Learning symmetric k-juntas in time $n^{o(k)}$*

Mihail N. Kolountzakis† Evangelos Markakis‡ Aranyak Mehta§

April 2005

Abstract

We give an algorithm for learning symmetric k-juntas (boolean functions of n boolean variables which depend only on an unknown set of k of these variables) in the PAC model under the uniform distribution, which runs in time $n^{O(k/\log k)}$. Our bound is obtained by proving the following result: Every symmetric boolean function on k variables, except for the parity and the constant functions, has a non-zero Fourier coefficient of order at least 1 and at most $O(k/\log k)$. This improves the previously best known bound of $3k/31$ [11], and provides the first $n^{o(k)}$ time algorithm for learning symmetric juntas.

1 Introduction

We consider a fundamental problem in computational learning theory: learning in the presence of irrelevant information. One formalization of the problem is as follows: We want to learn an unknown boolean function of n variables, which depends only on $k \ll n$ variables (typically k is $O(\log n)$). We call such a function a k-junta. We are provided with a set of labelled examples $\langle x, f(x)\rangle$, where the x's are picked uniformly and independently at random from the domain $\{0,1\}^n$ (this is the PAC model with uniform distribution). We wish to identify the k relevant variables and the truth table of the function.

The problem was first posed by Blum [1] and Blum and Langley [4], and it is considered [2, 13] to be one of the most important open problems in the theory of uniform distribution learning. It has connections with learning DNF formulas and decision trees of super-constant size, see [5, 8, 12, 15, 16] for details. The general case is believed to be hard and has even been used to propose a cryptosystem [3]. A trivial algorithm runs in time roughly n^k by doing an exhaustive search over all possible sets of relevant variables. Two important classes of juntas are learnable in polynomial time: parity and monotone functions. Learning parity functions can be reduced to solving a system of linear equations over $\mathbb{F}_2$ [7]. Monotone functions have non-zero singleton Fourier coefficients (e.g., see [13]). For the general case, the first significant breakthrough was given in [13] - learning with confidence $1-\delta$ in time $n^{0.7k}\text{poly}(2^k, n, \log 1/\delta)$. Note that we allow the running

*This work was done when all authors were at the Georgia Institute of Technology.

†School of Mathematics, Georgia Institute of Technology, Atlanta GA 30332, USA, and Department of Mathematics, Univ. of Crete, GR-71409 Iraklio, Greece. E-mail: kolount@gmail.com. Partially supported by European Commission IHP Network HARP (Harmonic Analysis and Related Problems), Contract Number: HPRN-CT-2001-00273 - HARP, and by grant INTAS 03-51-5070 (2004) (Analytical and Combinatorial Methods in Number Theory and Geometry).

‡University of Toronto, Department of Computer Science, Toronto, ON M5S3G4, Canada, E-mail: vangelis@cs.toronto.edu

§IBM Almaden Research Center, 650 Harry Rd, San Jose, CA 95120, USA, E-mail: mehtaa@us.ibm.com

time to be polynomial in 2^k, since this is the size of the truth-table which is output. In the typical setting of $k = O(\log n)$, this becomes polynomial in n.

In this paper we consider the class of *symmetric* k-juntas, functions which are symmetric on their relevant variables. The only non-trivial algorithm known for this case is the standard Fourier based algorithm, described in Section 2. The analysis of the running time of this algorithm reduces to the following question:

> What is the smallest t such that every symmetric boolean function on k variables, which is not a constant or a parity function, has a non-zero Fourier coefficient of order at least 1 and at most t?

A bound of t_0 implies a running time of roughly n^{t_0}. A bound of $\frac{2k}{3}$ was provided in [13]. This was improved to $\frac{3k}{31}$ in [11]. Here we show a bound of $O(k/\log k)$ (Theorem 3.3), giving the first algorithm for learning symmetric k-juntas in time $n^{o(k)}$.

Techniques

Our techniques involve a mix of number theory, combinatorics and probability. We start by reducing our problem to finding 0/1 solutions to a system of Diophantine equations involving binomial coefficients, as in [11]. We then take a departure from [11] by further reducing this to the problem of showing that a certain integer-valued polynomial P is constant over the set $\{0, 1, ..., k\}$. We manage to prove this in two steps: First, we show that P is constant over the union of two small intervals $\{0, ..., t\} \cup \{k - t, ..., k\}$. This is obtained by looking at P modulo carefully chosen prime numbers. To choose these prime numbers we use the Siegel-Walfisz theorem on the density of primes in arithmetic progressions with modulus of moderate growth. In the second step, we extend the constant nature of P to the whole interval $\{0, ..., k\}$ by repeated applications of Lucas' Theorem. One additional interesting aspect of our proof is the use of an equivalence between a) the vanishing of Fourier coefficients and b) the equality of moments of certain random variables under the uniform measure on the hypercube and under the measure defined by the function itself. This equivalence helps us eliminate a lot of case analysis.

2 Preliminaries

Symmetric Juntas

Given a boolean function f on n variables $x_1, ..., x_n$, we will say that x_i is a *relevant* variable for f if there exist $x, y \in \{0, 1\}^n$ which differ only in the i-th coordinate and $f(x) \neq f(y)$. Variables that are not relevant are called irrelevant. We will call f a k-junta if f has at most k relevant variables.

