Electronic Colloquium on Computational Complexity, Revision 1 of Report No. 49 (2013)

On the Structure of Boolean Functions with Small Spectral Norm

Amir Shpilka[*] Avishay Tal[†] Ben lee Volk[*]

Abstract

In this paper we prove results regarding Boolean functions with small spectral norm (the spectral norm of f is $\|\hat{f}\|_1 = \sum_\alpha |\hat{f}(\alpha)|$). Specifically, we prove the following results for functions $f : \{0,1\}^n \to \{0,1\}$ with $\|\hat{f}\|_1 = A$.

1. There is a subspace V of co-dimension at most A^2 such that $f|_V$ is constant.
2. f can be computed by a parity decision tree of size $2^{A^2} n^{2A}$. (a parity decision tree is a decision tree whose nodes are labeled with arbitrary linear functions.)
3. If in addition f has at most s nonzero Fourier coefficients, then f can be computed by a parity decision tree of depth $A^2 \log s$.
4. For every $0 < \epsilon$ there is a parity decision tree of depth $O(A^2 + \log(1/\epsilon))$ and size $2^{O(A^2)} \cdot \min\{1/\epsilon^2, O(\log(1/\epsilon))^{2A}\}$ that ϵ-approximates f. Furthermore, this tree can be learned, with probability $1 - \delta$, using $\mathsf{poly}(n, \exp(A^2), 1/\epsilon, \log(1/\delta))$ membership queries.

All the results above also hold (with a slight change in parameters) for functions $f : \mathbb{Z}_p^n \to \{0,1\}$.

[*] Faculty of Computer Science, Technion — Israel Institute of Technology, Haifa, Israel, {shpilka,benlee}@cs.technion.ac.il. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 257575.

[†] Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel. Email: avishay.tal@weizmann.ac.il. Research supported by an Adams Fellowship of the Israel Academy of Sciences and Humanities, by an Israel Science Foundation grant and by the I-CORE Program of the Planning and Budgeting Committee and the Israel Science Foundation

ISSN 1433-8092

1 Introduction

The Fourier transform is one of the most useful tools in the analysis of Boolean functions. It is a household name in many areas of theoretical computer science: Learning theory (cf. [KM93, LMN93, Man94]); Hardness of approximation (cf. [Hås01]); Property testing (cf. [BLR93, BCH$^+$96, GOS$^+$11]); Social choice (cf. [KKL88, Kal02]) and more. The reader interested in the Fourier transform and its applications is referred to the online book [O'D12].

A common theme in the study of Fourier transform is the question of classifying all Boolean functions whose Fourier transforms share some natural property. For example, Friedgut proved that Boolean functions that have *small influence* are close to being juntas (i.e. functions that depend on a small number of coordinates) [Fri98]. Friedgut, Kalai and Naor proved that Boolean functions whose Fourier spectrum is concentrated on the first two levels are close to dictator functions (i.e. functions of the form $f(x_1, \ldots, x_n) = x_i$ or $1 - x_i$). In [ZS10, MO09] it was conjectured that a Boolean function that has a *sparse* Fourier spectrum (i.e. that has only s nonzero Fourier coefficients), can be computed by a parity decision tree (for short we denote parity decision tree by $\oplus$-DT) of depth $\mathsf{poly}(\log s)$. Recall that in a $\oplus$-DT nodes are labeled by linear functions (over $\mathbb{Z}_2$) rather than by variables. It is well known that a function that is computed by a depth d $\oplus$-DT has sparsity at most $\exp(d)$ (see Lemma 2.5), so this conjecture implies a (more or less) tight result. This conjecture was raised in the context of the log-rank conjecture in communication complexity and, if true, it would imply that the log-rank conjecture is true for functions of the form $F(x, y) = f(x \oplus y)$, for some Boolean function f.

In this paper we are interested in the structure of functions that have small spectral norm. Namely, in Boolean functions $f : \{0, 1\}^n \to \{0, 1\}$ that for some number A satisfy

$$\|\hat{f}\|_1 \stackrel{\text{def}}{=} \sum_{\alpha} |\hat{f}(\alpha)| \leq A \,, \tag{1}$$

where A may depend on the number of variables n (for definitions see Section 2). Such functions were studied in the context of circuit complexity (cf. [Gro97]) and, more notably, in learning theory, where it is one of the most general family of Boolean functions that can be learned efficiently [KM93, Man94, ABF$^+$08]. In particular, Kushilevitz and Mansour proved that any Boolean function satisfying (1), can be well approximated by a sparse polynomial [KM93]. This already gives some rough structure for functions with small spectral norm, however one may ask for a more refined structure that captures the function exactly. Green and Sanders were the first to obtain such a result (and until this work this was the only such result). They proved that if f satisfies Equation (1) then it can be expressed as a sum of at most $2^{2^{O(A^4)}}$ characteristic functions of subspaces, that is,

$$f = \sum_{i=1}^{2^{2^{O(A^4)}}} \pm \mathbb{1}_{V_i}, \tag{2}$$

where each V_i is a subspace. Thus, when A is constant this gives a very strong result on the structure of such a function f. This result can be seen as an *inverse* theorem, as it is well known and easy to see that the spectral norm of the characteristic function of a subspace is constant. Thus, [GS08a] show that in general, any function with a small spectral norm is a linear combination of a (relatively) small number of such characteristic functions. Of course, ideally one would like to show that the number of functions in the sum is at most $\mathsf{poly}(A)$ and not doubly exponential in A, however, Green and Sanders note that "it seems to us that it would be difficult to use our method to reduce the number of exponentials below two."

It is possible that another classification of Boolean functions with small spectral norm could be achieved using decision trees, or more generally, parity decision trees. It is not hard to show that if a Boolean function g is computed by a $\oplus$-DT with s leaves then the spectral norm of g is at most s (see Lemma 2.5). Interestingly, we are not aware of any Boolean function that has a small spectral norm and that cannot be computed by a small $\oplus$-DT. It is thus an interesting question whether this is indeed the general case, namely, that any function of small spectral norm can be computed by a small $\oplus$-DT. We note that the result of [GS08a] does not yield such a structure. Indeed, if we were to represent the function given by Equation (2) as a $\oplus$-DT then, without knowing anything more about the function, then we do not see a more efficient representation than the brute-force one that yields a $\oplus$-DT of size $n^{2^{2^{O(A^4)}}}$.

Another interesting question concerning functions with small spectral norm comes from the learning theory perspective. As mentioned above, Kushilevitz and Mansour proved that for any Boolean function satisfying Equation (1) there is some sparse polynomial $g = \sum_{i=1}^{A^2/\epsilon} \hat{f}(\alpha_i)\chi_{\alpha_i}(x)$ (where the coefficients in the summation are the A^2/ϵ largest Fourier coefficient of f) such that $\Pr_x[f(x) \neq \mathrm{sgn}(g(x)] \leq \epsilon$. Thus, their learning algorithm outputs as hypothesis the function $\mathrm{sgn}(g(x))$. This is the case even if f is computed by a small decision tree or a small $\oplus$-DT. It would be desirable to output a hypothesis coming from the same complexity class as f, i.e. to output a decision tree or a $\oplus$-DT. However, a hardness result of [ABF$^+$08] shows that under reasonable complexity assumptions, one cannot hope to output a small decision tree approximating f. So, a refinement of the question should be to try and output the smallest tree one can find for a function approximating f. For example, the function

$$\mathrm{sgn}(g) = \mathrm{sgn}\left(\sum_{i=1}^{A^2/\epsilon} \hat{f}(\alpha_i)\chi_{\alpha_i}(x)\right) \tag{3}$$

can be computed by a $\oplus$-DT of depth $O(A^2/\epsilon)$ in the natural way. Even when A is a constant and ϵ is polynomially small this does not give much information. Thus, a natural question is to try and find a better representation for such a range of parameters.

1.1 Our results

Our first result identifies a *local* structure shared by Boolean functions with small spectral norm.

Theorem 1.1. *Let $f : \{0,1\}^n \to \{0,1\}$ be such that $\|\hat{f}\|_1 = A$, then, there is an affine subspace $V \subset \{0,1\}^n$ of co-dimension at most A^2 such that f is constant on V.*

We note that the proof of [GS08a] does not imply the existence of such an affine subspace V of such a high dimension. Our next result gives a $\oplus$-DT computing f.

Theorem 1.2. *Let $f : \{0,1\}^n \to \{0,1\}$ be such that $\|\hat{f}\|_1 = A$, then, f can be computed by a $\oplus$-DT of size $2^{A^2}n^{2A}$.*

In particular, the theorem implies that $f = \sum_{i=1}^{2^{A^2}n^{2A}} \pm \mathbb{1}_{V_i}$, where each V_i is a subspace.

