathbb{N}$*, or is k-*granular*, if it is of the form $(integer)/2^k$. We say a function $f : \mathbb{F}_2^n \to \mathbb{R}$ is k-*granular *if $\widehat{f}(\alpha)$ is k-granular for every α. We say that a number v is μ-*close to k-granular *if $|v - j/2^k| \leq \mu$ for some integer j.*

The following structural result is the key theorem for the completeness of our sparsity test; it says that in any function that is close to being sparse in ℓ_2, all the large Fourier coefficients are close to being granular.

Theorem 1 *[Completeness Theorem.] If f is μ-close to s-sparse in ℓ_2, then each $\hat{f}(\alpha)$ for $\alpha \in B$ is $\frac{\mu}{\sqrt{s}}$-close to $\lceil \log s \rceil$-granular.*

Proof. Pick a set of $k = \lceil \log s \rceil + 1$ equations $A\alpha = b$ at random (i.e. pick a $k \times n$ random matrix A and a random vector $b \in \mathbb{F}_2^k$). Let $A^\perp \subset \mathbb{F}_2^n$ be the set of solutions to $A\alpha = 0$. Define H to be the coset of $A^\perp$ of solutions to $A\alpha = b$. We have

$$\mathrm{P}_H f(x) = \sum_{\alpha \in H} \hat{f}(\alpha)\chi_\alpha(x).$$

Fix $\alpha_i \in B$. We will show that with non-zero probability the following two events happen together: the vector α_i is the unique coefficient in $B \cap H$, and the ℓ_2 Fourier mass of the set $S \cap H$ is bounded by $\frac{\mu^2}{s}$. Clearly, $\mathbf{Pr}_{A,b}[A\alpha_i = b] = 2^{-k}$. Let us condition on this event. By pairwise independence, for any $j \neq i$, $\mathbf{Pr}_{A,b}[A\alpha_j = b | A\alpha_i = b] = 2^{-k} \leq \frac{1}{2s}$. Thus $\mathbf{E}_{A,b}\left[|\{j \neq i \text{ such that } A\alpha_j = b\}| \,\middle|\, A\alpha_i = b\right] = \frac{(s-1)}{2^k} < \frac{1}{2}$. Hence by Markov's inequality

$$\mathbf{Pr}_{A,b}[\exists j \neq i \text{ such that } A\alpha_j = b \,|\, A\alpha_i = b] < \frac{1}{2}. \tag{4}$$

Now consider the coefficients from S. We have

$$\mathbf{E}_{A,b}\left[\sum_{\beta \in S \cap H} \hat{f}(\beta)^2 \big| A\alpha_i = b\right] = \sum_{\beta \in S} \mathbf{Pr}[\beta \in H | A\alpha_i = b]\hat{f}(\beta)^2 \leq 2^{-k}\mu^2 \leq \frac{\mu^2}{2s}.$$

Hence by Markov's inequality,

$$\mathbf{Pr}_{A,b}\left[\sum_{\beta \in S \cap H} \hat{f}(\beta)^2 \geq \frac{\mu^2}{s} \big| A\alpha_i = b\right] \leq \frac{1}{2}. \tag{5}$$

Thus by applying the union bound to Equations 4 and 5, we have both the desired events (α_i being the unique solution from B, and small ℓ_2 mass from S) happening with non-zero probability over the choice of A, b. Fixing this choice, we have

$$\mathrm{P}_H f(x) = \hat{f}(\alpha_i)\chi_{\alpha_i}(x) + \sum_{\beta \in S \cap H} \hat{f}(\beta)\chi_\beta(x) \text{ where } \sum_{\beta \in S \cap H} \hat{f}(\beta)^2 \leq \frac{\mu^2}{s}.$$

But by Fact 4 we also have $\mathrm{P}_H f(x) = \mathbf{E}_{y\in A}[\chi_b(y) f(x+y)]$ (here we abuse notations and think of A as both the matrix A and the space spanned by the rows of A. In particular, $A = (A^\perp)^\perp$). Thus the function $\mathrm{P}_H f(x)$ is the average of a Boolean function over 2^k points, hence it is $(k-1)$-granular.

