Electronic Colloquium on Computational Complexity, Report No. 178 (2010)

On the Minimal Fourier Degree of Symmetric Boolean Functions

Amir Shpilka [*] Avishay Tal[†]

Abstract

In this paper we give a new upper bound on the minimal degree of a nonzero Fourier coefficient in any non-linear symmetric Boolean function. Specifically, we prove that for every non-linear and symmetric $f : \{0,1\}^k \to \{0,1\}$ there exists a set $\emptyset \neq S \subset [k]$ such that $|S| = O(\Gamma(k) + \sqrt{k})$, and $\hat{f}(S) \neq 0$, where $\Gamma(m) \leq m^{0.525}$ is the largest gap between consecutive prime numbers in $\{1,\ldots,m\}$. As an application we obtain a new analysis of the PAC learning algorithm for symmetric juntas, under the uniform distribution, of Mossel et al. [MOS04]. Namely, we show that the running time of their algorithm is at most $n^{O(k^{0.525})} \cdot \mathrm{poly}(n, 2^k, \log(1/\delta))$ where n is the number of variables, k is the size of the junta (i.e. number of relevant variables) and δ is the error probability. In particular, for $k \geq \log(n)^{1/(1-0.525)} \approx \log(n)^{2.1}$ our analysis matches the lower bound 2^k (up to polynomial factors).

Our bound on the degree greatly improves the previous result of Kolountzakis et al. [KLM⁺09] who proved that $|S| = O(k/\log k)$.

[*] Faculty of Computer Science, Technion-Israel Institute of Technology, Haifa, Israel and Microsoft Research, Cambridge MA. Email: shpilka@cs.technion.ac.il. This research was partially supported by the Israel Science Foundation (grant number 339/10).

[†] Faculty of Computer Science, Technion-Israel Institute of Technology, Haifa, Israel. Email: avishay.tal@gmail.com.

ISSN 1433-8092

1 Introduction

One of the most important tools in the analysis of Boolean functions is the Fourier transform of the function. Roughly, the Fourier transform studies the correlation that the function has with linear functions. Although the Fourier transform is nothing but a linear transformation on the space of functions, it has found many applications in different areas of theoretical computer science, a partial list of includes learning theory, hardness of approximation, pseudo-randomness, social choice theory and more.

A typical question concerning the Fourier transform is: given a family of Boolean functions, what can we say about the Fourier spectrum of members in the family. For example, is most of the weight of the Fourier spectrum concentrated on the first few levels? Is the Fourier spectrum spread? Does the function have a nonzero Fourier coefficient at a certain level?

In this paper we consider the family of symmetric Boolean functions and study the following problem: What is the minimal degree such that *any* nonlinear symmetric Boolean function $f : \{0,1\}^k \to \{0,1\}$ has a nonzero Fourier coefficient of (at most) that degree. In other words, what is the minimal size of a set $\emptyset \neq S \subseteq [k]$ such that $\hat{f}(S) \neq 0$.

This problem was first studied (although implicitly) in [MOS04] in the context of giving PAC learning algorithms for Boolean juntas. It was later explicitly discussed in [KLM$^+$09], where improved bounds were obtained. A related question was studied in [GR97]. There the authors studied the question of what is the *maximal* degree such that *any* nonconstant symmetric Boolean function $f : \{0,1\}^k \to \{0,1\}$ has a nonzero Fourier coefficient of that degree. Although this question seems the complete opposite of the question that we study here, note that if f is *balanced* (i.e. obtains the values 0 and 1 equally often) and $\hat{f}(S) \neq 0$ then $\hat{g}([k] \setminus S) \neq 0$, and vice versa, where $g = f \oplus \mathrm{PARITY}$. Thus, a lower bound on the maximal degree translates to an upper bound on the minimal degree. We discuss these results in more detail in Section 1.3.

Besides being a very natural question that continues the investigation of Fourier spectrum of Boolean functions, our work is also motivated by the problem of giving learning algorithms for symmetric juntas.

Learning juntas. One of the most important open problems in learning theory is learning in the presence of irrelevant information. The problem can be described in the following way: we are given as input a set of labelled data points, coming from some high dimensional space and we have to come up with a (small) hypothesis that correctly labels the data points. However, it may be the case that only a small fraction of the data is actually relevant, and so, in order to be able to find such an hypothesis efficiently, we have to discover what the relevant variables are. This problem appears in many real-life applications. For example, when trying to learn how some genetic attribute depends on the DNA, we expect only a small number of DNA letters to affect this attribute, while the rest are irrelevant.

In this paper we consider a (special case of a) question that was proposed by Blum and Langley [Blu94, BL97] as a clean formulation of learning in the presence of irrelevant information. The general question is: Let $f : \{0,1\}^n \to \{0,1\}$ be an unknown Boolean function depending on $k \ll n$ variables. Henceforth we refer to such a function as a k-junta. We get as input a set of labelled examples $\langle x, f(x) \rangle$ where the data points $x = (x_1, \ldots, x_n)$ are chosen independently and uniformly at random from $\{0,1\}^n$. Our goal is to efficiently identify the k relevant variables and the truth table of the function (we shall describe the learning model in more detail in Section 3). It is clear that by going over all $\binom{n}{k}$ possible choices of k variables we can learn f. However, the main question is whether this can be done faster. Specifically, Blum and Langley [BL97] asked the following, still unsolved, question: "Does there exists a polynomial time algorithm for learning the class of Boolean functions over $\{0,1\}^n$ that have $\log(n)$ relevant features, in the PAC or uniform distribution models?" Note, that for this setting of parameters, this is a sub-problem of the notoriously hard questions of learning polynomial size DNF formulas and decision trees. Another evidence for the central role that the junta learning problem plays in computational learning theory can be found in the words of Mossel et al. [MOS04]: "We believe that the problem of efficiently learning k-juntas is the single most important open question in uniform distribution learning." For more background and applications we refer the reader to [Blu94, BL97, MOS04]. In this work we shall consider the case where the underlying junta is a symmetric function.

1.1 Our results

Our main result is a new theorem on the degree of the first (non-empty) non-zero Fourier coefficient, of a nonlinear symmetric Boolean function f. We shall need the following notation. For an integer m, denote with $\Gamma(m)$ the size of the largest interval inside $\{1,\ldots,m\}$ that does not contain a prime number. In other words,

$$\Gamma(m) = \max\{b-a \mid 1 \leq a < b \leq m \text{ and there is no prime number in the interval } (a,b]\}.$$

The best bound on Γ was given in [BHP01] where it was shown that $\Gamma(m) \leq m^{0.525}$. We also let $\hat{f}(S)$ be the Fourier coefficient of f at S (see definition in Section 2.1).