Another result settles the conjecture of [ZS10, MO09] for the case of sparse Boolean functions with small spectral norm.

Theorem 1.3. *Let $f : \{0,1\}^n \to \{0,1\}$ be such that $\|\hat{f}\|_1 = A$ and $|\{\alpha \mid \hat{f}(\alpha) \neq 0\}| = s$. Then f can be computed by a $\oplus$-DT of depth $A^2 \log s$.*

Thus, if the spectral norm of f is constant (or $\mathsf{poly}(\log s)$), Theorem 1.3 settles the conjecture affirmatively. The conjecture is still open for the case where the spectral norm of f is large.

Our last result (for functions over the Boolean cube) fits into the context of learning theory and provides a bound on the depth of a $\oplus$-DT *approximating* a function with a small spectral norm. Here, the distance between two Boolean functions is measured with respect to the uniform distribution, namely, $\mathrm{dist}(f,g) = \Pr_{x\in\{0,1\}^n}[f(x) \neq g(x)]$.

Theorem 1.4. *Let $f : \{0,1\}^n \to \{0,1\}$ be such that $\|\hat{f}\|_1 = A$. Then for every $\delta, \epsilon > 0$ there is a randomized algorithm that, given a query oracle to f, outputs (with probability at least $1-\delta$) a $\oplus$-DT of depth $O(A^2 + \log(1/\epsilon))$ and size $2^{O(A^2)} \min\{1/\epsilon^2, O(\log(1/\epsilon))^{2A}\}$, which computes a Boolean function g_ϵ such that dist$(f, g_\epsilon) \leq \epsilon$. The algorithm runs in time polynomial in $n, \exp(A^2), 1/\epsilon$ and $\log(1/\delta)$.*

Thus, when A is a constant and ϵ is polynomially small, the depth is $O(\log n)$ and the size is only poly-logarithmic in n. This greatly improves upon the representation guaranteed by Equation (3). If one insists on outputting a $\oplus$-DT, then, for all ranges of parameters, the tree that we obtain is much smaller than the tree guaranteed by Equation (3).

We also prove analogs of the theorems above for functions $f : \mathbb{Z}_p^n \to \{+1,-1\}$ having small spectral norm. Namely, in the theorems above one could instead talk of $f : \mathbb{Z}_p^n \to \{0,1\}$ and obtain essentially the same results.[1] Theorems 4.7, 4.8, 4.10 and 4.11 are the $\mathbb{Z}_p$ analogs to Theorems 1.1, 1.2, 1.3 and 1.4, respectively. We note that in [GS08b] Green and Sanders extended their result to hold for functions mapping an abelian group G to $\{0,1\}$, obtaining the same bound as in [GS08a], so our result for functions on $\mathbb{Z}_p^n$ could be seen as an analog to their result for such groups.

1.2 Comparison with [GS08a]

Comparing Theorem 1.2 to Equation (2) (that was proved in [GS08a]), we note that while Equation (2) does not involve the number of variables (i.e. the upper bound on the number of subspaces only involves A), our result does involve n. On the other hand, we give a more refined structure - that of a parity decision tree - which is not implied by Equation (2) (see also the discussion above). Moreover, when $A = \Omega((\log\log n)^{1/4})$, our bound is much better than the one given in Equation (2).

Our proof technique is also quite different than that of [GS08a]. Their proof idea is to represent f as $f = f_1 + f_2$ where the Fourier supports of f_1 and f_2 are disjoint, and such that f_1 and f_2 are *close to being integer valued* and have a somewhat smaller spectral norm. Then, using recursion, they represent each f_i as a sum of a small number of characteristic functions of subspaces. In particular, Green and Sanders do not restrict their treatment to Boolean functions but rather study functions that at every point of the Boolean cube obtain a value that is almost an integer. Thus, they prove a more general result, namely, that $f_{\mathbb{Z}}$, the integer part of f, can be represented in the form of Equation (2). We on the other hand only work with Boolean functions, so their result is stronger from that respect. However, while their proof was a bit involved and required using results from additive combinatorics, our approach is more elementary and is based on exploiting the fact that f is Boolean. In particular, our starting point is an analysis of the simple equation $f^2 = 1$ (when we think of f as mapping $\{0,1\}^n$ to $\{\pm 1\}$). Furthermore, we are able to use the fact that f is Boolean in order to show that it can be computed by a small $\oplus$-DT, which does not seem to

[1] Of course, one would have to speak about the analog of a $\oplus$-DT for the case where the inputs come from $\mathbb{Z}_p^n$.

follow from [GS08a].

Green and Sanders later extended their technique and proved a similar result for functions over general abelian groups $f : G \to \{0, 1\}$ [GS08b]. Our technique do not extend to general groups, but we do obtain results for the case that $G = \mathbb{Z}_p^n$, which again has the same advantages and disadvantages compared to the result of [GS08b] (although, the simplicity of our approach is even more evident here).

1.3 Proof idea

As mentioned above, our proof relies on the simple equation $f^2 = 1$ (when we think of $f : \{0, 1\}^n \to \{\pm 1\}$). By expanding the Fourier representations (See Section 2 for definitions) of both sides we reach the identity

$$\sum_{\gamma} \hat{f}(\gamma)\hat{f}(\delta + \gamma) = 0,$$

that holds for all $\delta \neq 0$ (See Lemma 3.2). This identity could be interpreted as saying that the mass on pairs whose product is positive is the same as the mass on pairs whose product is negative. In particular, if we consider the two heaviest elements in the Fourier spectrum, say, $\hat{f}(\alpha)$ and $\hat{f}(\beta)$, and let $\delta = \alpha + \beta$, then by restricting f to one of the subspaces $\chi_\delta(x) = 1$ or $\chi_\delta(x) = -1$, we get a substantial saving in the spectral norm (see Lemma 3.1). This happens since there is a significant L_1 mass on pairs $\hat{f}(\gamma), \hat{f}(\delta + \gamma)$ that have different signs. By repeating this process we manage to prove the existence of small $\oplus$-DT for f.

The argument for functions over $\mathbb{Z}_p^n$ is similar, but requires more technical work. For that reason we decided to give a separate proof for the case of functions over the Boolean cube, and then, after the ideas were laid out in their simpler form, to prove the results in the more general case.

1.4 The work of Tsang et al. [TWXZ13]

Independently and simultaneously to our work, Tsang et al. [TWXZ13] obtained related results. The main objective of the work [TWXZ13] was to study the communication complexity of sparse Boolean functions. These are functions f such that the *communication matrix* of the function $F(x, y) = f(x \oplus y)$ has low rank. Resolving the log-rank conjecture from communication complexity for such functions was the main motivation for the conjecture raised in [MO09] and [ZS10].

Tsang et al. managed to prove a stronger version of our Theorem 1.1, namely, they proved that f is constant on a subspace of co-dimension at most $O(A)$. Their argument is identical to ours (namely, to the one given in Lemma 3.1) except that they observe that after $O(1/A)$ steps of increasing the largest Fourier coefficient of f, it grows to at least $1/2$. From that point on they make use of the simple observation that the proof of (their equivalent of) Lemma 3.1 actually guarantees that the restriction that saves the most in the spectral norm keeps increasing the largest coefficient. Thus, now at each step the spectral norm goes down by some constant factor and hence additional $O(1/A)$ many steps would make f constant.[2]

This immediately improves the results in Theorems 1.1 and 1.3; we can now change the factor A^2 to A in both.

The work [TWXZ13] does not contain analogs for Theorems 1.2 and 1.4. We also note that Tsang et al. did not study the case of functions from $\mathbb{Z}_p^n$ to $\{0, 1\}$, and so they do not have analogs of Theorems 4.7, 4.8, 4.10 and 4.11.

[2] Our Lemma 3.1 only speaks about the spectral norm, but the effect on the largest Fourier coefficient is obvious from the proof.

1.5 Organization

Section 2 contains the basic background and definitions. In Section 3 we prove our results for functions $f : \mathbb{Z}_2^n \to \{+1, -1\}$. The results for functions on $\mathbb{Z}_p^n$ are given in Section 4. Finally, in Section 5 we discuss problems left open by this work.

2 Notation and Basic Results

It will be more convenient for us to talk about functions $f : \{0,1\}^n \to \{\pm 1\}$. Note that if $f : \{0,1\}^n \to \{0,1\}$ then $1-2f : \{0,1\}^n \to \{\pm 1\}$ and $1-2f$ and f have roughly the same spectral norm (up to a multiplicative factor of 2) and the same Fourier sparsity (up to ± 1).