We consider the class of symmetric juntas. A boolean function $f : \{0, 1\}^k \to \{0, 1\}$ on k variables is a symmetric function if for any permutation $\pi \in S_k$, $f(x_1, ..., x_k) = f(\pi(x_1), ..., \pi(x_k))$. Hence the value of f at $(x_1, ..., x_k)$ depends only on the *weight* of $(x_1, ..., x_k)$, which is the number of variables that are set to 1. A symmetric k-junta is a function on n variables which is symmetric on the k variables it depends on.

We will describe a symmetric boolean function on k variables by a $(k + 1)$-bit string $f_0 f_1 ... f_k$, where f_i is the value of f on an input of weight i. The following four special symmetric functions on k variables will appear often: the two constant functions $\mathbf{0}$ and $\mathbf{1}$, the parity function $\oplus$, and its complement $\overline{\oplus}$.

Learning in the PAC model

We consider the PAC learning model [14], in which we wish to learn a *Concept Class* $\mathcal{C} = \bigcup_n \mathcal{C}_n$, where each $\mathcal{C}_n$ is a collection of boolean functions from $\{0,1\}^n \to \{0,1\}$. In our case, $\mathcal{C}_n$ is the class of symmetric k-juntas on n variables. Let ϵ be an *accuracy parameter* and δ a *confidence parameter*. A learning algorithm $\mathcal{A}$ for $\mathcal{C}$ has access to an *oracle* for $f \in \mathcal{C}_n$. A query to the oracle outputs a labeled example $\langle x, f(x)\rangle$, where x is drawn from $\{0,1\}^n$ according to some probability distribution $\mathcal{D}$. $\mathcal{A}$ is said to be a learning algorithm for the class $\mathcal{C}$ under the distribution $\mathcal{D}$ if for all $f \in \mathcal{C}$, it outputs, with probability at least $1-\delta$, a hypothesis h such that $\Pr_x[h(x) = f(x)] \geq 1-\epsilon$. We will be concerned only with the uniform distribution and we will obtain an algorithm with accuracy parameter $\epsilon = 0$, i.e., we identify the exact function f.

Fourier Transform

We will consider functions of the form: $f : \{0,1\}^n \to \mathbb{R}$. An orthonormal basis for the functions defined on the Boolean cube can be given by the *characters* of the group $\mathbb{Z}_2^n$. In particular, for every $S \subseteq \{1,...,n\}$, define the following function:

$$\chi_S(x) = (-1)^{\sum_{i \in S} x_i}.$$

Any real-valued function on the Boolean cube can be expressed as a linear combination of the functions χ_S. Given f, we have that $f(x) = \sum_S \hat{f}(S)\chi_S(x)$, where $\hat{f}(S)$ is the Fourier coefficient of f at S and is equal to the inner product of f with χ_S:

$$\hat{f}(S) = \frac{1}{2^n} \sum_{x \in \{0,1\}^n} f(x)\chi_S(x).$$

Fourier-based Learning

Let f be a k-junta. It is known that we can exactly calculate the Fourier coefficients of f in the uniform distribution PAC model, with confidence $1-\delta$ in time $poly(2^k, n, \log\frac{1}{\delta})$, using standard Chernoff-Hoeffding bounds (see [10, 13]). Observe further, that if x_i is an irrelevant variable for a k-junta f, then for any $S \subseteq \{x_1,...,x_n\}$ containing x_i, $\hat{f}(S) = 0$. Hence if $\hat{f}(S) \neq 0$, for some S, then S contains only relevant variables.

This suggests the following algorithm: Starting with $l = 1$, compute the Fourier coefficients of all subsets of $\{x_1,...,x_n\}$ of size l. Collect the union of all relevant variables that correspond to subsets with non-zero Fourier coefficients. Stop as soon as you collect all k relevant variables.

Since the function is symmetric, for any two sets S, T of relevant variables such that $|S| = |T|$, we have $\hat{f}(S) = \hat{f}(T)$. Hence the first time that we will identify some relevant variables in the algorithm, we will actually be able to identify all the relevant variables. Once we find the relevant variables, finding the truth-table of the function can be done in time $poly(2^k, \log\frac{1}{\delta})$.

The above algorithm would take time roughly n^k for $f \in \{\mathbf{0}, \mathbf{1}, \oplus, \overline{\oplus}\}$. However, these particular functions are well known to be learnable in time $poly(n, \log\frac{1}{\delta})$. Hence the following is true:

Fact 2.1. *If every symmetric function $f \notin \{\mathbf{0}, \mathbf{1}, \oplus, \overline{\oplus}\}$ has a non-zero Fourier coefficient of order between 1 and t, then we can learn symmetric k-juntas in time n^t $poly(2^k, n, \log\frac{1}{\delta})$.*

3 Main Section

3.1 An Equivalent Formulation

We state an equivalent condition for the existence of a non-zero Fourier coefficient of a boolean function f, as proved in [11]. Let $f : \{0,1\}^k \to \{0,1\}$ be a boolean function. For a vector $\mathbf{x} = (x_1, \ldots, x_k)$, and a set $S \subseteq [k]$, let $\mathbf{x}_S$ be the projection of $\mathbf{x}$ on the indices of S. Let $\sigma \in \{0,1\}^{|S|}$. Define the following probabilities:

$$p_{S,\sigma}(f) := \Pr\left[f(\mathbf{x}) = 1 \mid \mathbf{x}_S = \sigma\right]$$

Unless mentioned, all probabilities are over the uniform distribution on $\{0,1\}^k$. For $t \geq 1$, call a boolean function f on k variables *t-null*, if for all sets $S \subseteq [k]$, with $|S| = t$, and for all $\sigma \in \{0,1\}^t$, the probabilities $p_{S,\sigma}(f)$ are all equal to each other. The following lemma reveals the connection with the Fourier coefficients of f.