Theorem 1.1. *Let $f : \{0,1\}^k \to \{0,1\}$ be a non-linear symmetric Boolean function (i.e. f is not constant and is not parity nor its negation). Then, there exists a set $\emptyset \neq S \subset [k]$, of size $|S| = O(\Gamma(k)+\sqrt{k})$, such that $\hat{f}(S) \neq 0$.*

Our second result concerns an interesting subcase of the general junta learning problem that was first discussed in [MOS04], learning symmetric juntas. Here we are guaranteed that the unknown function is symmetric in its k variables. For this model we obtain the following learning result.

Theorem 1.2. *The class of symmetric k-juntas over n bits can be exactly learned, from random examples sampled from the uniform distribution, with confidence $1-\delta$, in time $n^{O(k^{0.525})} \cdot \mathrm{poly}(2^k, n, \log(1/\delta))$.*

Using standard learning tools, Theorem 1.2 follows immediately from our main result, Theorem 1.1.

Cramér proved that the Riemann hypothesis implies that $\Gamma(m) = O(\sqrt{m}\log m)$ (which is slightly weaker than Legendre's conjecture that $\Gamma(m) = O(\sqrt{m})$) and conjectured that $\Gamma(m) = O(\log^2 m)$ [Cra36]. Thus, if either Cramér's conjecture or Legendre's conjecture is true then Theorem 1.1 may be improved to give a set S of size $O(\sqrt{k})$, which will imply a similar improvement to Theorem 1.2.

1.2 Proof technique

A basic idea that appears in previous works is that if all non-empty Fourier coefficients of f, up to size t, are zero, then no matter how we fix any t variables from f, its bias remains the same. Namely, the probability that f assumes the value 0 is unchanged under any such fixing of at most t variables. This is formally stated in Lemma 4.4. The natural idea now is to consider many different restrictions and to try and combine all the information obtained from them to show that the bias cannot remain unchanged, unless f is a linear function.

Denote with $F(i)$ the value that f obtains on inputs that contain exactly i ones and $k-i$ zeros. It follows that $\mathrm{bias}(f) = \frac{1}{2^k}\sum_{i=0}^{k}\binom{k}{i}F(i)$ (see Definition 4.2). If we fix ℓ variables to 1 and $t-\ell$ variables to 0, and the bias is unchanged then we get that $\mathrm{bias}(f) = \frac{1}{2^{k-t}}\sum_{i=0}^{k-t}\binom{k-t}{i}F(i+\ell)$. Assume now that this holds for every $\ell \leq t$ and that $p = k-t > 2$ is a prime number. It follows that for every $\ell \leq t$,

$$2^{k-t}\cdot \mathrm{bias}(f) = \sum_{i=0}^{k-t}\binom{k-t}{i}F(i+\ell) \equiv_p F(\ell) + F(p+\ell) \ ,$$

where $\equiv_p$ means equality modulo p. Since $p > 2$ we actually get that there exists some constant c such that for every $\ell \leq t$, $F(\ell) + F(p+\ell) = c$. Thus, by considering those restrictions we learn that, for every $\ell \leq t$, $F(\ell)$ and $F(p+\ell)$ satisfy a certain linear relation.

By considering such relations for two primes $k - \sqrt{k} - 2\Gamma(k) < q < k - \sqrt{k} - \Gamma(k)$ and $\frac{k-\sqrt{k}}{2} - \Gamma(k) < p < \frac{k-\sqrt{k}}{2}$ we obtain enough information to conclude that if f does not have a non-zero small Fourier coefficient, then it must be a linear function (see Lemma 4.8).

We shall not briefly describe the learning algorithm of [MOS04]. In a nutshell, the algorithm first assumes that f is a linear function and then, by solving (or, more accurately, trying to solve) a system of linear equations, it will find the relevant variables of f (if the assumption is true). If the algorithm failed to find a relevant variable in the first step, then it will search for the sparsest non-zero Fourier coefficient that is supported on a small set S. In this way a subset of the relevant variables will be found, and it is not hard to see that one can use the same argument several times to recover all the relevant variables.

We note that our proof technique is very similar in nature to that of [KLM+09]. There the polynomial $G(z) = F(z+1) - F(z)$ was studied modulo different primes, however the information obtained from those primes was used in a different way than it is used here.

1.3 Related work

In [GR97], following [NS94], von zur Gathen and Roche studied the question of giving a lower bound on the *real* degree of non-constant symmetric Boolean functions. In other words, the problem is proving that there is a large set S such that $\hat{f}(S) \neq 0$. They were able to prove that the degree of a symmetric function on k bits is always at least $k - \Gamma(k)$, and conjectured that actually the degree is at least $k - O(1)$. This conjecture is still open. In [CST] the related question of providing lower bounds on the degree of symmetric functions from $\{0,1\}^k$ to $\{0,1,\ldots,m\}$ was considered and lower bounds of the form $k - o(k)$ on the degree were proved (when $m < k$). We shall later see the connection between bounding the degree of functions that take values in $\{0,1,2\}$ to proving the existence of a not too large S such that $\hat{f}(S) \neq 0$. We note that the result of [GR97] actually implies the following corollary. We say that a Boolean function f is balanced if $\Pr_x[f(x) = 0] = 1/2$. I.e. if f gets the values 0 and 1 equally often. In other words, $\hat{f}(\emptyset) = 0$, when f is viewed as a function to $\{-1,1\}$.

Corollary 1.3 ([GR97])**.** *Let $f : \{0,1\}^k \to \{0,1\}$ be a balanced symmetric Boolean function. Then, there exists a set $\emptyset \neq S \subset [k]$ of size $|S| \leq \Gamma(k)$ such that $\hat{f}(S) \neq 0$.*

Thus, Theorem 1.1 can be viewed as proving a similar bound for the case of *unbalanced* symmetric functions.

Mossel et al made the first breakthrough in PAC learning of juntas under the uniform distribution [MOS04]. They gave a learning algorithm whose running time is $n^{\frac{w}{w+1}k} \cdot \mathrm{poly}(n, 2^k, \log(1/\delta))$, where w is the matrix multiplication exponent. Currently the best bound on w gives $w < 2.376$ and so their algorithm runs in time (roughly) $n^{0.7k} \cdot \mathrm{poly}(n, 2^k, \log(1/\delta))$, which is better than the trivial algorithm that runs in time $n^k \cdot \mathrm{poly}(n, 2^k, \log(1/\delta))$. For the case of symmetric juntas, the algorithm of [MOS04] runs in time an $n^{2k/3} \cdot \mathrm{poly}(n, 2^k, \log(1/\delta))$. Their analysis for the case of symmetric juntas was greatly improved by Kolountzakis et al. [KLM+09] who gave an $n^{O(k/\log k)} \cdot \mathrm{poly}(n, 2^k, \log(1/\delta))$ upper bound on the running time of the algorithm for that case. Both results are based on the fact that every non-linear symmetric function f on k variables, has a non-zero Fourier coefficient that is supported on a somewhat small non-empty set S. Namely, on weaker versions of Theorem 1.1.