2.1 Decision trees and parity decision trees

In this section we define the basic computational models that we shall consider in the paper.

Definition 2.1 (Decision tree)**.** *A decision tree is a labeled binary tree T. Each internal node of T is labeled with a variable x_i, and each leaf by a bit $b \in \{+1, -1\}$. Given an input $x \in \mathbb{Z}_2^n$, a computation over the tree is executed as follows: Starting at the root, stop if it's a leaf, and output its label. Otherwise, query its label x_i. If $x_i = 0$, then recursively evaluate the left subtree, and if $x_i = 1$, evaluate the right subtree.*

A decision tree T computes a function f if for every $x \in \mathbb{Z}_2^n$, the computation of x over T outputs $f(x)$. The *depth* of a decision tree is the maximal length of a path from the root to a leaf. The decision tree complexity of f, denoted $D(f)$, is the depth of a minimal-depth tree computing f. Since one can always simply query all the variables of the input, it holds that for any Boolean function f, $D(f) \leq n$. A comprehensive survey of decision tree complexity can be found in [BdW02].

In the context of Fourier analysis, even a function with simple Fourier spectrum, such as the parity function over n bits, which has only 1 nonzero Fourier coefficient, requires a full binary decision tree for its computation, and in particular its depth is n. This example suggests that a more suitable computational model for understanding the connection between the computational complexity and the Fourier expansion of a function is the *parity decision tree* model, first presented by Kushilevitz and Mansour ([KM93]).

Definition 2.2 ($\oplus$-DT)**.** *A parity decision tree is a labeled binary tree T, in which every internal node is labeled by a linear function $\alpha \in \mathbb{Z}_2^n$, and each leaf with a bit $b \in \{+1, -1\}$. Whenever a computation over an input x arrives at an internal node, it queries $\langle \alpha, x \rangle$ (where the inner product is carried modulo 2). If $\langle \alpha, x \rangle = 0$ it recursively evaluates the left subtree, and if $\langle \alpha, x \rangle = 1$, it evaluates the right subtree. When the computation reaches a leaf it outputs its label.*

Namely, a $\oplus$-DT can make an arbitrary linear query in every internal node (and in particular, compute the parity of n bits using a single query). Since a query of a single variable is linear, this model is an extension of the regular decision tree model.

The depth of the minimal-depth parity decision tree which computes f is denoted $D^{\oplus}(f)$, thus $D^{\oplus}(f) \leq D(f)$. As the example of the parity function shows, the parity decision tree model is strictly stronger than the model of decision trees. We also denote by $\text{size}_{\oplus}(f)$ the size (i.e. number of leaves) of a minimal-size $\oplus$-DT computing f.

As a helpful tool, we extend the parity decision tree model to a *functional parity decision tree* model, in which we allow every leaf to be labeled with a Boolean function, rather than only by a constant. A functional $\oplus$-DT T then computes a function f if for every leaf ℓ of T, its label equals the restriction of f to the affine subspace defined by the constraints that appear on the path from T's root to ℓ.

2.2 Fourier Transform

We represent Boolean functions as functions $f : \mathbb{Z}_2^n \to \{+1, -1\} \subseteq \mathbb{R}$ where -1 represents the Boolean value "True" and 1 represents the Boolean value "False". For a vector of n bits α, α_i denotes its i-th coordinate. The set of 2^n group characters $\{\chi_\alpha : \mathbb{Z}_2^n \to \{+1, -1\} \mid \alpha \in \mathbb{Z}_2^n\}$, with $\chi_\alpha(x) = (-1)^{\sum_{i=1}^n \alpha_i x_i}$ for every $\alpha \in \mathbb{Z}_2^n$, forms a basis of the vector space of functions from $\mathbb{Z}_2^n$ into $\mathbb{R}$. Furthermore, the basis is orthonormal with respect to the inner product[3]

$$\langle f, g\rangle = \mathop{\mathbb{E}}_x [f(x)g(x)]$$

where the expectation is taken over the uniform distribution over $\mathbb{Z}_2^n$. The *Fourier expansion* of a function $f : \mathbb{Z}_2^n \to \{+1, -1\}$ is its unique representation as a linear combination of those group characters:

$$f(x) = \sum_{\alpha\in\mathbb{Z}_2^n} \hat{f}(\alpha)\chi_\alpha(x).$$

Two of the basic identities of Fourier analysis, which follow from the orthonormality of the basis, are:

1. $\hat{f}(\alpha) = \langle f, \chi_\alpha\rangle = \mathbb{E}_x [f(x)\chi_\alpha(x)]$
2. (Plancherel's Theorem) $\langle f, g\rangle = \mathbb{E}_x [f(x)g(x)] = \sum_{\alpha\in\mathbb{Z}_2^n} \hat{f}(\alpha)\hat{g}(\alpha)$.

The case $f = g$ in Plancherel's theorem is called *Parseval's Identity.* Furthermore, when f is Boolean, $f^2 = 1$, which implies

$$\sum_{\alpha\in\mathbb{Z}_2^n} \hat{f}(\alpha)^2 = 1. \tag{4}$$

We define two basic complexity measures for Boolean functions:

Definition 2.3. *Let $f : \mathbb{Z}_2^n \to \{+1, -1\}$ be a Boolean function. The* sparsity *of f, denoted* $\mathrm{spar}(f)$*, is the number of non-zero Fourier coefficients, namely*

$$\mathrm{spar}(f) = \#\left\{\alpha \in \mathbb{Z}_2^n \mid \hat{f}(\alpha) \neq 0\right\}.$$

A function f is said to be *s-sparse* if $\mathrm{spar}(f) \leq s$.

Definition 2.4. *Let $f : \mathbb{Z}_2^n \to \{+1, -1\}$ be a Boolean function. The L_1 norm (also dubbed the* spectral norm*) of f is defined as*

$$\|\hat{f}\|_1 = \sum_{\alpha\in\mathbb{Z}_2^n} |\hat{f}(\alpha)|.$$

For every $f : \mathbb{Z}_2^n \to \{+1, -1\}$ it holds that $\|\hat{f}\|_1 \geq \|f\|_\infty = 1$ (where $\|f\|_\infty = \max_{x\in\mathbb{Z}_2^n} |f(x)|$). We later show (Lemma 3.5) that equality is obtained if and only if $f = \pm\chi_\alpha$ for some $\alpha \in \mathbb{Z}_2^n$.

These measure are related to parity decision trees using the following simple lemma. For completeness we give the proof of the lemma in Appendix A.

[3] Later when we study of functions over $\mathbb{Z}_p^n$ we define the inner product to be $\mathbb{E}_x\left[f(x)\overline{g(x)}\right]$.

Lemma 2.5. *Let $f : \mathbb{Z}_2^n \to \{+1, -1\}$ be a Boolean function computed by a $\oplus$-DT T of depth k and size m. Then:*

1. $\mathrm{spar}(f) \le m2^k \le 4^k$.
2. $\|\hat{f}\|_1 \le m \le 2^k$.

In the upcoming sections we consider restrictions of Boolean functions to (affine) subspaces of $\mathbb{Z}_2^n$. We denote by $f|_V$ the restriction of f to a subspace $V \subseteq \mathbb{Z}_2^n$. For any $\alpha \neq 0$, the set $\{x \mid \chi_\alpha(x) = 1\}$ is a subspace of $\mathbb{Z}_2^n$ of co-dimension 1. The restriction of f to this subspace is denoted $f|_{\chi_\alpha=1}$. Similarly, the set $\{x \mid \chi_\alpha(x) = -1\}$ is an affine subspace of co-dimension 1, and we denote with $f|_{\chi_\alpha=-1}$ the restriction of f to this subspace. It can be shown (cf. [O'D12], Chapter 3, Section 3.3) that under such a restriction, the coefficients $\hat{f}(\beta)$ and $\hat{f}(\alpha+\beta)$ (for every $\beta \in \mathbb{Z}_2^n$) collapse to a single Fourier coefficient whose absolute value is $|\hat{f}(\beta) + \hat{f}(\alpha+\beta)|$. Similarly, in the Fourier transform of $f|_{\chi_\alpha=-1}$, they collapse to a single coefficient whose absolute value is $|\hat{f}(\beta) - \hat{f}(\alpha+\beta)|$. This in particular implies that $\|\hat{f}\|_1$ and $\mathrm{spar}(f)$ do not increase when f is restricted to such a subspace. Indeed, both facts follow easily from the representation

$$f(x) = \sum_{\beta \in \mathbb{Z}_2^n/\langle\alpha\rangle} \left(\hat{f}(\beta) + \hat{f}(\beta+\alpha)\chi_\alpha(x)\right) \chi_\beta(x) \,, \tag{5}$$

where $\mathbb{Z}_2^n/\langle\alpha\rangle$ denotes the cosets of the group $\langle\alpha\rangle = \{0, \alpha\}$ in $\mathbb{Z}_2^n$. When studying a restricted function, say $f' = f|_{\chi_\alpha(x)=1}$, we shall abuse notation and denote with $\widehat{f'}(\beta)$ the term corresponding to the coset $\beta + \langle\alpha\rangle$. Namely, $\widehat{f'}(\beta) = \hat{f}(\beta) + \hat{f}(\beta+\alpha)$. (similarly, for $f'' = f|_{\chi_\alpha(x)=-1}$, we shall denote $\widehat{f''}(\beta) = \hat{f}(\beta) - \hat{f}(\beta+\alpha)$.) Thus, in f' both $\widehat{f'}(\beta)$ and $\widehat{f'}(\beta+\alpha)$ refer to the same Fourier coefficient as we only consider coefficients modulo $\langle\alpha\rangle$ (similarly for f'').