Lemma 3.1. *[11] Let f be a boolean function on k variables. Then f is t-null for some $1 \leq t \leq k$, if and only if, for all $\emptyset \neq S \subseteq [k]$ with cardinality at most t, $\hat{f}(S) = 0$.*

It is clear that if $s \leq t$ and f is t-null then it is also s-null.

When we consider the case of symmetric functions, $p_{S,\sigma}(f)$ just depends on $t := |S|$ and the weight w of σ. We denote this by $p_{t,w}(f)$. It is clear that:

$$p_{t,w}(f) = \frac{1}{2^{k-t}} \sum_{i=0}^{k} f_i \binom{k-t}{i-w} \tag{1}$$

where $\binom{l}{m}$ is 0 if $m < 0$ or $m > l$, and $\binom{0}{0}$ is 1. It follows that f is t-null if for $0 \leq w \leq t$, $p_{t,w}(f)$ are all equal. It is easy to see that the constant boolean functions $\{\mathbf{0}, \mathbf{1}\}$ are t-null for all t with $1 \leq t \leq k$. The parity functions $\{\oplus, \overline{\oplus}\}$ are also t-null for all t satisfying $1 \leq t < k$. From Lemma 3.1 and Equation 1 we get:

Corollary 3.2. *All symmetric boolean functions $f \notin \{\mathbf{0}, \mathbf{1}, \oplus, \overline{\oplus}\}$ have a non-zero Fourier coefficient of order at most t_0 (and at least 1) iff $\{\mathbf{0}, \mathbf{1}, \oplus, \overline{\oplus}\}$ are the only solutions to*

$$\sum_{i=0}^{k-t_0} f_i \binom{k-t_0}{i} = \sum_{i=1}^{k-t_0+1} f_i \binom{k-t_0}{i-1} = \cdots = \sum_{i=t_0}^{k} f_i \binom{k-t_0}{i-t_0} \tag{2}$$

In the next section, we show that this is true for $t_0 \leq Ck/\log k$ for large k and some positive constant C.

3.2 A bound of $O(k/\log k)$.

This section is devoted to the proof of our main result:

Theorem 3.3. *There is an absolute constant $k_0 > 0$ such that for $k \geq k_0$, every symmetric boolean function f on k bits with $f \notin \{\mathbf{0}, \mathbf{1}, \oplus, \overline{\oplus}\}$ has a non-zero Fourier coefficient of order at most $4k/\log k$ and at least 1.*

Let us start with some general discussion about the proof.

The preliminary setup is the following. Suppose f is a boolean function on $G = \mathbb{Z}_2^k$, such that all its non-constant Fourier coefficients of order up to $\epsilon k = k - N$ are 0. Then the values f_j of f satisfy (2) with $t_0 = k - N$, which, changing indices, can be rewritten as:

$$\sum_j \binom{N}{j} f_{\nu+j} = c_N, \quad \text{for all } \nu = 0, \ldots, k-N. \tag{3}$$

It is easy to show by induction on N, starting with $N = k$ and going down, that

$$c_N = 2^N \mathrm{Avg}\, f = 2^{N-k} \sum_{x \in \{0,1\}^k} f(x). \tag{4}$$

We want to show that if $k - N = \epsilon k = 4k/\log k$, then f_j is either constant or alternates between 0 and 1. We prove this for all k sufficiently large.

Define $D_j = f_{j+1} - f_j$, for $j = 0, \ldots, k-1$, and observe that the sequence D_j satisfies the homogeneous version of (3):

$$\sum_j \binom{N}{j} D_{\nu+j} = 0, \quad \text{for all } \nu = 0, \ldots, k-N-1. \tag{5}$$

Remark. In (5) the number N can be replaced by any other integer N_1 in the interval $[N, k]$. This follows since all the non-constant Fourier coefficients up to order $k - N$ are 0.

From (5) the sequence D_j may be defined for all $j \in \mathbb{Z}$ and $D_j \in \mathbb{Z}$ for all j. From the theory of recurrence relations we know then that the sequence D_j may be written as a linear combination of the following sequences:

$$(-1)^j, (-1)^j j, (-1)^j j^2, \ldots, (-1)^j j^{N-1}.$$

The reason for this is that -1 is the only root of the characteristic polynomial of the recurrence, $\phi(z) = \sum_j \binom{N}{j} z^j = (1+z)^N$. Therefore there is a polynomial $P(x)$, of degree at most $N-1$, such that

$$D_j = (-1)^j P(j), \quad \text{for all } j \in \mathbb{Z}.$$

Clearly $P(x)$ takes integer values on integers and in particular $P(j) \in \{-1, 0, 1\}$ for $j = 0, \ldots, k-1$. From the well known characterization of integer-valued polynomials [**?**, p. 129, Problem 85] it follows that we may write

$$P(x) = \sum_{j=0}^{N-1} a_j \binom{x}{j}, \quad \text{with } a_j \in \mathbb{Z}. \tag{6}$$

At this point it is instructive to give a proof, in this framework, of a result of [13]. This proof will also serve to clarify the relation of our method to that of [17]. A boolean function is called *balanced* if it takes the value 1 as often as it takes the value 0.