1.4 Organization

The paper is organized as follows. In Section 2 we give the basic definitions and discuss representations of Boolean functions as polynomials. In Section 3 we consider the PAC learning model and prove Theorem 1.2 assuming Theorem 1.1. The proof of Theorem 1.1 is given in Section 4. Finally, in Section 5 we present a possible approach towards obtaining an improved analysis of the learning algorithm of [MOS04] (for symmetric juntas).

2 Preliminaries

We denote $[m] = \{1, \ldots, m\}$. For $x \in \{0,1\}^n$ we denote with $|x|$ the weight of x, i.e., the number of non-zero entries in x. In other words, $|x| = x_1 + \ldots + x_n$. All logarithms in this paper are taken to base 2. We denote $\binom{n}{\leq r} \triangleq \sum_{i=0}^{r} \binom{n}{i}$. To ease the reading we will drop floors and ceilings, as it will be obvious that they do not affect the results.

Definition 2.1. *A Boolean function $f : \{0,1\}^n \to \{0,1\}$ is called a k-junta if it depends on only k of the input bits (usually $k \ll n$). Namely, there exists a function $g : \{0,1\}^k \to \{0,1\}$ and k indices $1 \leq i_1 < i_2 < \ldots < i_k \leq n$ such that*

$$f(x_1, x_2, ..., x_n) = g(x_{i_1}, x_{i_2}, ..., x_{i_k})$$

We will be studying integer equations modulo prime numbers and so the following two claims will be useful. The first is the well known Lucas' theorem.

Theorem 2.2 (Lucas)**.** *Let $a, b \in \mathbb{N} \setminus \{0\}$ and let p be a prime number. Denote with $a = a_0 + a_1 p + a_2 p^2 + \cdots + a_k p^k$ and $b = b_0 + b_1 p + b_2 p^2 + \cdots + b_k p^k$ their base p representations. Then $\binom{a}{b} \equiv_p \prod_{i=0}^{k} \binom{a_i}{b_i}$, where $\binom{a_i}{b_i} = 0$ if $a_i < b_i$.*

The second theorem guarantees the existence of a prime number in any large enough interval.

Theorem 2.3 ([BHP01])**.** *For all large m, the interval $[m - m^{0.525}, m]$ contains prime numbers.*

We note the famous conjectures of Legendre and Cramér stating that the gap between consecutive primes in $[m]$ is at most $O(\sqrt{m})$ (Legendre) or even just $O(\log^2(m))$ (Cramér), and that Cramér showed that the Riemann hypothesis implies that the gap is at most $O(\sqrt{m}\log m)$ [Cra36].

2.1 Representations of Boolean functions

The basic objects that we study in this paper are symmetric Boolean functions.

Definition 2.4. *A function $f : \{0,1\}^k \to \{0,1\}$ is symmetric if $f(x) = f(y)$ for all x and y such that $|x| = |y|$.*

In other words, a function is symmetric if permuting the coordinates of the input does not change the value of the function.

We shall consider two equivalent ways of representing symmetric Boolean functions. One common and useful representation is the Fourier transform (which applies to non-symmetric functions as well). For this representation it is convenient to think of our function f as mapping $\{-1,1\}^k$ to $\{-1,1\}$, by applying the linear transformation $b \to 1-2b$ from $\{0,1\}$ to $\{-1,1\}$.

For a subset $S \subseteq [n]$ denote $\chi_S(x) = \prod_{i\in S} x_i$. It is a well known fact that $\{\chi_S\}_{S\subseteq[k]}$ form an orthonormal basis to the space of functions from $\{-1,1\}^k$ to $\mathbb{C}$ under the inner product $\langle f,g\rangle = \mathbf{E}_{x\in\{-1,1\}^k}[f(x)\cdot\overline{g(x)}]$, where x is distributed uniformly on $\{-1,1\}^k$. In particular, every Boolean function $f : \{-1,1\}^k \to \{-1,1\}$ can be represented as

$$f(x) = \sum_{S\subseteq[k]} \hat{f}(S)\chi_S(x) = \sum_{S\subseteq[k]} \hat{f}(S)\prod_{i\in S} x_i \ , \tag{1}$$

where

$$\hat{f}(S) = \mathbf{E}_{x\in\{0,1\}^k}[f(x)\cdot\overline{\chi_S(x)}] \ . \tag{2}$$

We call $\hat{f}(S)$ the Fourier coefficient of f at S. Note that Equation (1) gives a representation of f as a polynomial over the reals. For example, if we denote $\text{PARITY}_k = \oplus_{i=1}^k x_i$ then, as a polynomial from $\{-1,1\}^k$ to $\{-1,1\}$, we have $\text{PARITY}_k = x_1\cdot x_2\cdots x_k = \chi_{[k]}$. When f is a symmetric polynomial it follows that $\hat{f}(S) = \hat{f}(T)$ whenever $|S| = |T|$. Parseval's identity implies that for $f : \{-1,1\}^k \to \{-1,1\}$ it holds that $\sum_{S\subseteq[k]} \hat{f}(S)^2 = 1$.

We note that whenever $\hat{f}(S) \neq 0$ then f depends on all variables x_i such that $i \in S$ (and possibly on other variables as well). In particular, in order to find a relevant variable for a junta f it is sufficient to find a non-empty S such that $\hat{f}(S) \neq 0$.

Symmetric Boolean function can also be represented by univariate real polynomials of degree at most n. Indeed, recall that $f(x)$ is actually a function of $|x| = \sum_{i=1}^k x_i$. Hence, there exists a degree $\leq k$ polynomial $F : \{0,\dots,k\} \to \{0,1\}$ such that $F(|x|) = f(x)$. Similarly to the Fourier representation we shall represent F using a specific basis, $\{1, x, \binom{x}{2}, \dots, \binom{x}{k}\}$. This basis is sometimes called the Newton basis. We can express f as:

$$f(x) = F(|x|) = \sum_{d=0}^{k} \gamma_d \cdot \binom{|x|}{d} \ . \tag{3}$$

The coefficients γ_d are given in the following lemma. For completeness, we give the proof of the lemma in Appendix A.

Lemma 2.5. $\gamma_d = \sum_{i=0}^d (-1)^{d-i}\cdot\binom{d}{i}\cdot F(i)$.

In particular, all γ_d's are integers and γ_d only depends on the first $d+1$ values of F. Another obvious fact is that the degree of f as a real polynomial in Equation (1) is the same as the degree of F in Equation (3) (even though the domain is different in the two representations).

It is obvious that f determines F and vice versa and so, in what follows, we shall have both representations in mind and will move freely between them.

We shall denote symmetric functions on the Boolean cube with the letters f, g, h and their corresponding integer polynomials with F, G, H, respectively.

3 Learning symmetric juntas

In this section we prove Theorem 1.2, assuming Theorem 1.1. We start by describing the PAC learning model in more detail.