We now consider the function $g(x) = \sum_{\beta\in S\cap H} \hat{f}(\beta)\chi_\beta(x)$. Since $\mathbf{E}_x[g(x)^2] \leq \frac{\mu^2}{s}$, for some $x_0 \in \mathbb{F}_2^n$ we have $g(x_0)^2 \leq \frac{\mu^2}{s}$, hence $|g(x_0)| \leq \frac{\mu}{\sqrt{s}}$. Fixing this x_0, we have $\mathrm{P}_H f(x_0) = \hat{f}(\alpha_i)\chi_{\alpha_i}(x_0) + g(x_0)$, and hence $|\hat{f}(\alpha_i)| = |\mathrm{P}_H f(x_0) - g(x_0)|$. Since $\mathrm{P}_H f(x_0)$ is $(k-1)$-granular and $|g(x_0)| \leq \frac{\mu}{\sqrt{s}}$, the claim follows. $\square$

Thus, if f has its Fourier mass concentrated on s coefficients, then it is close in ℓ_2 to an s-sparse, $\lceil \log s \rceil$ granular real-valued function. We next show that this real-valued function must in fact be Boolean.

Theorem 6. *[Soundness Theorem.] Let $f : \mathbb{F}_2^n \to \{-1, 1\}$ be $\mu \leq \frac{1}{20s^2}$ close to s-sparse in ℓ_2. Then there is an s-sparse Boolean function $F : \mathbb{F}_2^n \to \{-1, 1\}$ within Hamming distance $\frac{\mu^2}{2}$ from f.*

Proof. Let $B = \{\alpha_1, \cdots, \alpha_s\}$ be the s largest Fourier coefficients of f and let $k = \lceil \log s \rceil$. By Theorem 1, each $\hat{f}(\alpha_i)$ is $\frac{\mu}{\sqrt{s}}$ close to k-granular. So we can write

$$\hat{f}(\alpha_i) = \hat{F}(\alpha_i) + \hat{G}(\alpha_i)$$

where $\hat{F}(\alpha_i)$ is k-granular and $|\hat{G}(\alpha_i)| \leq \frac{\mu}{\sqrt{s}}$. Set $\hat{F}(\beta) = 0$ and $\hat{G}(\beta) = \hat{f}(\beta)$ for $\beta \in S = \bar{B}$. Thus we have $f(x) = F(x) + G(x)$, further F is s-sparse and k-granular, while

$$\mathbf{E}[G(x)^2] \leq s\frac{\mu^2}{s} + \mu^2 \leq 2\mu^2.$$

It suffices to show that F's range is $\{-1, 1\}$. In this case, G's range must be $\{-2, 0, 2\}$, the value $G(x)^2$ is exactly 4 whenever f and F differ, and therefore f and F satisfy

$$\mathbf{Pr}_x[f(x) \neq F(x)] = \mathbf{Pr}[|G(x)| = 2] = \frac{1}{4}\mathbf{E}_x[G(x)^2] \leq \frac{\mu^2}{2}.$$

As f is a Boolean function on $\mathbb{F}_2^n$ we have

$$1 = f^2 = F^2 + 2FG + G^2 = F^2 + G(2f - G). \tag{6}$$

Writing $H = G(2f - G)$, from Fact 7 below we have that for all α,

$$|\widehat{H}(\alpha)| \leq \|G\|_2 \|2f - G\|_2 \leq \|G\|_2(\|2f\|_2 + \|G\|_2) \leq 2\sqrt{2}\mu + 2\mu^2 < 4\mu \leq \frac{1}{5s^2}.$$