3 Boolean functions with small spectral Norm

In this section we prove our main results for functions over the Boolean cube. While many of the proofs and techniques used for general primes also apply to the case $p = 2$, we find the case $p = 2$ substantially simpler, so we present the proofs for this case separately.

3.1 Basic tools

In this section we prove the following lemma, which states that for every Boolean function $f : \mathbb{Z}_2^n \to \{+1, -1\}$, with small spectral norm, there exists a linear function χ_γ such that both restrictions $f|_{\chi_\gamma=1}$ and $f|_{\chi_\gamma=-1}$ have noticeable smaller spectral norms compared to f. In Section 4 we give a generalization of the lemma for functions $f : \mathbb{Z}_p^n \to \{+1, -1\}$ (Lemma 4.1).

Lemma 3.1 (Main Lemma for functions over $\mathbb{Z}_2^n$)**.** *Let $f : \mathbb{Z}_2^n \to \{+1, -1\}$ be a Boolean function. Let $\hat{f}(\alpha)$ be f's maximal Fourier coefficient in absolute value, and $\hat{f}(\beta)$ be the second largest, and suppose $\hat{f}(\beta) \neq 0$. Let $f' = f|_{\chi_{\alpha+\beta}=1}$ and $f'' = f|_{\chi_{\alpha+\beta}=-1}$. Then, if $\hat{f}(\alpha)\hat{f}(\beta) > 0$ then it holds that*

$$\|\hat{f'}\|_1 \le \|\hat{f}\|_1 - |\hat{f}(\alpha)| \quad \textit{and} \quad \|\hat{f''}\|_1 \le \|\hat{f}\|_1 - |\hat{f}(\beta)|.$$

If $\hat{f}(\alpha)\hat{f}(\beta) < 0$ then

$$\|\hat{f'}\|_1 \le \|\hat{f}\|_1 - |\hat{f}(\beta)| \quad \textit{and} \quad \|\hat{f''}\|_1 \le \|\hat{f}\|_1 - |\hat{f}(\alpha)|.$$

The proof of the lemma follows from analyzing the simple equation $f^2 = 1$.

Lemma 3.2. *Let $f : \mathbb{Z}_2^n \to \{+1, -1\}$ be a Boolean function. For all $\alpha \neq 0$, it holds that*

$$\sum_{\gamma} \hat{f}(\gamma)\hat{f}(\alpha+\gamma) = 0.$$

Proof. Since f is Boolean we have that $f^2 = 1$. In the Fourier representation,

$$\left(\sum_{\gamma} \hat{f}(\gamma)\chi_{\gamma}(x)\right)\left(\sum_{\beta} \hat{f}(\beta)\chi_{\beta}(x)\right) = 1.$$

Then $\sum_{\gamma} \hat{f}(\gamma)\hat{f}(\alpha+\gamma)$ is the Fourier coefficient $\widehat{f^2}(\alpha)$ of the function f^2 at α. However, if $\alpha \neq 0$ then this coefficient equals 0 by the uniqueness of the Fourier expansion of the function $f^2 = 1$. □

Proof of Lemma 3.1. Without loss of generality assume that $\hat{f}(\alpha)\hat{f}(\beta) > 0$, i.e. they have the same sign (the other case is completely analogous.) By Lemma 3.2,

$$\sum_{\gamma \in \mathbb{Z}_2^n} \hat{f}(\gamma)\hat{f}(\alpha+\beta+\gamma) = 0. \tag{6}$$

Let $N_{\alpha+\beta} \subseteq \mathbb{Z}_2^n$ be the set of vectors γ such that $\hat{f}(\gamma)\hat{f}(\alpha+\beta+\gamma) < 0$ (Note that by assumption, $\alpha, \beta \notin N_{\alpha+\beta}$). Switching sides in (6), we get:

$$2\left|\hat{f}(\alpha)\hat{f}(\beta)\right| = \sum_{\gamma \in N_{\alpha+\beta}} \left|\hat{f}(\gamma)\hat{f}(\alpha+\beta+\gamma)\right| - \sum_{\substack{\gamma \notin N_{\alpha+\beta} \\ \gamma \neq \alpha,\beta}} \left|\hat{f}(\gamma)\hat{f}(\alpha+\beta+\gamma)\right|.$$

In particular,

$$|\hat{f}(\alpha)||\hat{f}(\beta)| \leq \frac{1}{2} \sum_{\gamma \in N_{\alpha+\beta}} \left|\hat{f}(\gamma)\hat{f}(\alpha+\beta+\gamma)\right|. \tag{7}$$

We now use the fact that that $\hat{f}(\beta)$ is the second largest in absolute value, and $\hat{f}(\alpha)$ does not appear in the sum, to bound the right hand side:

$$\sum_{\gamma \in N_{\alpha+\beta}} \left|\hat{f}(\gamma)\hat{f}(\alpha+\beta+\gamma)\right| \leq |\hat{f}(\beta)| \sum_{\gamma \in N_{\alpha+\beta}} \min\left\{|\hat{f}(\gamma)|, |\hat{f}(\alpha+\beta+\gamma)|\right\}. \tag{8}$$

Then (7) and (8) (as well as the assumption $|\hat{f}(\beta)| > 0$) together imply

$$|\hat{f}(\alpha)| \leq \frac{1}{2} \sum_{\gamma \in N_{\alpha+\beta}} \min\left\{|\hat{f}(\gamma)|, |\hat{f}(\alpha+\beta+\gamma)|\right\}. \tag{9}$$

Let $f' = f|_{\chi_{\alpha+\beta}=1}$. Then for every γ the coefficients $\hat{f}(\gamma)$ and $\hat{f}(\alpha+\beta+\gamma)$ collapse to a single coefficient whose absolute value is $|\hat{f}(\gamma) + \hat{f}(\alpha+\beta+\gamma)|$ (recall Equation (5)). For $\gamma \in N_{\alpha+\beta}$,

$$|\hat{f}(\gamma) + \hat{f}(\alpha+\beta+\gamma)| = \left||\hat{f}(\gamma)| - |\hat{f}(\alpha+\beta+\gamma)|\right|$$

which reduces the L_1 norm of f' compared to that of f by at least $\min(|\hat{f}(\gamma)|, |\hat{f}(\alpha+\beta+\gamma)|)$. In total, since both γ and $\alpha+\beta+\gamma$ belong to $N_{\alpha+\beta}$, we get:

$$\|\widehat{f'}\|_1 \leq \|\hat{f}\|_1 - \frac{1}{2}\sum_{\gamma\in N_{\alpha+\beta}} \min\left\{|\hat{f}(\gamma)|, |\hat{f}(\alpha+\beta+\gamma)|\right\}.$$

Therefore by (9) we have

$$\|\widehat{f'}\|_1 \leq \|\hat{f}\|_1 - |\hat{f}(\alpha)|.$$

When we consider $f'' = f|_{\chi_{\alpha+\beta}=-1}$ we clearly have that for $\gamma = \alpha$,

$$|\widehat{f''}(\gamma)| = |\hat{f}(\gamma) - \hat{f}(\alpha+\beta+\gamma)| = |\hat{f}(\alpha)| - |\hat{f}(\beta)|.$$

Hence,

$$\|\widehat{f''}\|_1 \leq \|\hat{f}\|_1 - |\hat{f}(\beta)|.$$

□

Next, we show that any Boolean function with small spectral norm has a large Fourier coefficient.