The common model for learning juntas is the PAC-model that was introduced by Valiant in his seminal work [Val84]. In this model, the learner gets a set of labelled examples $\langle x, f(x) \rangle$ where x is drawn from a certain distribution D over $\{0,1\}^n$, and has to come up with a hypothesis h that approximates f with respect to the distribution D. When learning juntas we restrict our attention to the case where D, the underlying distribution, is the uniform distribution over $\{0,1\}^n$. Furthermore, our goal is to find the relevant k variables $x_{i_1}, \ldots, x_{i_k}$ and output h such that $h(x_{i_1}, \ldots, x_{i,k}) = f(x_1, \ldots, x_n)$. As the inputs to our learning algorithm are randomly distributed, we allow failure with a small probability δ.

The following lemma of [MOS04] shows that for the purpose of learning juntas, it is enough to find one relevant variable.

Lemma 3.1 ([MOS04])**.** *Suppose that A is an algorithm running in time $n^r \cdot \mathrm{poly}(2^k, n, \log(1/\delta))$ which can identify at least one variable relevant to f with confidence $1-\delta$ (assuming f is non-constant). Then there is an algorithm for exactly learning f which runs in time $n^r \cdot \mathrm{poly}(2^k, n, \log(1/\delta))$.*

Note that the only difference in the running time comes from the term $\mathrm{poly}(2^k, n, \log(1/\delta))$.

A common technique in PAC learning over the uniform distribution is to estimate Fourier coefficients of the unknown function f. Indeed, notice that as x is drawn uniformly at random, we can get a very good estimate of $\hat{f}(S)$, with high probability. Since f depends on k variables, if we compute $\hat{f}(S)$ to precision $1/2^{k+1}$ then by rounding to the nearest integer multiple of $1/2^k$ we can exactly compute $\hat{f}(S)$. This is captured by the following lemma of [MOS04].

Lemma 3.2 ([MOS04])**.** *For any set $S \subseteq [k]$, we can exactly calculate the Fourier coefficient $\hat{f}(S)$ with confidence $1-\delta$ in time $\mathrm{poly}(2^k, n, \log(1/\delta))$.*

Theorem 1.1 and Lemmas 3.2, 3.1 are almost all that is needed for the learning algorithm. All that is left is to handle the special case of linear functions. If f is a linear function (i.e., f is constant, parity or its negation) then by solving a system of linear equations we can easily find all the relevant variables of f. This is formally stated in the following simple lemma of [MOS04].

Lemma 3.3 ([MOS04])**.** *If f is a linear function, then we can learn f exactly in time $\mathrm{poly}(2^k, n, \log(1/\delta))$ with confidence $1-\delta$.*

We now formally present the learning algorithm, which is similar to the one given in [MOS04].

The proof of Theorem 1.2 easily follows, assuming Theorem 1.1 and the lemmas above.

Proof of Theorem 1.2. Let $\delta > 0$ be given. If f is linear then Lemma 3.3 implies that Step 1 will return the junta set with probability at least $1-\delta$. If f is not a linear function, Theorem 1.1 shows that there exists a non-empty set S of size s_k, $s_k = O(\Gamma(k) + \sqrt{k})$, such that $\hat{f}(S) \neq 0$. Set $\epsilon = \delta / \binom{n}{\leq s_k} = O(\delta / n^{s_k})$. By lemma 3.2, for any $S \subseteq [n]$ we can compute the Fourier coefficient $\hat{f}(S)$ exactly, with confidence $1-\epsilon$ in time $\mathrm{poly}(2^k, n, \log(1/\epsilon)) = \mathrm{poly}(2^k, n, \log(1/\delta))$. Thus, by the union bound, in time $n^{s_k} \cdot \mathrm{poly}(2^k, n, \log(1/\delta))$ we can compute exactly all Fourier coefficients $\hat{f}(S)$ for every $|S| \leq s_k$, with confidence $1-\delta$ (i.e. the probability that we do not compute all of them correctly is at most δ). Therefore, if f is not a linear function then we can find a non-empty set S with $\hat{f}(S) \neq 0$. It is clear that all the variables in S are relevant variables of f. Lemma 3.1

```
1. Run the algorithm guaranteed by Lemma 3.3 on $f$.
   Using $O(2^k \cdot \log(1/\delta))$ additional samples check that the linear function returned by the algorithm agrees with $f$.
   If this step was successful then return the linear function and halt.

2. Otherwise, for every $t \in [k]$:
   Go over all $S \subseteq \{1, 2, ..., n\}$ of size $t$ and compute the Fourier coefficient $\hat{f}(S)$, using Lemma 3.2.
   Output the first set $S$ such that $\hat{f}(S) \neq 0$ and halt.
```

Algorithm 1: Learning symmetric juntas

implies that we can learn f, with confidence $1 - \delta$, in time $n^{s_k} \cdot \mathrm{poly}(2^k, n, \log(1/\delta))$. □

4 Fourier spectrum of symmetric Boolean functions

In this section we prove Theorem 1.1. Our approach is similar to the approach taken by [GR97, KLM+09]. We study the bias of f after restricting some of the variables. From this point on we identify a symmetric function $f : \{0,1\}^k \to \{0,1\}$ with its corresponding integer polynomial $F : \{0, \dots, k\} \to \{0,1\}$. Recall that $F(i)$ is the value that f obtains on inputs of weight i.

Definition 4.1. *Let $F : \{0, 1, \dots, k\} \to \{0,1\}$ be a symmetric function on k bits. The (m, ℓ)-fixing of F, is a symmetric function on $k - m$ bits $F|_{(m,\ell)} : \{0, 1, \dots, k - m\} \to \{0,1\}$ defined by*

$$F|_{(m,\ell)}(i) \triangleq F(i + \ell) \; .$$

In other words, $f|_{(m,\ell)}$ is the symmetric function obtained by fixing ℓ variables to 1 and $m - \ell$ variables to 0 (again, we identify $f|_{(m,\ell)}$ with $F|_{(m,\ell)}$). We shall study the bias of F under different restrictions.

Definition 4.2. *The bias of a function $f : \{0,1\}^k \to \{0,1\}$ is defined as bias$(f) \triangleq E_{x \in \{0,1\}^k} f(x)$, where x is uniformly distributed.*

In other words, the bias is equal to the probability that $f(x) = 1$ (when x is picked uniformly at random). In particular, f is unbiased iff $\mathrm{bias}(f) = \frac{1}{2}$. Notice that when f is symmetric then the bias is given by

$$\mathrm{bias}(f) = \frac{1}{2^k} \sum_{i=0}^{k} \binom{k}{i} \cdot F(i) \; .$$

Similarly,

$$\mathrm{bias}(F|_{(m,\ell)}) = \frac{1}{2^{k-m}} \cdot \sum_{i=0}^{k-m} \binom{k-m}{i} \cdot F(i + \ell) \; . \qquad (4)$$

The following useful definition and lemma, relating the bias of $F|_{(m,\ell)}$ and the Fourier spectrum of f, was given in [KLM+09].