On the other hand, since F has granularity k it is easy to see that F^2 has granularity $2k$; in particular, $|\widehat{F^2}(\alpha)|$ is either an integer or at least $2^{-2k} \geq \frac{1}{4s^2}$-far from being an integer. But for (6) to hold as a functional identity, we must have $\widehat{F^2}(0) + \widehat{H}(0) = 1$ and $\widehat{F^2}(\alpha) + \widehat{H}(\alpha) = 0$ for all $\alpha \neq 0$. It follows then that we must have $\widehat{F^2}(0) = 1$ and $\widehat{F^2}(\alpha) = 0$ for all $\alpha \neq 0$; i.e., $F^2 = 1$ and hence F has range $\{-1, 1\}$, as claimed. $\square$

Fact 7 *Let $f, g : \mathbb{F}_2^n \to \mathbb{R}$. Then $|\widehat{fg}(\alpha)| \le \|f\|_2 \|g\|_2$ for every α.*

Proof. Using Cauchy-Schwarz and Parseval,

$$|\widehat{fg}(\alpha)| = |\sum_{\beta} \widehat{f}(\beta)\widehat{g}(\alpha+\beta)| \le \sqrt{\sum_{\beta} \widehat{f}(\beta)^2}\sqrt{\sum_{\beta} \widehat{g}(\alpha+\beta)^2} = \|f\|_2\|g\|_2. \qquad \square$$

4 Testing s-sparsity

The following is our algorithm for testing whether $f : \mathbb{F}_2^n \to \{-1, 1\}$ is s-sparse:

Algorithm 1 Testing s-sparsity

Inputs: s, ϵ

Parameters: $\mu = \min(\sqrt{2\epsilon}, \frac{1}{20s^2})$, $t = \lceil 2\log s + \log 100 \rceil$, $\tau = \frac{\mu^2}{100 \cdot 2^t}$.

1. Choose a random permuted t-dimensional coset structure $(H, \mathcal{C})$.
2. For each bucket $C \in \mathcal{C}$, estimate $\mathrm{wt}(C) = \sum_{\alpha \in C} \hat{f}(\alpha)^2$ to accuracy $\pm\tau$ with confidence $1 - (1/100)2^{-t}$, using Proposition 2.
3. Let $\mathcal{L}$ be the set of buckets where the estimate is at least 2τ. If $|\mathcal{L}| \ge s+1$, reject.

Roughly speaking, Step 1 pairwise independently hashes the Fourier coefficients of f into $\Theta(s^2)$ buckets. If f is s-sparse then at most s buckets have nonzero weight and the test accepts. On the other hand, if f passes the test with high probability then we show that almost all the Fourier mass of f is concentrated on at most s nonzero coefficients (one for each bucket in $\mathcal{L}$). Theorem 6 now shows that f is close to a sparse function. Our theorem about the test is the following:

Theorem 8. *Algorithm 1 ϵ-tests whether $f : \mathbb{F}_2^n \to \{-1, 1\}$ is s-sparse (with confidence $3/4$), making $O\left(\frac{s^6 \log s}{\epsilon^2} + s^{14}\log s\right)$ nonadaptive queries.*

The query complexity of Theorem 8 follows immediately from Proposition 2 and the fact that there are $2^t = O(s^2)$ buckets. In the remainder of this section we present the completeness (Lemma 1) and the soundness (Lemma 4) of the test. We begin with the completeness, which is straightforward.

Lemma 1. *If f is s-sparse then the test accepts with probability at least 0.9.*

Proof. Write $f = \sum_{i=1}^{s'} \hat{f}(\alpha_i)\chi_{\alpha_i}$, where each $\hat{f}(\alpha_i) \ne 0$ and $s' \le s$. Since there are 2^t buckets, all of the estimates in Step 2 are indeed τ-accurate, except with probability at most $1/100$. If the estimates are indeed accurate, the only buckets with weight at least τ are those that contain a nonzero Fourier coefficient, which are at most s in number. So f passes the test with probability at least 0.9. $\square$

We now analyze the soundness. We partition the Fourier coefficients of f into two sets: B of big coefficients and S of small coefficients. Formally, let

$$B \stackrel{\text{def}}{=} \{\alpha : \hat{f}(\alpha)^2 \geq 3\tau\}, \qquad S \stackrel{\text{def}}{=} \{\alpha : \hat{f}(\alpha)^2 < 3\tau\}.$$

We observe that if there are too many big coefficients the test will probably reject:

Lemma 2. *If* $|B| \geq s+1$ *then the test rejects with probability at least* $3/4$.