Lemma 3.3. *Let $f : \mathbb{Z}_2^n \to \{+1,-1\}$ be a Boolean function. Denote $A = \|\hat{f}\|_1$, and let $\hat{f}(\alpha)$ be f's maximal Fourier coefficient in absolute value. Then $|\hat{f}(\alpha)| \geq 1/A$. Furthermore, let $\hat{f}(\beta)$ be f's second largest Fourier coefficient in absolute value. Then $|\hat{f}(\beta)| > (1-\hat{f}(\alpha)^2)/\|\hat{f}\|_1 = (1-\hat{f}(\alpha)^2)/A$.*

Proof. By Parseval's identity,

$$1 = \mathbb{E}[f^2] = \sum_{\gamma} \hat{f}(\gamma)^2.$$

Now note that

$$1 = \sum_{\gamma} \hat{f}(\gamma)^2 \leq |\hat{f}(\alpha)| \sum_{\gamma} |\hat{f}(\gamma)| \leq A|\hat{f}(\alpha)|,$$

which implies that indeed $|\hat{f}(\alpha)| \geq 1/A$. The second statement follows similarly, since

$$1 - \hat{f}(\alpha)^2 = \sum_{\gamma\neq\alpha} \hat{f}(\gamma)^2 \leq |\hat{f}(\beta)| \sum_{\gamma\neq\alpha} |\hat{f}(\gamma)| < \|\hat{f}\|_1 \cdot |\hat{f}(\beta)| = A|\hat{f}(\beta)|.$$

□

Corollary 3.4. *Let $f : \mathbb{Z}_2^n \to \{+1,-1\}$ be a Boolean function such that $\|\hat{f}\|_1 = A > 1$. Then there exists $\gamma \in \mathbb{Z}_2^n$ and $b \in \{+1,-1\}$ such that $\|\widehat{f|_{\chi_\gamma=b}}\|_1 \leq A - 1/A$.*

Proof. The assumption $A > 1$ implies the second largest coefficient, $\hat{f}(\beta)$, is non-zero, and then the result is immediate from Lemma 3.1 and Lemma 3.3. □

3.2 Proofs of Theorems

We now show how Theorems 1.1,1.2,1.3 and 1.4 follow as simple consequences of Lemma 3.1.

Lemma 3.5. *Let $f : \mathbb{Z}_2^n \to \{+1,-1\}$ be a Boolean function such that $\|\hat{f}\|_1 = 1$. Then $f = \pm\chi_\alpha$ for some $\alpha \in \mathbb{Z}_2^n$.*

Proof. By Parseval's identity and the assumption, we get

$$\sum_{\gamma} \hat{f}(\gamma)^2 = 1 = \sum_{\gamma} |\hat{f}(\gamma)|.$$

For all γ we have that $|\hat{f}(\gamma)| \in [0,1]$, so $|\hat{f}(\gamma)| < \hat{f}(\gamma)^2$ unless $|\hat{f}(\gamma)| = 1$ or $\hat{f}(\gamma) = 0$, and the proposition follows. □

Corollary 3.4 and Lemma 3.5 imply Theorem 1.1:

Proof of Theorem 1.1. Apply Corollary 3.4 iteratively on f. After less than A^2 steps, we are left with a function g which is a restriction of f on an affine subspace defined by the restrictions so far, such that $\|\hat{g}\|_1 = 1$. By Lemma 3.5, $g = \pm\chi_\alpha$ for some $\alpha \in \mathbb{Z}_2^n$. If $\alpha \neq 0$ we further restrict g on $\chi_\alpha = 1$ to get a restriction of f which is constant. □

We note that the proof of Theorem 1.1 actually implies that f is constant on a subspace of co-dimension at most $\binom{A+1}{2}$. As mentioned earlier, a slight twist in the proof improves the co-dimension to $O(A)$ [TWXZ13].

Proof of Theorem 1.2. Let

$$L(n,A) \overset{\text{def}}{=} \max_{\substack{f:\mathbb{Z}_2^n \to \{+1,-1\} \\ \|\hat{f}\|_1 \leq A}} \text{size}_{\oplus}(f).$$

We show, by induction on n, that $L(n,A) \leq 2^{A^2} \cdot n^{2A}$.

For $n = 1$ the result is trivial.

Let $n > 1$ and further assume that $A > 1$ (if $A = 1$ then the claim follows from Lemma 3.5). Let $\hat{f}(\alpha), \hat{f}(\beta)$ be the first and second largest Fourier coefficients in absolute value, respectively. By Lemma 3.3 we are in one of the following cases:

1. $|\hat{f}(\alpha)| \geq 1/2$

2. $1/2 > |\hat{f}(\alpha)| \geq 1/A$ and $|\hat{f}(\beta)| > \frac{1-\hat{f}(\alpha)^2}{A} \geq \frac{3}{4A}$.

Consider the tree whose first query is the linear function χ_γ where $\gamma = \alpha + \beta$ (i.e. we branch left or right according to the value of $\langle x, \gamma \rangle$). By the choice of γ we obtain the following recursion: In case 1,

$$L(n,A) \leq L(n-1, A-1/2) + L(n-1, A);$$

While in case 2,

$$L(n,A) \leq L(n-1, A-1/A) + L(n-1, A-3/(4A)).$$

Note also that in the second case $A \geq 2$, or else $|\hat{f}(\alpha)| \geq 1/2$ by Lemma 3.3. Induction follows in the first case as

$$\begin{aligned} L(n,A) &\leq L(n-1, A-1/2) + L(n-1, A) \\ &\leq 2^{(A-1/2)^2} \cdot (n-1)^{2(A-1/2)} + 2^{A^2} \cdot (n-1)^{2A} \\ &\leq 2^{A^2} \cdot (n-1)^{2(A-1/2)} \left(1 + (n-1)\right) \\ &< 2^{A^2} \cdot n^{2A} \ . \end{aligned}$$

In the second case we have

$$\begin{aligned} L(n,A) &\leq L(n-1, A-1/A) + L(n-1, A-3/(4A)) \\ &\leq 2^{(A-1/A)^2}\cdot (n-1)^{2(A-1/A)} + 2^{(A-3/(4A))^2}\cdot (n-1)^{2(A-3/(4A))} \\ &\leq 2^{A^2}\cdot n^{2A}\left(2^{-2+1/A^2} + 2^{-3/2+(3/4)^2/A^2}\right) \\ &\leq 2^{A^2}\cdot n^{2A}\ , \end{aligned}$$

where in the last inequality we used the fact that $A \geq 2$. □

As the AND function demonstrates, this argument gives a result that is tight up to a polynomial factor in some cases.

Proof of Theorem 1.3. By Theorem 1.1, there exist A^2 linear functions $\alpha_1, \ldots, \alpha_{A^2}$ that can be fixed to values $b_1, \ldots, b_{A^2}$, respectively, where $b_i \in \{+1, -1\}$ for $1 \leq i \leq A^2$, such that f restricted to the subspace $\{x \mid \chi_{\alpha_i}(x) = b_i\ ,\ \forall 1 \leq i \leq A^2\}$ is constant. This implies that for any non-zero coefficient $\hat{f}(\beta)$ there exists at least one other non-zero coefficient $\hat{f}(\beta+\gamma)$ for $\gamma \in \mathrm{span}\{\alpha_1, \ldots, \alpha_{A^2}\}$. Indeed, if no such coefficient exists then the restriction $f|_{\chi_{\alpha_1}(x)=b_1,\ldots,\chi_{\alpha_{A^2}}=b^2}$ will have the non-constant term $\hat{f}(\beta)\cdot\chi_\beta$ (for example, this can be easily obtained from Equation (5)). Therefore, for any other fixing of $\chi_{\alpha_1}, \ldots, \chi_{\alpha_{A^2}}$, both $\hat{f}(\beta)\chi_\beta$ and $\hat{f}(\beta+\gamma)\chi_{\beta+\gamma}$ collapse to the same (perhaps non-zero) linear function, which implies that $\mathrm{spar}(f|_{\chi_{\alpha_1}=b'_1,\ldots,\chi_{\alpha_{A^2}}=b'_{A^2}}) \leq \mathrm{spar}(f)/2$ for any choice of $b'_1, \ldots, b'_{A^2}$. In other words, if we consider the tree of depth A^2 in which on level i all nodes branch according to $\langle \alpha_i, x\rangle$ then restricting f to any path yields a new function with half the sparsity. Thus, we can continue this process by induction for at most $\log s$ steps, until all the functions in the leaves are constant. The resulting tree has depth at most $A^2 \log s$ as claimed. □

Our next goal is proving Theorem 1.4. To this end, we use a lemma which shows there exists a low depth functional $\oplus$-DT which computes a function g such that $\Pr_x[f(x) \neq g(x)] \leq \epsilon$, where x is drawn from the uniform distribution over $\mathbb{Z}_2^n$. Recall that the *bias* of a Boolean function f is defined to be

$$\mathrm{bias}(f) \stackrel{\mathrm{def}}{=} \left|\Pr_x[f(x)=1] - \Pr_x[f(x)=-1]\right|.$$

Alternatively, $\mathrm{bias}(f) = |\hat{f}(0)|$.