Definition 4.3 ([MOS04, KLM+09])**.** *f is called t-null if for every set $S \subseteq [k]$ such that $1 \leq |S| \leq t$, it holds that $\hat{f}(S) = 0$.*

Lemma 4.4 ([KLM+09])**.** *The following are equivalent.*

1. *f is t-null.*
2. *For every $0 \le \ell \le m \le t$, bias$(F|_{(m,\ell)}) =$ bias(f).*
3. *For every $0 \le \ell \le t$, bias$(F|_{(t,\ell)}) =$ bias(f).*

In order to prove that a symmetric f is not t-null, we will look for a (t, ℓ) fixing that changes the bias. Towards this end we shall consider the bias of f modulo different prime numbers. Let $p < k$ be a prime number. If f is $(k - p)$-null then, by Lemma 4.4, there exists c_p such that for all $\ell \le k - p$ it holds that

$$c_p = \text{bias}(F|_{(k-p,\ell)}) \; .$$

In other words, according to Equation (4),

$$2^p \cdot c_p = \sum_{i=0}^{p} \binom{p}{i} \cdot F(i + \ell) \; .$$

Reducing this equation modulo p we get that for every $\ell \le k - p$

$$2^p \cdot c_p \equiv_p \sum_{i=0}^{p} \binom{p}{i} \cdot F(i + \ell) \equiv_p F(\ell) + F(p + \ell) \; . \tag{5}$$

Similarly, by considering the case that f is $(k - 2p)$-null we get that there exists c_{2p} such that for all $\ell \le k - 2p$ it holds that

$$2^{2p} \cdot c_{2p} = \sum_{i=0}^{2p} \binom{2p}{i} \cdot F(i + \ell) \; .$$

As before, reducing modulo p and using Lucas' theorem (Theorem 2.2), we obtain that for every $\ell \le k - 2p$

$$2^{2p} \cdot c_{2p} \equiv_p \sum_{i=0}^{2p} \binom{2p}{i} \cdot F(i + \ell) \equiv_p F(\ell) + 2F(p + \ell) + F(2p + \ell) \; . \tag{6}$$

In the next two sections we study the effect of fixing bits on the bias of f and prove Theorem 1.1.

4.1 Fixing 2 bits

In this subsection we present two classes of functions for which bias$(F) \neq$ bias$(F|_{(2,1)})$. In particular, every such function is not 2-null. For $i = 1, \ldots, k - 1$ the weight of $F(i)$ in bias(F) is $\frac{1}{2^k}\binom{k}{i}$, whereas the weight of $F(i)$ in bias$(F|_{(2,1)})$ is $\frac{1}{2^{k-2}}\binom{k-2}{i-1}$. The following is an easy observation.

Claim 4.5.

$$\frac{1}{2^k}\binom{k}{i} \le \frac{1}{2^{k-2}}\binom{k-2}{i-1} \quad \textit{iff} \quad \frac{k - \sqrt{k}}{2} \le i \le \frac{k + \sqrt{k}}{2} \; .$$

Proof. The LHS is equivalent to $k(k - 1) \le 4i(k - i)$. I.e. to $i^2 - ik + k(k - 1)/4 \le 0$. Solving we get the claimed result. □

Corollary 4.6. *Let F be a non-constant function $F : \{0, 1, \ldots, k\} \to \{0, 1\}$. If $F(i) = c$ for all $\frac{k-\sqrt{k}}{2} \le i \le \frac{k+\sqrt{k}}{2}$, then bias$(F) \neq$ bias$(F|_{(2,1)})$.*

Proof. Assume w.l.o.g. that $c = 0$. Hence, $F(i) = 1$ only for i such that $i < \frac{k-\sqrt{k}}{2}$ or $\frac{k+\sqrt{k}}{2} < i$, and because F is non-constant there exists some i such that $F(i) = 1$. Thus, the weight of each non-zero $F(i)$ decreases after the fixing, hence the probability that $F = 1$ decreases. Formally,

$$\begin{aligned}
\operatorname{bias}(F) = \frac{1}{2^k}\sum_{i=0}^{k}\binom{k}{i}F(i) &= \frac{1}{2^k}\sum_{i<\frac{k-\sqrt{k}}{2}}\binom{k}{i}F(i) + \frac{1}{2^k}\sum_{\frac{k+\sqrt{k}}{2}<i}\binom{k}{i}F(i) \\
&\geq \frac{1}{2^k}\sum_{1\leq i<\frac{k-\sqrt{k}}{2}}\binom{k}{i}F(i) + \frac{1}{2^k}\sum_{\frac{k+\sqrt{k}}{2}<i\leq k-1}\binom{k}{i}F(i) \\
&>^{(*)} \frac{1}{2^{k-2}}\sum_{1\leq i<\frac{k-\sqrt{k}}{2}}\binom{k-2}{i-1}F(i) + \frac{1}{2^{k-2}}\sum_{\frac{k+\sqrt{k}}{2}<i\leq k-1}\binom{k-2}{i-1}F(i) \\
&= \operatorname{bias}(F|_{(2,1)})\ ,
\end{aligned}$$

where inequality $(*)$ follows from Claim 4.5 □

In a similar way we obtain the following corollary.

Corollary 4.7. *Let F be a non-constant function $F : \{0, 1, \ldots, k\} \to \{0, 1\}$. If $F(i) = c$ for all $i < \frac{k-\sqrt{k}}{2}$ or $i > \frac{k+\sqrt{k}}{2}$, then bias$(F) \neq$ bias$(F|_{(2,1)})$.*

4.2 Proof of Theorem 1.1

In order to obtain Theorem 1.1 we shall combine the information obtained from considering restrictions modulo two different primes. For that end we shall prove the following lemma that shows correlation in the values of F between different regions of $[k]$.

Lemma 4.8. *Let $2 < q \leq k$ be a prime number. Let f be a* biased *non-constant symmetric function on k bits which is $(k - q + 1)$-null. Then, there exists a constant $c_{q-1} \in \{0, 1\}$ such that for every $\ell = 0, \ldots, k - q$*

$$F(\ell) = F(\ell + q) = c_{q-1}\ .$$

We now show how Theorem 1.1 follows from the above lemma. To ease the reading we repeat the statement of the theorem here.