Theorem 3.4. (Mossel, O'Donnel and Servedio, 2003) *If $f : \{0,1\}^k \to \{0,1\}$ is a balanced symmetric function which is not constant or a parity function then some of its Fourier coefficients of order at most $O(k^{0.548})$ are non-zero.*

Proof. Subtracting c_N from both sides of (3) and using (4) we obtain that the sequence $f_n - \frac{c_N}{2^N} = f_n - \operatorname{Avg} f = f_n - \frac{1}{2}$ satisfies the homogeneous recurrence relation (5) in place of D_n. By the same reasoning as above $(-1)^n(f_n - \frac{1}{2})$ is then a polynomial of degree at most $N-1$. But it only takes the values $\pm\frac{1}{2}$ for $n = 0, 1, \ldots, N, \ldots, k-1$. Von zur Gathen and Roche [17] have shown that any polynomial $Q(n)$ which takes only two values for $n = 0, 1, \ldots, k$ must have degree $d \geq k - O(k^{0.548})$, hence $k - N = O(k^{0.548})$, which is what we wanted to prove. □

Remark. The method of [17] says nothing about polynomials which may take 3 or 4 values. If one ommits the assumption that f is balanced then the sequence $(-1)^n(f_n - \operatorname{Avg} f)$ may take up to 4 possible values.

Plan of proof. We assume that f has all non-constant Fourier coefficients of order up to $k-N$ equal to 0 and we want to show that $f \in \{\mathbf{0}, \mathbf{1}, \oplus, \overline{\oplus}\}$. Since $D_j = f_{j+1} - f_j$ it is enough to show that either D_j is identically 0 or that $D_j = (-1)^j$ or $D_j = (-1)^{j+1}$. This is equivalent to showing that $P(j) = (-1)^j D_j$ is a constant polynomial, constantly equal to $-1, 0$ or 1.

We will first show that the polynomial P is constant in two "small" intervals at the endpoints of the interval $[0, k]$ (Lemma 3.6). To achieve this we will first show that P has period 2 in each of these intervals (Lemma 3.5). For this we use some elaborate number-theoretic results (Theorem A) on the distribution of primes. Many of the technicalities in that part would not be needed if one knew that there are plenty of twin primes, that is integers p such that p and $p+2$ are both primes.

Once we have that P is constant in these two intervals near the endpoints of $[0, k]$ we show using the modular approach first introduced in [11] that P is also constant on a similar interval around the midpoint of $[0, k]$ (Lemma 3.7). At this point a significant element of our method is to eliminate the possibility that P is 0 (we are assuming of course that f is not constant). To show this we interpret f as a probability measure on the discrete cube and the vanishing of Fourier coefficients up to order r becomes equivalent with r-wise independence of the marginals of that measure (Theorem 3.8). It follows that if P vanishes in the middle interval in question then the second moment of a certain random variable would be larger than we know it is (Corollary 3.9). This elimination of 0 as a possible value is what makes the method work. We repeatedly obtain that P is constant in more and more intervals of the same length, each in the middle of the existing gaps, until the whole interval $[0, k]$ is covered (Lemma 3.11).

Notation. In what follows we repeatedly use the letter C to denote a positive constant which depends on no parameter (unless we say otherwise). As is customary, this constant C need not be the same in all its occurences.

Lemma 3.5. *The polynomial P satisfies the 2-periodicity condition*

$$P(j) = P(j+2),$$

whenever $j, j+2 \in \mathcal{A} = [0, k-N-\Gamma] \cup [N+\Gamma, k-1]$.

Proof. If $p \geq N$ is a prime, and since all the factors that appear in denominators in (6) are strictly less than p (hence invertible mod p), it follows that the sequence $P(j) \bmod p$, $j \in \mathbb{Z}$, may be viewed as a polynomial with coefficients in $\mathbb{Z}_p$ and therefore is a p-periodic sequence mod p, i.e.

$$P(j+p) = P(j) \bmod p, \quad \text{for all } j \in \mathbb{Z} \text{ and } p \geq N. \tag{7}$$

If, in addition, $0 \leq j < j+p < k$, when all P-values that appear in (7) are in $\{-1, 0, 1\}$, it follows that we have the non-modular equality

$$P(j+p) = P(j), \quad (N \leq p \leq p+j < k). \tag{8}$$

We shall need various primes in intervals from now on. The version of the prime number theorem that we will be using is the Siegel-Walfisz theorem (see [9, Theorem 2]). Define the logarithmic integral

$$\operatorname{Li} x = \int_2^x \frac{dt}{\log t} \sim \frac{x}{\log x}, \quad (x \to \infty).$$

The Euler function $\varphi(q)$ below denotes the number of moduli mod q which are coprime to q.