Lemma 3.6. *Let $f : \mathbb{Z}_2^n \to \{+1, -1\}$ be a Boolean function with $\|\hat{f}\|_1 \leq A$. Then, there exists a functional $\oplus$-DT of depth at most $O(A^2 + \log(1/\epsilon))$ that computes a function g such that $\Pr_x[f(x) \neq g(x)] \leq \epsilon$. Furthermore, the size of the tree is at most $2^{O(A^2)}\min\{1/\epsilon^2, O(\log(1/\epsilon))^{2A}\}$.*

Proof. Let $K = \max\left\{10A^2, 2\log(1/\epsilon)\right\}$ be a bound on the depth of the tree. In order to construct the functional decision tree, we use a recursive argument that stops whenever we reach a constant leaf, or after K levels of recursion, and then show that for a uniformly random $x \in \mathbb{Z}_2^n$, x arrives at a highly biased leaf with probability $\geq 1-\epsilon$, hence proving the statement of the lemma.

Let $\hat{f}(\alpha)$ be f's largest coefficient in absolute value, and $\hat{f}(\beta)$ the second largest. Note that if $|\hat{f}(0)| > 1-\epsilon$ we are done. Hence, we consider two cases:

1. $|\hat{f}(\alpha)| > 1-\epsilon$ for $\alpha \neq 0$:

   We first show that if $|\hat{f}(\alpha)| > 1-\epsilon$ then $|\hat{f}(0)| < \epsilon$. By considering $-f$ instead of f, if needed, we may assume without the loss of generality $\hat{f}(\alpha) > 1-\epsilon$. Note that

   $$1-\epsilon < \hat{f}(\alpha) = \Pr[f = \chi_\alpha] - \Pr[f \neq \chi_\alpha] = (1 - \Pr[f \neq \chi_\alpha]) - \Pr[f \neq \chi_\alpha],$$

so $\Pr[f \neq \chi_\alpha] < \epsilon/2$. Now, since $\mathbb{E}[\chi_\alpha] = 0$, we have

$$|\hat{f}(0)| = |\mathbb{E}[f]| = |\mathbb{E}[f] - \mathbb{E}[\chi_\alpha]| = |\mathbb{E}[f - \chi_\alpha]| \leq \mathbb{E}[|f - \chi_\alpha|] = 2\Pr[f \neq \chi_\alpha] < \epsilon.$$

In this case we query on χ_α. Note that no matter what value χ_α obtains, the restricted function has bias at least $|\hat{f}(\alpha)| - |\hat{f}(0)| > 1 - 2\epsilon$, and we terminate the recursion.

2. $|\hat{f}(\alpha)| \leq 1 - \epsilon$:

   In this case we query on $\chi_{\alpha+\beta}$. Let $f' = f|_{\chi_{\alpha+\beta}=1}$ and $f'' = f|_{\chi_{\alpha+\beta}=-1}$ By Lemma 3.1, for at least one of f' and f'', the spectral norm drops by at least $1/A$. We continue by induction the construction on f' and f'', terminating when all the leaves are highly biased (in particular this includes the case of a constant leaf), or after at most K levels of recursion.

It remains to be shown that the fraction of inputs $x \in \mathbb{Z}_2^n$ that arrive at an unbiased leaf is at most ϵ. We say an internal node labeled χ_γ is norm-reducing for x, if $\chi_\gamma(x) = b$ and the restriction on $\chi_\gamma = b$ reduces the spectral norm by at least $1/A$. Clearly, a computation over any input x which traverses A^2 norm reducing nodes for x arrives at a constant leaf. Furthermore, by construction, all the leaves which are not highly biased appear in the K-th level of the tree. Hence, an input which arrives at an unbiased node satisfies K independent linear equations, for which at most A^2 are norm reducing. Since for every fixed $0 \neq \gamma \in \mathbb{Z}_2^n$ and $b \in \{+1, -1\}$ the probability that $\chi_\gamma(x) = b$ is exactly $1/2$, the probability that x arrives at a non highly biased node is bounded by[4]

$$\sum_{i=0}^{A^2-1} \frac{\binom{K}{i}}{2^K} \leq 2^{-K} A^2 \binom{K}{A^2} \leq 2^{(-K/2)} \leq \epsilon$$

by the choice of K.

To prove the upper bound on the size of the tree we first note that 2^K is a trivial upper bound. Moreover, as in the proof of Theorem 1.2, the construction that we have satisfies the recursion formula

$$\begin{aligned} S(K-d, B) \leq \max\{&S(K-(d+1), B-1/2) + S(K-(d+1), B), \\ &S(K-(d+1), B-1/B) + S(K-(d+1), B-3/(4B))\}, \end{aligned}$$

where $S(K-d, B)$ stands for the number of leaves in the tree rooted at a node v at depth d such that the function f_v computed at v satisfies $\|\hat{f_v}\|_1 \leq B$. As before, the solution to this recursion is $S(K, A) \leq 2^{A^2} K^{2A}$. Overall, we have that the size of the tree the approximating parity decision is at most:

$$\begin{aligned} \min\left\{2^K, 2^{A^2} K^{2A}\right\} &= \min\left\{\max\left\{2^{10A^2}, \epsilon^{-2}\right\}, 2^{A^2} \cdot \max\left\{(2\log(1/\epsilon))^{2A}, (10A^2)^{2A}\right\}\right\} \\ &\leq \min\left\{2^{10A^2} \cdot \epsilon^{-2}, 2^{A^2} \cdot (10A^2)^{2A} \cdot (2\log(1/\epsilon))^{2A}\right\} \\ &\leq 2^{O(A^2)} \cdot \min\left\{\epsilon^{-2}, O(\log(1/\epsilon))^{2A}\right\} \end{aligned}$$

as claimed.

□

[4] We count how many words in $\{0,1\}^K$ with fewer than A^2 1's are there.

Note that if we replace each highly biased function-labeled leaf in the functional $\oplus$-DT from Lemma 3.6 with the constant it is biased towards (i.e. by the sign of its constant term), the total error would increase by at most ϵ. That is, it can be easily converted to a regular $\oplus$-DT of a function g which ϵ-approximates f. In fact, in the proof of Lemma 3.6, we could have continued the recursion until reaching a constant leaf or depth K, but for the sake of understanding the proof of Theorem 1.4 it may be more clear to keep the current version in mind.

The proof of Theorem 1.4 follows by combining Lemma 3.6 with the well known result of Goldreich and Levin [GL89] and of Kushilevitz and Mansour [KM93], who showed that given a query oracle to a function f, with high probability, one can approximate its large Fourier coefficients in polynomial time.

Lemma 3.7 ([GL89, KM93])**.** *There exists a randomized algorithm, such that given a query oracle to a function $f : \mathbb{Z}_2^n \to \{+1, -1\}$, and parameters δ, θ, η, outputs, with probability at least $1 - \delta$, a list containing all of f's Fourier coefficients whose absolute value is at least θ. Furthermore, the algorithm outputs an additive approximation of at most η to each of these coefficients. The algorithm runs in polynomial time in n, $1/\theta$, $1/\eta$ and* $\log(1/\delta)$.

Proof of Theorem 1.4. We use the algorithm from Lemma 3.7 to find f's largest Fourier coefficient in absolute value, $\hat{f}(\alpha)$. Whenever $|\hat{f}(\alpha)| \leq 1 - \epsilon$, Lemma 3.3 implies $|\hat{f}(\beta)| > \frac{1-\hat{f}(\alpha)^2}{\|\hat{f}\|_1} \geq \frac{1-\hat{f}(\alpha)^2}{A} > \epsilon/A$, so the same algorithm can be used to find the second largest coefficient, $\hat{f}(\beta)$, in time $\mathsf{poly}(n, A, 1/\epsilon, \log(1/\delta))$. We use Lemma 3.6 to construct a functional $\oplus$-DT, and replace every function-labeled leaf with the constant it's biased towards. The bound on the running time follows from the size of the $\oplus$-DT and the running time of the algorithm from Lemma 3.7.