Theorem (Theorem 1.1)**.** *Let $f : \{0, 1\}^k \to \{0, 1\}$ be a non-linear symmetric Boolean function (i.e. f is not constant and is not parity nor its negation). Then, there exists a set $\emptyset \neq S \subset [k]$, of size $|S| = O(\Gamma(k) + \sqrt{k})$, such that $\hat{f}(S) \neq 0$.*

Proof of Theorem 1.1. If f is balanced then the claim follows from Corollary 1.3. Hence, we can assume that f is biased. In addition, assume for a contradiction that f is $(2\Gamma(k) + \sqrt{k})$-null. By the definition of Γ, there exist prime numbers p, q such that $\frac{k-\sqrt{k}}{2} - \Gamma(k) \leq p \leq \frac{k-\sqrt{k}}{2}$ and $k - \sqrt{k} - 2\Gamma(k) \leq q \leq k - \sqrt{k} - \Gamma(k)$. Since f is $(k - q + 1)$-null, Lemma 4.8 implies that there exists a constant $c_{q-1} \in \{0, 1\}$ such that

$$F(0) = F(1) = \ldots = F(k - q) = F(q) = F(q + 1) = \ldots = F(k) = c_{q-1}\ .$$

As f is also $(k - 2p)$-null, Equation (6) implies that there exists a constant $0 \leq c_{2p} < p$ such that for all $\ell = 0, 1, \ldots, k - 2p$

$$F(\ell) + 2 \cdot F(p + \ell) + F(2p + \ell) \equiv_p c_{2p}\ .$$

Assuming $4 < p$ (otherwise k is at most some fixed constant and the claim is not interesting), these equations hold over the integers and so we get that for every $\ell = 0, 1, \ldots, k - 2p$

$$F(\ell) + 2 \cdot F(p + \ell) + F(2p + \ell) = c_{2p} \ .$$

Note that for $\frac{k-\sqrt{k}}{2} - p \leq \ell \leq \frac{k+\sqrt{k}}{2} - p$, we have $\ell \leq \sqrt{k} + \frac{k-\sqrt{k}}{2} - p \leq \sqrt{k} + \Gamma(k) \leq k - q$ and $\ell + 2p \geq \frac{k-\sqrt{k}}{2} + p \geq k - \sqrt{k} - \Gamma(k) \geq q$. Therefore,

$$2 \cdot F(p + \ell) = c_{2p} - F(\ell) - F(2p + \ell) = c_{2p} - 2c_{q-1} \ .$$

In other words, F is constant in the interval $\left[\left\lceil \frac{k-\sqrt{k}}{2} \right\rceil, \left\lfloor \frac{k+\sqrt{k}}{2} \right\rfloor\right]$. By Corollary 4.6 we conclude that f is not 2-null, in contradiction. Therefore, f is not $(2\Gamma(k) + \sqrt{k})$-null, which is what we wanted to prove. □

We end this section by proving Lemma 4.8, which concludes the proof of Theorem 1.1.

Proof of Lemma 4.8. Lemma 4.4 implies that since f is $(k-q+1)$-null then for all $\ell = 0, 1, \ldots, k - q + 1$ it holds that

$$\sum_{i=0}^{q-1} \binom{q-1}{i} \cdot F(i + \ell) = 2^{q-1} \cdot \mathrm{bias}(F) \ .$$

Consider these equations modulo q. By Lucas' Theorem, $\binom{q-1}{i} \equiv_q (-1)^i$. Therefore, we get that there exists a number $0 \leq c_{q-1} < q$ such that

$$\sum_{i=0}^{q-1} (-1)^i \cdot F(i + \ell) \equiv_q c_{q-1} \ . \tag{7}$$

Hence, for all $\ell = 0, 1, \ldots, k - q$ it holds that

$$\sum_{i=0}^{q-1} (-1)^i \cdot F(i + \ell) \equiv_q c_{q-1} \equiv_q \sum_{i=0}^{q-1} (-1)^i \cdot F(i + \ell + 1) \ .$$

Adding the RHS to the LHS we obtain,

$$2c_{q-1} \equiv_q F(\ell) + \sum_{i=1}^{q-1} ((-1)^i + (-1)^{i-1}) \cdot F(i + \ell) + (-1)^{q-1} F(q + \ell) = F(\ell) + F(q + \ell) \ .$$

Hence, $2c_{q-1} \in \{0, 1, 2\} \mod q$. It follows that c_{q-1} is either $0, 1$ or $(q + 1)/2$. If $c_{q-1} = 0$ or 1 then clearly $F(\ell) = F(q + \ell) = c_{q-1}$ and we are done, so we only need to rule out the case $c_{q-1} = (q + 1)/2$. So assume that $c_{q-1} = (q + 1)/2$. Equation (7) gives

$$\sum_{i=0}^{q-1} (-1)^i \cdot F(i + \ell) \equiv_q (q + 1)/2 \ .$$

In other words,

$$\sum_{i<q \,:\, i \text{ even}} F(i + \ell) - \sum_{i<q \,:\, i \text{ odd}} F(i + \ell) \equiv_q (q + 1)/2 \ .$$

Therefore it must be the case that either $F(\ell) = F(\ell + 2) = \ldots = F(\ell + q - 1) = 1$ and $F(\ell + 1) = \ldots = F(\ell + q - 2) = 0$, or vice versa. This implies that $f|_{(k-q+1,\ell)}$ is parity or its negation, and in particular $f|_{(k-q+1,\ell)}$ is unbiased. As f is $(k - q + 1)$-null we have that $\mathrm{bias}(F|_{(k-q+1,\ell)}) = \mathrm{bias}(f)$, and so f is unbiased. This contradicts the assumption that f is biased.

□

5 On nullity and degree of polynomials taking three values

In this section we show a connection between the problem of upper bounding the minimal size of a non-zero Fourier coefficient of a symmetric function and the problem of giving a lower bound on the degree of a univariate polynomial $H : \{0, \dots, k\} \to \{0, 1, 2\}$, that was studied in [CST] (in the argument below we consider $H : \{0, \dots, k\} \to \{-1, 0, 1\}$, but the degrees of H and $H + 1$ are of course equal).

Using the observation that $\hat{f}(S) = (f \oplus \mathrm{PARITY})\hat{}\,(S^c)$, where S^c is the complement of S, Mossel et al. [MOS04] concluded that

$$\begin{aligned} \deg(F) < k - t &\iff \forall S : k - t \le |S|, \hat{f}(S) = 0 \\ &\iff \forall S : |S| \le t, (f \oplus \mathrm{PARITY})\hat{}\,(S) = 0 \\ &\iff f \oplus \mathrm{PARITY} \text{ is } t\text{-null and unbiased.} \end{aligned} \tag{8}$$

We shall prove a one directional reduction from *any* symmetric t-null function (i.e. even a biased one) to a low degree polynomial that maps $\{0, 1, \dots, k-2\}$ to $\{-1, 0, 1\}$. We first prove the following lemma that gives a relation between different coefficients in the Newton basis representation of a symmetric f such that $f \oplus \mathrm{PARITY}$ is t-null.