Proof. Proposition 5(3) implies that after Step 1, except with probability at most $1/100$ there are at least $s+1$ buckets C containing an element of B. In Step 2, except with probability at most $1/100$, we get an estimate of at least $3\tau - \tau \geq 2\tau$ for each such bucket. Then $|\mathcal{L}|$ will be at least $s+1$ in Step 3. Hence the overall rejection probability is at least $1 - 2/100$. $\square$

Next we show that if the weight on small coefficients, $\mathrm{wt}(S) = \sum_{\alpha \in S} \hat{f}(\alpha)^2$, is too large then the test will probably reject:

Lemma 3. *If* $\mathrm{wt}(S) \geq \mu^2$ *then the test rejects with probability at least* $3/4$.

Proof. Suppose that indeed $\mathrm{wt}(S) \geq \mu^2$. Fix a bucket index b and define the random variable $M_b := \mathrm{wt}(C(b) \cap S) = \sum_{\alpha \in C(b) \cap S} \hat{f}(\alpha)^2 = \sum_{\alpha \in S} \hat{f}(\alpha)^2 \cdot I_{\alpha \to b}$. Here the randomness is from the choice of $(H, \mathcal{C})$, and we have used the pairwise independent indicator random variables defined in Proposition 5(3). Let us say that the bucket $C(b)$ is *good* if $M_b \geq \frac{1}{2}\mathbf{E}[M_b]$. We have $\mathbf{E}[M_b] = 2^{-t}\,\mathrm{wt}(S) \geq 100\tau > 0$, and by Proposition 4 we deduce $\mathbf{Pr}[M_b \leq \frac{1}{2}\mathbf{E}[M_b]] \leq \frac{3\tau}{(1/2)^2\mathbf{E}[M_b]} \leq 3/25$. Thus the expected fraction of bad buckets is at most $3/25$, so by Markov's inequality there are at most $(3/5)2^t$ bad buckets except with probability at most $1/5$. But if there are at least $(2/5)2^t$ good buckets, we have at least $(2/5)(100s^2) \geq s+1$ buckets b with $\mathrm{wt}(C(b) \cap S) \geq \frac{1}{2}\mathbf{E}[M_b] \geq 50\tau$. Assuming all estimates in Step 2 of the test are accurate to within $\pm\tau$ (which fails with probability at most $1/100$), Step 3 of the test will reject. Thus we reject except with probability at most $1/5 + 1/100 < 1/4$. $\square$

Now we put together the pieces to establish soundness of the test:

Lemma 4. *Suppose the test accepts* f *with probability exceeding* $1/4$*. Then* f *is* ϵ*-close to an* s*-sparse Boolean function.*

Proof. Assuming the test accepts f with probability exceeding $1/4$, by Lemma 2 we have $|B| \leq s$, by Lemma 3 we have $\mathrm{wt}(S) \leq \mu^2$. Thus f is $\mu \leq \frac{1}{20s^2}$ close in ℓ_2 to being s-sparse. We now apply the soundness theorem, Theorem 6 to conclude that f must be $\frac{\mu^2}{2} \leq \epsilon$-close in Hamming distance to an s-sparse Boolean function. $\square$

References

1. ALON, N., FISCHER, E., NEWMAN, I., AND SHAPIRA, A. A combinatorial characterization of the testable graph properties: It's all about regularity. In *Proc. STOC* (2006).