Theorem A (Siegel-Walfisz) Let $\pi(x; M, a)$ be the number of primes $\leq x$ which are equal to $a \bmod M$ and assume that $(M, a) = 1$. Then if $M \leq (\log x)^A$, A a constant, we have

$$\pi(x; M, a) = \frac{\operatorname{Li} x}{\varphi(M)} + O(x \exp(-c\sqrt{\log x})), \quad (\text{as } x \to \infty). \tag{9}$$

where c depends on A only (the constant in the $O(\cdot)$ term is absolute).

For $\pi(x)$, the number of primes up to x without any restriction, we thus have $\pi(x) = \operatorname{Li}(x) + O(x \exp(-c\sqrt{\log x})$, for some absolute constant c.

These theorems guarantee that, for $x \to \infty$, the interval $[x, x + \Delta]$ has the "expected" number of primes whenever $\Delta \geq Cx/(\log x)^A$, whatever the constant A, even if we impose the condition that these primes are equal to $a \bmod M$, as long as $M \leq (\log x)^B$, for any constant B.

We use the above theorems along with the p-periodicity of P to deduce that P is in fact 2-periodic on the union of 2 small sub-intervals of $[0, k-1]$.

Definition 3.1. Γ *denotes the maximum difference between succesive primes in the interval* $[0, k]$.

From Theorem A it follows, for instance, that $\Gamma = O(k/\log^{10} k)$ which is $o(k - N)$.

Assume $q < r$ are two primes in $[N, N + h]$, where $h = (k - N)/3 = \frac{\epsilon}{3}k$. (The length of the interval $[N, N + h]$ is large enough to guarantee the existence of many primes in it.) From (8) it follows that the finite sequences

$$P(0), \ldots, P(k - q) \quad \text{and} \quad P(q), \ldots, P(k)$$

are identical. Applying (8) again with r we get that the finite sequences

$$P(0), \ldots, P(k - r) \quad \text{and} \quad P(r), \ldots, P(k)$$

are identical. It follows that

$$P(j + r - q) = P(j), \quad \text{for all } j \text{ with } N + h \leq j \leq N + 2h \text{ and } r > q \text{ primes in } [N, N + h]. \tag{10}$$

We now assume, as we may, that the difference $M = r - q$ is the smallest difference between two primes in $[N, N + h]$. By the prime number theorem $M \leq C \log k$. Hence, we can apply Theorem A with modulus M. Since $\varphi(M) \leq M \leq C \log k$ in that case Theorem A guarantees that the number of primes equal to $a \bmod M$ in $[N, N + h]$ is at least

$$C\frac{h}{\log^2 k} \sim C\frac{k}{\log^3 k},$$

whenever $(M, a) = 1$. All that matters here is that this number is positive for large k.

Let $t \in [N, N + h]$ be the smallest prime which is equal to $-1 \bmod M$. By Theorem A, applied to modulus M and residue -1, its existence is guaranteed and furthermore that $t \sim N$. The same theorem guarantees that we can find a prime $s \in (t, N + h]$ such that $s = 1 \bmod M$. Then

$s-t = 2 \bmod M$ or $s-t = \ell M+2$, for some nonnegative integer ℓ. Therefore, for $N+h \le j \le N+2h$ we have

$$\begin{aligned} P(j) &= P(j+s-t) \ \text{(applying (10) for the primes } s,t) \\ &= P(j+\ell M+2) \\ &= P(j+(\ell-1)M+2) \ \text{(applying (10) for the primes } r,q) \\ &\cdots \\ &= P(j+2). \end{aligned}$$

This 2-periodicity

$$P(j) = P(j+2) \tag{11}$$

is now transferred to all $j, j+2 \in \mathcal{A}$ by using (8) repeatedly for appropriate primes p.

We use the following observation: if $P(j)$ is 2-periodic in an interval $[a,b] \subseteq [0,k]$ and $j \in [0,k]$ is such that there exists a prime $p \ge N$ for which $j+p, j+2+p \in [a,b]$ or $j-p, j+2-p \in [a,b]$ then $P(j) = P(j+2)$.

Since we know that P is 2-periodic in the interval $[N+h, N+2h]$, we first apply the observation to obtain the 2-periodicity in the interval $[0, 2h-\Gamma]$, since for any j in that interval we can find an appropriate prime to apply the observation.

Using this new interval we now get the 2-periodicity in the interval $[N+\Gamma, k]$. Next we deduce the 2-periodicity in the interval $[0, k-N-\Gamma]$.

□

Notice that in the sequence D_j, if one erases the 0's, one sees an alternation of -1 and 1 (this follows from the fact that $f_j \in \{0,1\}$). This property greatly reduces the number of allowed patterns in D_j and in fact it implies that P is constant in $\mathcal{A}$.