Lemma 5.1. *Let $f : \{0, 1\}^k \to \{0, 1\}$ be a symmetric function. If $f \oplus PARITY$ is t-null, then, when representing f in Newton's basis, $F(|x|) = \sum_{i=0}^{k} \gamma_i \cdot \binom{|x|}{i}$, we have $\gamma_{i+1} = -2\gamma_i$ for $i = k-t, ..., k-1$.*

Proof. Denote $g = f \oplus \mathrm{PARITY}$ and let $G : \{0, \dots, k\} \to \{0, 1\}$ be its univariate representation. Since we assume that g is t-null, it follows that $\mathrm{bias}(G|_{(\ell,0)}) = \mathrm{bias}(G|_{(\ell+1,0)})$ for $\ell = 0, \dots, t - 1$. Therefore,

$$\begin{aligned} \frac{1}{2^{k-\ell}} \cdot \sum_{i=0}^{k-\ell} \binom{k-\ell}{i} \cdot (-1)^{G(i)} &= \frac{1}{2^{k-\ell}} \cdot \sum_{i=0}^{k-\ell} \binom{k-\ell}{i} \cdot (1 - 2G(i)) \\ &= 1 - 2\mathrm{bias}(G|_{(\ell,0)}) = 1 - 2\mathrm{bias}(G|_{(\ell+1,0)}) \\ = \frac{1}{2^{k-\ell-1}} \cdot \sum_{i=0}^{k-\ell-1} \binom{k-\ell-1}{i} \cdot (1 - 2G(i)) &= \frac{1}{2^{k-\ell-1}} \cdot \sum_{i=0}^{k-\ell-1} \binom{k-\ell-1}{i} \cdot (-1)^{G(i)} . \end{aligned}$$

Multiplying both sides by $2^{k-\ell}$ and using the fact that $(-1)^{G(i)} = (-1)^{F(i)} \cdot (-1)^i$ we get

$$\sum_{i=0}^{k-\ell} \binom{k-\ell}{i} \cdot (-1)^{F(i)} \cdot (-1)^i = 2 \cdot \sum_{i=0}^{k-\ell-1} \binom{k-\ell-1}{i} \cdot (-1)^{F(i)} \cdot (-1)^i .$$

Since $F(i) = (1 - (-1)^{F(i)})/2$, and $\sum_{i=0}^{d} \binom{d}{i} \cdot (-1)^i = 0$, it follows that

$$\sum_{i=0}^{k-\ell} \binom{k-\ell}{i} \cdot F(i) \cdot (-1)^i = 2 \cdot \sum_{i=0}^{k-\ell-1} \binom{k-\ell-1}{i} \cdot F(i) \cdot (-1)^i . \tag{9}$$

By Lemma 2.5 we have $(-1)^d \gamma_d = \sum_{i=0}^{d} \binom{d}{i} \cdot F(i) \cdot (-1)^i$. Hence, Equation (9) is equivalent to

$$(-1)^{k-\ell} \cdot \gamma_{k-\ell} = 2 \cdot (-1)^{k-\ell-1} \cdot \gamma_{k-\ell-1} ,$$

i.e. $\gamma_{k-\ell} = -2 \cdot \gamma_{k-\ell-1}$. The claim now follows as this holds for every $\ell = 0, \dots, t - 1$. □

We now show the connection between t-null functions and polynomials to $\{-1, 0, 1\}$.

Theorem 5.2. *If $f \oplus PARITY$ is t-null then the interpolation polynomial of $F(|x|+2) - F(|x|)$ on the range $\{0, 1, \ldots, k-2\}$ is of degree smaller than $k - t - 1$.*

Proof. Let $G(|x|) = F(|x|+2) - F(|x|)$. We compute G's representation in the Newton basis using F's representation. As before, denote $F(|x|) = \sum_{i=0}^{k} \gamma_i \binom{|x|}{i}$. Since $\binom{|x|+2}{i} = \binom{|x|}{i-2} + 2 \cdot \binom{|x|}{i-1} + \binom{|x|}{i}$, we have that

$$\begin{aligned} G(|x|) &= \sum_{i=0}^{k} \gamma_i \cdot \left[\binom{|x|+2}{i} - \binom{|x|}{i} \right] \\ &= \gamma_0 \cdot 0 + \gamma_1 \cdot 2 + \sum_{i=2}^{k} \gamma_i \cdot \left[\binom{|x|}{i-2} + 2 \cdot \binom{|x|}{i-1} + \binom{|x|}{i} - \binom{|x|}{i} \right] \\ &= \gamma_1 \cdot 2 + \sum_{i=2}^{k} \gamma_i \cdot \left[\binom{|x|}{i-2} + 2 \cdot \binom{|x|}{i-1} \right] \\ &= \gamma_1 \cdot 2 + \binom{|x|}{0} \cdot \gamma_2 + \sum_{i=1}^{k-2} \binom{|x|}{i} \cdot (\gamma_{i+2} + 2 \cdot \gamma_{i+1}) + \binom{|x|}{k-1} \cdot 2 \cdot \gamma_k \\ &=^{(*)} \gamma_1 \cdot 2 + \binom{|x|}{0} \cdot \gamma_2 + \sum_{i=1}^{k-t-2} \binom{|x|}{i} \cdot (\gamma_{i+2} + 2 \cdot \gamma_{i+1}) + \binom{|x|}{k-1} \cdot 2 \cdot \gamma_k \, , \end{aligned}$$

where equality $(*)$ follows from Lemma 5.1, as F is t-null. Let $H(|x|)$ be the interpolation polynomial at the points $\{0, 1, \ldots, k-2\}$. In other words, $H(i) = G(i)$ for $i = 0, 1, \ldots, k-2$, and $\deg(H) \le k-2$. By Lemma 2.5 the coefficients of $\binom{|x|}{i}$ in G and H (for $i = 0, 1, \ldots, k-2$) are equal as they depend on the same set of values. Since $\deg(H) \le k-2$ it must be the case that

$$H(|x|) = G(|x|) - \binom{|x|}{k-1} \cdot 2 \cdot \gamma_k = \gamma_1 \cdot 2 + \binom{|x|}{0} \cdot \gamma_2 + \sum_{i=1}^{k-t-2} \binom{|x|}{i} \cdot (\gamma_{i+2} + 2 \cdot \gamma_{i+1}) \, .$$

In particular, $\deg(H) < k - t - 1$. Finally, as $H(i)$ and $G(i)$ agree on $i = 0, 1, \ldots, k-2$ and $G(i) = F(i+2) - F(i)$, we have that H maps $\{0, \ldots, k-2\}$ to $\{-1, 0, 1\}$. □

Strengthening a conjecture of von zur Gathen and Roche [GR97], Mossel et al. [MOS04] conjectured that any non-linear symmetric function must have a Fourier coefficient of size $O(1)$. Theorem 5.2, suggests the following approach.

Corollary 5.3. *If the degree of any non-constant polynomial $H : \{0, \ldots, k-2\} \to \{-1, 0, 1\}$ is at least $k - t$, then every non-linear symmetric function $f : \{0,1\}^k \to \{0,1\}$ must satisfy $\hat{f}(S) \neq 0$ for some non-empty S of size $|S| < t$.*

Proof. Assume for a contradiction that there exists a non-linear symmetric function f which is $(t-1)$-null. Let $g = f \oplus \text{PARITY}$. Hence, $f = g \oplus \text{PARITY}$. Theorem 5.2 implies that the degree of the polynomial agreeing with $G(y+2) - G(y)$ on $\{0, \ldots, k-2\}$ is smaller than $k - t$. By our assumption, it follows that $G(y+2) - G(y)$ is constant on $\{0, \ldots, k-2\}$. Since G only attains the values 0 and 1, it must be the case that $G(y+2) - G(y) = 0$ on $\{0, \ldots, k-2\}$ (assuming[1] $k \ge 4$).