In fact, there is a slight inaccuracy in the argument above. Note that Lemma 3.7 only guarantees that we find a coefficient that is approximately the largest one. However, if it is the case that the second largest coefficient is very close to the largest one, then in Lemma 3.6 when we branch according to $\chi_{\alpha+\beta}$ both children have significantly smaller spectral norm.

If it is the case that we correctly identified the largest Fourier coefficient but failed to identify the second largest then we note that if our approximation is good enough, say better than $\epsilon/2A$, then even if we are mistaken and branch according to $\chi_{\alpha+\beta'}$ where $\left| |\hat{f}(\beta)| - |\hat{f}(\beta')| \right| < \epsilon/2A$, the the argument in Lemma 3.6 still works, perhaps with a slightly worse constant in the big O. □

4 Functions over $\mathbb{Z}_p^n$ with small spectral norm

In this section, we extend our results to functions $f : \mathbb{Z}_p^n \to \{+1, -1\}$ where p is any fixed prime. Throughout this section we assume $p > 2$. We start by giving some basic facts on the Fourier transform over $\mathbb{Z}_p^n$.

4.1 Preliminaries

Let $\omega = e^{\frac{2\pi i}{p}} \in \mathbb{C}$ be a primitive root of unity of order p. The set of p^n group characters

$$\{\chi_\alpha : \mathbb{Z}_p^n \to \mathbb{C} \mid \alpha \in \mathbb{Z}_p^n\}$$

where $\chi_\alpha(x) = \omega^{\langle \alpha, x \rangle}$, is a basis for the vector space of functions from $\mathbb{Z}_p^n$ into $\mathbb{C}$, and is orthonormal with respect to the inner product $\langle f, g \rangle = \mathbb{E}_x[f(x)\overline{g(x)}]$.[5] We now have that $\hat{f}(\alpha) = \mathbb{E}_x[f(x)\overline{\chi_\alpha(x)}]$

[5] For a complex number z, we denote by $\overline{z}$ its complex conjugate.

and $f = \sum_{\alpha \in \mathbb{Z}_p^n} \hat{f}(\alpha)\chi_\alpha$. Plancherel's theorem holds here as well and the sparsity and L_1 norm are defined in the same way as they were defined for functions $f : \mathbb{Z}_2^n \to \{+1, -1\}$. Lemma 3.3 also extends to functions $f : \mathbb{Z}_p^n \to \{+1, -1\}$, with virtually the same proof. When f is real-valued (and in particular, a Boolean function), then $\hat{f}(0) = \mathbb{E}[f]$ is real, and it can also be directly verified that $\hat{f}(\alpha) = \overline{\hat{f}(-\alpha)}$.

We have the analog to Equation (5):

$$f(x) = \sum_{\beta \in \mathbb{Z}_p^n / \langle \alpha \rangle} \left(\sum_{k=0}^{p-1} \hat{f}(\beta + k \cdot \alpha)(\chi_\alpha(x))^k \right) \chi_\beta(x) \ . \tag{10}$$

Hence, when f is restricted to an affine subspace on which $\chi_\alpha = \omega^\lambda$ (where $0 \le \lambda \le p-1$), then for every[6] $\beta \in \mathbb{Z}_p^n / \langle \alpha \rangle$ we have

$$\hat{g}(\beta) = \sum_{k=0}^{p-1} \omega^{\lambda k} \hat{f}(\beta + k\alpha).$$

For every $\beta \in \mathbb{Z}_p^n$, we denote by $[\beta]_\alpha = \beta + \langle \alpha \rangle$ the coset of $\langle \alpha \rangle$ in which β resides.

Lemma 3.2 now becomes:

$$\sum_{\alpha \in \mathbb{Z}_p^n} \hat{f}(\alpha)\hat{f}(\beta - \alpha) = 0 \tag{11}$$

for all $0 \neq \beta \in \mathbb{Z}_p^n$.

As a generalization of the $\oplus$-DT model, we define a p-ary linear decision tree, denoted $\oplus_p$-DT, to be a computation tree where every internal node v is labeled by a linear function $\gamma \in \mathbb{Z}_p^n$ and has p children. The edges between v and its children are labeled $0, 1, \ldots, p-1$, and on an input x, it computes $\langle \gamma, x \rangle \mod p$ and branches accordingly. We carry along from the binary case the notation $D^{\oplus_p}(f)$ and $\mathrm{size}_{\oplus_p}(f)$, and define them to be the depth (respectively, size) of a minimal-depth (resp. size) $\oplus_p$-DT computing f.

4.2 Basic tools

In this section we prove the basic tools required for generalizing the theorems for functions defined on $\mathbb{Z}_2^n$ to functions $f : \mathbb{Z}_p^n \to \{+1, -1\}$. As a generalization of Lemma 3.1, we show a slightly more complex and detailed argument:

Lemma 4.1 (Main Lemma for functions over $\mathbb{Z}_p^n$)**.** *Let $f : \mathbb{Z}_p^n \to \{+1, -1\}$ be a non-constant Boolean function such that $\|\hat{f}\|_1 = A$. Let $\hat{f}(\alpha)$ be its largest coefficient in absolute value, and $\hat{f}(\beta)$ be the second largest. Then there exist a universal constant c_0 and a constant $c_1 = c_1(p) = O(1/p^2)$ such that*

1. *For all $\lambda \in \mathbb{Z}_p$, $\|\widehat{f|_{\chi_{\beta-\alpha}=\omega^\lambda}}\|_1 \le A - c_0|\hat{f}(\beta)|$.*
2. *There exists at least $m := \lfloor p/3 \rfloor$ distinct elements $\lambda_1, \ldots, \lambda_m \in \mathbb{Z}_p$ such that $\|\widehat{f|_{\chi_{\beta-\alpha}=\omega^{\lambda_k}}}\|_1 \le A - c_0|\hat{f}(\alpha)| \le A - c_0/A$ for $k = 1, \ldots, m$.*
3. *There exists at least $p-1$ distinct elements $\lambda_1, \ldots, \lambda_{p-1} \in \mathbb{Z}_p$ such that $\|\widehat{f|_{\chi_{\beta-\alpha}=\omega^{\lambda_k}}}\|_1 \le A - c_1 \cdot |\hat{f}(\alpha)| \le A - c_1/A$ for $k = 1, \ldots, p-1$.*

[6] Recall that $\langle \alpha \rangle$ is the additive group generated by α and $\mathbb{Z}_p^n / \langle \alpha \rangle$ is the set of cosets of $\langle \alpha \rangle$.

As before we first prove a claim characterizing functions with very small spectral norm. Observe that when $p > 2$, the characters themselves are not Boolean functions any more. The following is a variant of Lemma 3.5 for $\mathbb{Z}_p^n$ with $p > 2$.

Lemma 4.2. *Let $f : \mathbb{Z}_p^n \to \{+1, -1\}$ be a Boolean function such that $\|\hat{f}\|_1 = 1$. Then $f = \pm 1$.*

Proof. Once more, using Parseval's identity and the assumption:

$$\sum_{\gamma} |\hat{f}(\gamma)|^2 = 1 = \sum_{\gamma} |\hat{f}(\gamma)|.$$

As before, $|\hat{f}(\gamma)| \in [0, 1]$, which implies $|\hat{f}(\alpha)| = 1$ for exactly one $\alpha \in \mathbb{Z}_p^n$, i.e. $f = z \cdot \chi_\alpha$ where $z \in \mathbb{C}$ and $|z| = 1$. Since f is Boolean and $f(0) = z$, we get $z = \pm 1$, and $\pm\chi_\alpha$ is Boolean (when $p > 2$) only when $\alpha = 0$. □

The following is a purely geometric lemma we use in our analysis. Since the Fourier coefficients now are complex numbers we need to bound the decrease in the spectral norm when two coefficients that are not aligned in the same direction collapse to the same coefficient.

Lemma 4.3. *Let $z_1, z_2 \in \mathbb{C}$ such that $|z_1| = R$, $|z_2| = r$ and $r \leq R$. Suppose the angle between z_1 and z_2 is θ. Then, for $C = C(\theta) = (1 - \cos(\theta))/2$ it holds that*

$$|z_1| + |z_2| - |z_1 + z_2| \geq Cr.$$

We give the simple proof in Appendix B.

The next lemma is similar to the inequalities of the type we used in the proof of Lemma 3.1.