Lemma 3.6. *The polynomial P is constant in $\mathcal{A}$ (defined in Lemma 3.5).*

Proof. From Lemma 3.5 the values of P in $[N+\Gamma, k-1]$ must be a 2-periodic sequence. The only essentially different non-constant 2-periodic patterns for the values of P in $[N+\Gamma, k-1]$ are $010101\ldots$ and $(-1)1(-1)1\ldots$ and they both violate the property that $D_j = (-1)^j P(j)$ must satisfy, namely that if one erases the 0's then one must see an alternation of 1 and -1. Therefore P is constant in each of the two intervals of $\mathcal{A}$. From the p-periodicity (8), applied, say, for some $p \sim (k+N)/2$ it follows that the constant is the same in both intervals. □

We now extend the set on which P is constant to a superset of $\mathcal{A}$ that contains a small interval around $k/2$.

Lemma 3.7. *Let $a = \frac{N}{2} + \frac{3\Gamma}{2}$ and $b = \frac{N}{2} + (k-N) - \frac{5\Gamma}{2}$. Then $P(l) = P(0)$ for $a \le l \le b$.*

Proof. We will make use of the following theorem which follows from Lucas' Theorem [6, Ch. 3].

Theorem B If r is a prime which does not divide n then $\binom{mr}{n} = 0 \bmod r$. Also, if $0 \le m < r$ then $\binom{mr}{lr} = \binom{m}{l} \bmod r$.

We shall apply Theorem B with $m = 2$ and with a prime r such that $2r$ is the least possible such number larger than $N+\Gamma$. It follows that $2r \le (N+\Gamma) + 2\Gamma = N+3\Gamma$. And it follows from the remark after (5) that

$$\sum_j (-1)^j \binom{2r}{j} P(j+\nu) = 0, \quad (\nu \in \mathbb{Z}). \tag{12}$$

Taking residues mod r and using Theorem B for $m = 2$ we obtain

$$P(\nu) - 2P(\nu + r) + P(\nu + 2r) = 0 \bmod r, \quad (\nu \in \mathbb{Z}).$$

By our particular choice of r we have $P(\nu) = P(\nu + 2r) = P(0)$ whenever $\nu \in [0, k - N - 3\Gamma]$. It follows that $P(\nu + r) = P(0)$ for all such ν so we get $P(l) = P(0)$ for all l in the interval

$$\left[\frac{N}{2} + \frac{3\Gamma}{2}, \frac{N}{2} + (k - N) - \frac{5\Gamma}{2}\right].$$

□

So far we have proved $P(l) = P(0)$ on the set (a, b are defined in Lemma 3.7)

$$\mathcal{A}_2 = [0, k - N - \Gamma] \cup [a, b] \cup [N + \Gamma, k - 1],$$

which consists of three asymptotically equispaced intervals of asymptotic size ϵk. We consider two cases for P. The first is when P is 0 on $\mathcal{A}_2$ and the second is when P is 1 or -1.

To eliminate the case that P is 0 on $\mathcal{A}_2$, we shall need the following theorem, which already gives a lot of significant information about the function f. It should be thought of as analogous to the fact that the moments of a vector random variable can be read off the Fourier Transform of its distribution (the *characteristic function*) by looking at partial derivatives at 0.

Theorem 3.8. *Suppose $f : G = \mathbb{Z}_2^k = \{0, 1\}^k \to \mathbb{R}$ is nonnegative and not identically 0 and has all its Fourier coefficients of order at most r (and at least 1) equal to 0. Let μ denote the uniform probability measure on the cube G and ν denote the probability measure on G defined by*

$$\nu(A) = \sum_{x \in A} f(x) \Big/ \sum_{x \in G} f(x), \quad (A \subseteq G).$$

Let also $X_1, \ldots, X_k$ denote the coordinate functions on G, which we view as random variables. Then for all $i_1 < i_2 < \cdots < i_s$, $0 \le s \le r$, we have

$$\mathbf{E}_\nu(X_{i_1} \cdots X_{i_s}) = \mathbf{E}_\mu(X_{i_1} \cdots X_{i_s}).$$

Proof. Let $F = \sum_{x \in G} f(x)$. We assume for simplicity that $i_1 = 1, \ldots, i_s = s$. Then, writing $x = (x_1, x_2, \ldots, x_k)$ and $[s] = \{1, \ldots, s\}$, we have

$$\begin{aligned}
\mathbf{E}_\nu(X_1 \cdots X_s) &= \frac{1}{F} \sum_{x \in G} f(x) x_1 \cdots x_s \\
&= \frac{1}{F} \sum_{x \in G} f(x) \frac{1 + (-1)^{x_1 + 1}}{2} \cdots \frac{1 + (-1)^{x_s + 1}}{2} \\
&= \frac{1}{2^s F} \sum_{x \in G} f(x) \sum_{S \subseteq [s]} (-1)^{|S| + \sum_{i \in S} x_i} \\
&= \frac{|G|}{2^s F} \sum_{S \subseteq [s]} (-1)^{|S|} \frac{1}{|G|} \sum_{x \in G} f(x) (-1)^{\sum_{i \in S} x_i} \\
&= \frac{|G|}{2^s F} \sum_{S \subseteq [s]} (-1)^{|S|} \widehat{f}(S) \\
&= \frac{|G|}{2^s F} \widehat{f}(0) \quad \text{(by the vanishing of } \widehat{f}(S) \text{ for } \emptyset \ne S \subseteq [s]) \\
&= 2^{-s} \\
&= \mathbf{E}_\mu(X_1 \cdots X_s)
\end{aligned}$$

□

Remarks.
1. For functions $f : \{0,1\}^k \to \{0,1\}$, which is all we shall need here, the above theorem also follows directly from the definition of t-nullity in Section 3.1.
2. If the nonnegative function f is symmetric then the identity of moments up to order r with those of the uniform distribution (r-wise independence) and the vanishing of the non-constant Fourier coefficients of weight up to r are equivalent. This can be proved by induction on r. We do not use this here.