[1] When $k = 3$ the claim follows by inspection, noticing that $t = 2$ and that any symmetric function on 3 bits has a nonzero Fourier coefficient of degree 1. For $k = 2$, while the assumption is meaningless, it is easy to verify that a nonlinear f has a degree 1 nonzero Fourier coefficient.

Hence, G is equal to some constant on all the even elements in $\{0,\ldots,k\}$ and to some (possibly different) constant on all the odd elements there. From the definition of g it follows that f has the same property. This can only happen if f is linear, which contradicts our assumption. □

Thus, if one could prove strong lower bounds on the degree of non-constant polynomials $H : \{0,\ldots,k-2\} \to \{-1,0,1\}$ then one would get improved learning algorithms for symmetric juntas. We note, however, that obtaining better bounds, even when the range of H is $\{0,1\}$ is still open. The best bounds that are currently known are $\deg(H) \geq k - \Gamma(k)$ when $H : \{0,\ldots,k-2\} \to \{0,1\}$ [GR97], and $\deg(H) \geq k - O(\frac{k}{\log\log k})$ when $H : \{0,\ldots,k-2\} \to \{-1,0,1\}$ [CST].

References

[BHP01] R. C. Baker, G. Harman, and J. Pintz. The difference between consecutive primes, II. *Proceedings of the London Mathematical Society*, 83:532–562, 2001.

[BL97] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artif. Intell.*, 97(1-2):245–271, 1997.

[Blu94] A. L. Blum. Relevant examples and relevant features: Thoughts from computational learning theory. In *In AAAI Fall Symposium on 'Relevance*, 1994.

[Cra36] H. Cramér. On the order of magnitude of the difference between consecutive prime numbers. *Acta Arithmetica*, 2:23–46, 1936.

[CST] G. Cohen, A. Shpilka, and A. Tal. On the degree of univariate integer polynomials. This paper supersedes the eralier http://www.eccc.uni-trier.de/report/2010/039/.

[GR97] J. von zur Gathen and J. R. Roche. Polynomials with two values. *Combinatorica*, 17(3):345–362, 1997.

[KLM$^+$09] M. N. Kolountzakis, R. J. Lipton, E. Markakis, A. Mehta, and N. K. Vishnoi. On the fourier spectrum of symmetric boolean functions. *Combinatorica*, 29(3):363–387, 2009.

[MOS04] E. Mossel, R. O'Donnell, and R. A. Servedio. Learning functions of k relevant variables. *J. Comput. Syst. Sci.*, 69(3):421–434, 2004.

[NS94] N. Nisan and M. Szegedy. On the degree of Boolean functions as real polynomials. *Computational Complexity*, 4:301–313, 1994.

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

A Proof of Lemma 2.5

For completeness we give the proof of Lemma 2.5. To ease the reading we repeat it here.

Lemma (Lemma 2.5)**.** *Let $F : \{0,1,\ldots,n\} \to \mathbb{N}$ be a polynomial. Then F can be represented as $F(x) = \sum_{d=0}^{n} \gamma_d \cdot \binom{x}{d}$ and $\gamma_d = \sum_{i=0}^{d} (-1)^{d-i} \cdot \binom{d}{i} \cdot F(i)$.*

Proof. The set of polynomials $\{\binom{x}{0}, \binom{x}{1}, \ldots, \binom{x}{n}\}$ is a basis of the vector space of real polynomials with degree at most n. Thus, we can express F as a linear combination of them. We prove by induction on d that $\gamma_d = \sum_{i=0}^{d} (-1)^{d-i} \cdot \binom{d}{i} \cdot F(i)$. The basis for the induction is $d = 0$. Clearly, $F(0) = \sum_{i=0}^{n} \gamma_0 \cdot \binom{0}{i} = \gamma_0$. We now prove the induction step. The value $F(d)$ is given by

$$F(d) = \sum_{i=0}^{n} \gamma_i \cdot \binom{d}{i} = \sum_{i=0}^{d} \gamma_i \cdot \binom{d}{i} = \gamma_d + \sum_{i=0}^{d-1} \gamma_i \cdot \binom{d}{i} .$$

Rearranging the equation (isolating γ_d) we get

$$\gamma_d = F(d) - \sum_{i=0}^{d-1} \gamma_i \cdot \binom{d}{i} . \tag{10}$$

By the induction assumption we have that $\gamma_\ell = \sum_{j=0}^{\ell} (-1)^{\ell-j} \cdot \binom{\ell}{j} \cdot F(j)$, for $\ell = 0, 1, \ldots, d-1$. Plugging this to Equation (10) we obtain

$$\begin{aligned} \gamma_d &= F(d) - \sum_{i=0}^{d-1} \binom{d}{i} \cdot \sum_{j=0}^{i} (-1)^{i-j} \cdot \binom{i}{j} \cdot F(j) \\ &= F(d) - \sum_{j=0}^{d-1} F(j) \cdot \sum_{i=j}^{d-1} (-1)^{i-j} \cdot \binom{d}{i}\binom{i}{j} \end{aligned}$$

From the identity $\binom{d}{i} \cdot \binom{i}{j} = \binom{d}{j} \cdot \binom{d-j}{i-j}$ it follows that

$$\begin{aligned} \gamma_d &= F(d) - \sum_{j=0}^{d-1} F(j) \cdot \binom{d}{j} \cdot \sum_{i=j}^{d-1} (-1)^{i-j} \cdot \binom{d-j}{i-j} \\ &= F(d) - \sum_{j=0}^{d-1} F(j) \cdot \binom{d}{j} \cdot \sum_{r=0}^{d-1-j} (-1)^{r} \cdot \binom{d-j}{r} \end{aligned} \tag{11}$$

Since $\sum_{r=0}^{d-j} (-1)^r \cdot \binom{d-j}{r} = (1 + (-1))^{d-j} = 0$, we conclude that $\sum_{r=0}^{d-j-1} (-1)^r \cdot \binom{d-j}{r} = -(-1)^{d-j}$. Rewriting Equation (11) we obtain

$$\begin{aligned} \gamma_d &= F(d) + \sum_{j=0}^{d-1} F(j) \cdot \binom{d}{j} \cdot (-1)^{d-j} \\ &= \sum_{j=0}^{d} F(j) \cdot \binom{d}{j} \cdot (-1)^{d-j} , \end{aligned}$$

as required. □

ECCC ISSN 1433-8092

http://eccc.hpi-web.de