2. ALON, N., KAUFMAN, T., KRIVELEVICH, M., LITSYN, S., AND RON, D. Testing low-degree polynomials over GF(2). In *Proc. RANDOM* (2003), pp. 188–199.
3. ALON, N., AND SHAPIRA, A. A characterization of the (natural) graph properties testable with one-sided error. In *Proc. FOCS* (2005), pp. 429–438.
4. ALON, N., AND SHAPIRA, A. Every monotone graph property is testable. In *Proc. STOC* (2005), pp. 128–137.
5. AUSTIN, T., AND TAO, T. On the testability and repair of hereditary hypergraph properties. *Submitted to Random Structures and Algorithms* (2008).
6. BELLARE, M., COPPERSMITH, D., HASTAD, J., KIWI, M., AND SUDAN, M. Linearity testing in characteristic two. *IEEE Trans. on Information Theory 42*, 6 (1996), 1781–1795.
7. BELLARE, M., GOLDREICH, O., AND SUDAN, M. Free bits, pcps and non-approximability-towards tight results. *SIAM J. Comput. 27(3)* (1998), 804–915.
8. BERNASCONI, A., AND CODENOTTI, B. Spectral analysis of boolean functions as a graph eigenvalue problem. *IEEE Trans. Computers 48*, 3 (1999), 345–351.
9. BLAIS, E. Improved bounds for testing juntas. In *Proc. RANDOM* (2008), pp. 317–330.
10. BLUM, M., LUBY, M., AND RUBINFELD, R. Self-testing/correcting with applications to numerical problems. *J. Comp. Sys. Sci. 47* (1993), 549–595. Earlier version in STOC'90.
11. DIAKONIKOLAS, I., LEE, H., MATULEF, K., ONAK, K., RUBINFELD, R., SERVEDIO, R., AND WAN, A. Testing for concise representations. In *Proc. FOCS* (2007), pp. 549–558.
12. FELDMAN, V., GOPALAN, P., KHOT, S., AND PONNUSWAMI, A. New results for learning noisy parities and halfspaces. In *Proc. FOCS* (2006), pp. 563–576.
13. FISCHER, E. The art of uninformed decisions: A primer to property testing. *Bulletin of the European Association for Theoretical Computer Science 75* (2001), 97–126.
14. FISCHER, E., KINDLER, G., RON, D., SAFRA, S., AND SAMORODNITSKY, A. Testing juntas. *J. Computer & System Sciences 68*, 4 (2004), 753–787.
15. JACKSON, J. An efficient membership-query algorithm for learning DNF with respect to the uniform distribution. *Journal of Computer and System Sciences 55* (1997), 414–440.
16. KAUFMAN, T., AND SUDAN, M. Sparse random linear codes are locally decodable and testable. In *Proc. FOCS* (2007), pp. 590–600.
17. KAUFMAN, T., AND SUDAN, M. Algebraic property testing: the role of invariance. In *Proc. 40th Annual ACM Symposium on Theory of Computing (STOC)* (2008), pp. 403–412.
18. KUSHILEVITZ, E., AND MANSOUR, Y. Learning decision trees using the fourier spectrum. *SIAM Journal on Computing 22*, 6 (Dec. 1993), 1331–1348.
19. LINIAL, N., MANSOUR, Y., AND NISAN, N. Constant depth circuits, Fourier transform and learnability. *Journal of the ACM 40*, 3 (1993), 607–620.
20. MATULEF, K., O'DONNELL, R., RUBINFELD, R., AND SERVEDIO, R. Testing halfspaces. In *Proc. SODA* (2009), pp. 256–264.
21. PARNAS, M., RON, D., AND SAMORODNITSKY, A. Testing basic boolean formulae. *SIAM J. Disc. Math. 16* (2002), 20–46.
22. SAMORODNITSKY, A. Low-degree tests at large distances. In *Proc.* 39^{th} *ACM Symposium on the Theory of Computing (STOC'07)* (2007), pp. 506–515.