Corollary 3.9. *Under the assumptions and definitions of Theorem 3.8 the random variable $S = X_1 + \cdots + X_k$ has the same power moments $\mathbf{E}(S^s)$ under the probability measures μ and ν, up to order $s \leq r$.*

Proof. The power S^s, $s \leq r$, can be written as a sum of terms of the type $X_{i_1} \cdots X_{i_t}$, for $t \leq s$. One uses the fact that $X_j^2 = X_j$. □

Lemma 3.10. *If P is 0 on $\mathcal{A}_2$, then f is constant.*

Proof. Suppose the polynomial P is constantly equal to 0 on the set $\mathcal{A}_2$ and that f is not constant. The sequence f_j is then constant in each of the three intervals of $\mathcal{A}_2$. By possibly considering $1 - f$ (whose Fourier coefficients vanish exactly where those of f do, if f is not a constant function), we may assume that $f_j = 0$ on the middle interval (a, b). Let τ be the distribution of the random variable $S = X_1 + \cdots + X_k$ under the measure induced by f on G (each vertex $x \in G$ has probability proportional to $f(x)$), where $X_1, \ldots, X_k$ are the coordinate functions on G. Note that this is a well defined probability distribution since we assumed that f is not the $\mathbf{0}$ function.

The s-th moment with respect to the measure τ of the variable S in Corollary 3.9 is the expression

$$M(\tau, s) = \frac{1}{F} \sum_j f_j \binom{k}{j} j^s,$$

where again $F = \sum_j f_j \binom{k}{j}$. By Corollary 3.9, if $s \leq k - N$ this moment must equal the s-th moment with respect to the binomial measure μ, which is the quantity

$$M(\mu, s) = 2^{-k} \sum_j \binom{k}{j} j^s.$$

But the variance of S under μ is

$$M(\mu, 2) - M(\mu, 1)^2 = k, \tag{13}$$

since under μ the random variables $X_1, \ldots, X_k$ are independent, while the variance of S under τ is

$$\mathbf{E}_\tau (S - \mathbf{E}_\tau S)^2 = \mathbf{E}_\tau (S - \mathbf{E}_\mu S)^2 = \mathbf{E}_\tau (S - k/2)^2 \geq C\epsilon^2 k^2 \tag{14}$$

as the mass of τ sits to the left of $a \sim k/2 - \epsilon k/2$ and to the right of $b \sim k/2 + \epsilon k/2$. The orders of magnitude in (13) and (14) are different whenever $\epsilon \geq C/\sqrt{k}$, which is true in our case as $\epsilon = 4/\log k$. This contradiction proves that P cannot equal 0 on $\mathcal{A}_2$. □

Extending $\mathcal{A}_2$ to $[0, k-1]$.

For $2^l = m = 2, 4, \ldots$, we define the sets

$$B_m = \bigcup_{j=0}^{m} \left[\frac{j}{m} N + \Delta(m), \frac{j}{m} N + \epsilon k - \Delta(m) \right],$$

where $\Delta(m) = \Delta(m/2) + m\Gamma$, for $m \geq 4$, and $\Delta(2) = 3\Gamma$. (These intervals will be overlapping when m is large.)

Lemma 3.11. *There is a constant $k_0 > 0$ such that if $k \geq k_0$ and $\epsilon = 4/\log k$ then*
(a) the polynomial P is equal to 1 on $B_m \cap [0, k-1]$, for $m = 2, 4, 8, \ldots$ with $m \leq \frac{1}{2} \log k$, and
(b) if m takes the highest value allowed in (a) then B_m covers $[0, k-1]$, hence $P = 1$ on $[0, k-1]$.

Proof. To prove (a) we work by induction on $m = 2, 4, \ldots$. The base case $m = 2$ is settled since we have $B_2 \subseteq \mathcal{A}_2$ (that's why we chose $\Delta(2)$ large enough).

Assume now that we have proved $P = 1$ on $B_{m/2} \cap [0, k-1]$. We apply Theorem B for m and we choose a prime r such that mr is the least possible larger than N. Thus

$$N/m \leq r \leq N/m + \Gamma. \tag{15}$$

Theorem B together with relation (12) gives for all $\nu \in \mathbb{Z}$

$$P(\nu) - mP(\nu + r) + \binom{m}{2} P(\nu + 2r) - \cdots + (-1)^m P(\nu + mr) = 0 \bmod r. \tag{16}$$

We would like, for j even, the number $\nu + jr$ to belong to $B_{m/2}$, for most values of ν in the interval $[0, \epsilon k]$. That is we want

$$\frac{j}{m} N + \Delta(m/2) \leq \nu + jr \leq \frac{j}{m} N + \epsilon k - \Delta(m/2),$$

for $0 \leq j \leq m$, j even. Given (15) this follows from

$$\Delta(m/2) \leq \nu \leq \epsilon k - \Delta(m/2) - m\Gamma. \tag{17}$$

For ν satisfying (17) the range of the expression $\nu + jr$ (j fixed) contains the interval

$$[jr + \Delta(m/2), jr + \epsilon k - \Delta(m/2) - m\Gamma],$$

which, using (15) again, contains the interval

$$\left[\frac{j}{m} N + m\Gamma + \Delta(m/2), \frac{j}{m} N + \epsilon k - \Delta(m/2) - m\Gamma \right].$$

From the relation $\Delta(m) = \Delta(m/2) + m\Gamma$ it follows that this last interval is the j-th interval of B_m.

We have shown that whenever ν satisfies (17) the numbers $\nu + jr$, $0 \leq j \leq m$, j even, are all in $B_{m/2}$ so, by the induction hypothesis, the polynomial P takes the value 1 on them.

In the left hand side of (16) the sum of the absolute values of the coefficients is at most 2^m and as long as $2^m < r$ it follows that (mod r) can be dropped from (16). If (17) is satisfied it is clear that the sum of the terms of (16) corresponding to even j is 2^{m-1}, since these P terms are all 1. If, in addition $2^m < r$, we obtain that the terms corresponding to odd j must all have their P term

equal to 1. The reason for this is that the sum of absolute values of the odd terms is at most 2^{m-1} and is equal to that only in case all P's are equal to 1.

Letting ν run through all terms allowed by (17) we obtain that P has the value of 1 on all intervals of B_m corresponding to odd j. Since the intervals corresponding to even j are already contained in $B_{m/2}$ we obtain the desired conclusion, that P is equal to 1 on B_m, as long as $2^m < r$, which is clearly satisfied if $2^m < N/m$ or

$$m \leq \frac{1}{2} \log k. \tag{18}$$

This concludes the proof of (a).

To prove (b) observe that $\Delta(m) \leq 2m\Gamma$. Letting $\epsilon = 4/\log k$, we observe that if we let m be as large as part (a) allows then each of the intervals of B_m overlaps with the next one thus covering all of the interval $[0, k-1]$, which proves (b) and that P is constantly equal to 1, as we had to prove. □

This concludes the proof of the Theorem 3.3, which implies:

Corollary 3.12. *The class of symmetric k-juntas can be learned exactly under the uniform distribution with confidence $1-\delta$ in time $n^{O(k/\log k)} \cdot \mathrm{poly}(2^k, n, \log(1/\delta))$.*

4 Discussion

The main open question is to obtain tight upper and lower bounds on the running time of the Fourier-based algorithm for symmetric juntas. It may even be that for large k, every symmetric function has a non-zero Fourier coefficient of constant order.

It should also be noted that in the case of balanced symmetric functions, i.e., symmetric functions with $Pr[f(x) = 1] = 1/2$, a bound of $O(k^{0.548})$ follows from [17] (see [13]). Hence to improve our result, one may focus on finding new techniques for unbalanced functions.

References

[1] A. Blum. Relevant examples and relevant features: Thoughts from computational learning theory. In *AAAI Symposium on Relevance*, 1994.

[2] A. Blum. Open problems. COLT, 2003.

[3] A. Blum, M. Furst, M. Kearns, and R. J. Lipton. Cryptographic primitives based on hard learning problems. In *CRYPTO*, pages 278–291, 1993.

[4] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, 1997.

[5] N. Bshouty, J. Jackson, and C. Tamon. More efficient PAC learning of DNF with membership queries under the uniform distribution. In *Annual Conference on Computational Learning Theory*, pages 286–295, 1999.

[6] P. Cameron. *Combinatorics: topics, techniques, algorithms.* Cambridge University Press, 1994.

[7] D. Helmbold, R. Sloan, and M. Warmuth. Learning integer lattices. *SIAM Journal of Computing*, 21(2):240–266, 1992.

[8] J. Jackson. An efficient membership-query algorithm for learning dnf with respect to the uniform distribution. *Journal of Computer and System Sciences*, 55:414–440, 1997.

[9] A. Kumchev. *The distribution of prime numbers.* manuscript, 2005.

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

[11] R. Lipton, E. Markakis, A. Mehta, and N. Vishnoi. On the fourier spectrum of symmetric boolean functions with applications to learning symmetric juntas. In *IEEE Conference on Computational Complexity*, 2005.

[12] Y. Mansour. An $o(n^{\log\log n})$ learning algorithm for DNF under the uniform distribution. *Journal of Computer and System Sciences*, 50:543–550, 1995.

[13] E. Mossel, R. O'Donnel, and R. Servedio. Learning juntas. In *STOC*, pages 206–212, 2003.

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

[15] K. Verbeurgt. Learning DNF under the uniform distribution in quasi-polynomial time. In *Annual Workshop on Computational Learning Theory*, pages 314–326, 1990.

[16] K. Verbeurgt. Learning sub-classes of monotone DNF on the uniform distribution. In *Michael M. Richter, Carl H. Smith, Rolf Wiehagen, and Thomas Zeugmann, editors, Algorithmic Learning Theory, 9th International Conference*, pages 385–399, 1998.

[17] J. von zur Gathen and J. Roche. Polynomials with two values. *Combinatorica*, 17(3):345–362, 1997.