Lemma 4.4. *Let $f : \mathbb{Z}_p^n \to \{+1, -1\}$ be a non-constant Boolean function, and suppose $\hat{f}(0)$ is the largest Fourier coefficient in absolute value and $\hat{f}(\beta)$ is the second largest. Then*

$$2|\hat{f}(0)| \leq \sum_{\substack{\gamma \in \mathbb{Z}_p^n \\ \gamma \neq 0, \beta}} \min\left\{|\hat{f}(\gamma)|, |\hat{f}(\gamma - \beta)|\right\}.$$

Proof. By rearranging Equation (11) with respect to β, we get:

$$|2\hat{f}(0)\hat{f}(\beta)| = \left|\sum_{\substack{\gamma \in \mathbb{Z}_p^n \\ \gamma \neq 0, \beta}} \hat{f}(\gamma)\hat{f}(\beta - \gamma)\right|$$

Now apply the triangle inequality to the right hand side, and then utilize the fact that $\hat{f}(\beta)$ is the second largest in absolute value and $\hat{f}(0)$ does not appear in the right hand side, to obtain

$$2|\hat{f}(0)||\hat{f}(\beta)| \leq |\hat{f}(\beta)| \sum_{\substack{\gamma \in \mathbb{Z}_p^n \\ \gamma \neq 0, \beta}} \min\left\{|\hat{f}(\gamma)|, |\hat{f}(\beta - \gamma)|\right\}.$$

Since f is real-valued, $\hat{f}(\beta - \gamma) = \overline{\hat{f}(\gamma - \beta)}$ (and in particular, they have the same absolute value), and since f is non-constant, by Lemma 4.2 we have $\|\hat{f}\|_1 > 1$, i.e. $\hat{f}(\beta) \neq 0$, which implies the desired inequality. □

When analyzing the loss in the L_1 norm which is caused by restriction on χ_η, it will be convenient to sum over the individual losses on pairs $\hat{f}(\gamma), \hat{f}(\gamma+\eta)$ that collapse to the same coefficient. However, letting γ run over all of $\mathbb{Z}_p^n$, these pairs are not pairwise disjoint, so we might over-count the losses. The following lemma generously accounts for such over-counting issues, by showing that summing over all (not pairwise disjoint) pairs differs from the true counting by at most a constant factor.

Lemma 4.5. *Let $f : \mathbb{Z}_p^n \to \{+1, -1\}$, $0 \neq \eta \in \mathbb{Z}_p^n$, and $\lambda \in \mathbb{Z}_p$. If*

$$\sum_{\gamma \in \mathbb{Z}_p^n} |\hat{f}(\gamma)| + |\hat{f}(\eta+\gamma)| - |\hat{f}(\gamma) + \omega^\lambda \hat{f}(\eta+\gamma)| = B,$$

then

$$\sum_{\gamma \in \mathbb{Z}_p^n/\langle \eta \rangle} \left(\sum_{k=0}^{p-1} \left| \hat{f}(\gamma + k\eta) \right| - \left| \sum_{k=0}^{p-1} \hat{f}(\gamma+k\eta)\omega^{\lambda k} \right| \right) \geq B/3.$$

Note that the left hand side of the last inequality is exactly the loss in the L_1 norm when restricting f on $\chi_\eta = \omega^\lambda$. We defer the proof of Lemma 4.5 to Appendix C.

Lemma 4.6. *Let $f : \mathbb{Z}_p^n \to \{+1, -1\}$ be a non-constant Boolean function such that $\|\hat{f}\|_1 = A$. Let $\hat{f}(\alpha)$ be its largest coefficient in absolute value, and $\hat{f}(\beta)$ be the second largest. Suppose $\lambda \in \mathbb{Z}_p$ is such that the angle between $\hat{f}(\alpha)$ and $\omega^\lambda \hat{f}(\beta)$ in absolute value is at most $\pi/3$. Then there is a universal constant $c > 0$ such that $\|\widehat{f|_{\chi_{\beta-\alpha}=\omega^\lambda}}\|_1 \leq A - c|\hat{f}(\alpha)| \leq A - c/A$.*

Proof. Denote $\eta \overset{\text{def}}{=} \beta - \alpha$. Under the assumption of the lemma, noting that[7] $\hat{f}(\alpha) \cdot \overline{\omega^\lambda \hat{f}(\beta)} = \hat{f}(\alpha) \cdot \omega^{-\lambda} \hat{f}(-\beta)$

$$\operatorname{Re}\left(\hat{f}(\alpha) \cdot \omega^{-\lambda} \hat{f}(-\beta) \right) \geq \cos(\pi/3) \cdot |\hat{f}(\alpha) \cdot \omega^{-\lambda} \hat{f}(-\beta)| = \cos(\pi/3)|\hat{f}(\alpha)||\hat{f}(\beta)|. \tag{12}$$

By equation (11), with respect to $-\eta \neq 0$, we have

$$\sum_\gamma \hat{f}(\gamma)\hat{f}(-\eta-\gamma) = 0.$$

Hence

$$c_0 \cdot \hat{f}(\alpha)\hat{f}(-\beta) = -\sum_{\gamma \neq \alpha, -\beta} \hat{f}(\gamma)\hat{f}(-\eta-\gamma),$$

where $c_0 = 1$ if $\alpha = -\beta$ and $c_0 = 2$ otherwise. Multiplying by $\omega^{-\lambda}$ and taking the real part of both sides gives

$$\operatorname{Re}\left(c_0 \cdot \omega^{-\lambda} \hat{f}(\alpha)\hat{f}(-\beta) \right) = \sum_{\gamma \neq \alpha, -\beta} -\operatorname{Re}\left(\hat{f}(\gamma)\omega^{-\lambda}\hat{f}(-\eta-\gamma) \right). \tag{13}$$

Let $N_\eta = \left\{ \gamma \mid \operatorname{Re}\left(\hat{f}(\gamma)\omega^{-\lambda}\hat{f}(-\eta-\gamma) \right) < 0, \gamma \neq \alpha, -\beta \right\}$. Then (13), as well as the fact that $c_0 \in \{1, 2\}$ and the left hand side is positive (by (12)), imply

$$\operatorname{Re}\left(\omega^{-\lambda} \hat{f}(\alpha)\hat{f}(-\beta) \right) \leq \sum_{\gamma \in N_\eta} -\operatorname{Re}\left(\hat{f}(\gamma)\omega^{-\lambda}\hat{f}(-\eta-\gamma) \right). \tag{14}$$

[7] $\operatorname{Re}(z)$ is the real part of a complex number z and $\overline{z}$ is its conjugate.

Note that for every pair $(\hat{f}(\gamma),\ \omega^{-\lambda}\hat{f}(-\eta-\gamma))$, where $\gamma \in N_\eta$, the angle between $\hat{f}(\gamma)$ and $\omega^{-\lambda}\hat{f}(-\eta-\gamma)$ $(= \omega^{\lambda}\hat{f}(\eta+\gamma))$ is in the range $[\pi/2, 3\pi/2]$. Furthermore, when applying the restriction $\chi_\eta = \omega^\lambda$ each such pair collapses to the same coefficient, and since the angle between the two coefficients (in absolute value) is at least $\pi/2$, by Lemma 4.3 it follows that for all $\gamma \in N_\eta$,

$$|\hat{f}(\gamma)| + |\hat{f}(\eta+\gamma)| - |\hat{f}(\gamma) + \omega^\lambda \hat{f}(\eta+\gamma)| \geq \min\{|\hat{f}(\gamma)|, |\hat{f}(\eta+\gamma)|\} \cdot C(\pi/2). \tag{15}$$

(where $C(\theta) = (1-\cos(\theta))/2$ is as defined in Lemma 4.3). Using (15) to bound the loss on every coefficient, and summing over all $\gamma \in \mathbb{Z}_p^n$ (while bearing in mind that every summand is non-negative), we have

$$\begin{aligned}
&\sum_\gamma |\hat{f}(\gamma)| + |\hat{f}(\eta+\gamma)| - |\hat{f}(\gamma) + \omega^\lambda \hat{f}(\eta+\gamma)| \\
&\quad \geq \sum_{\gamma \in N_\eta} |\hat{f}(\gamma)| + |\hat{f}(\eta+\gamma)| - |\hat{f}(\gamma) + \omega^\lambda \hat{f}(\eta+\gamma)| \\
&\quad \geq \sum_{\gamma \in N_\eta} \min\{|\hat{f}(\gamma)|, |\hat{f}(\eta+\gamma)|\} \cdot C(\pi/2)
\end{aligned}$$

Since neither α nor $-\beta$ appear in N_η, and $\hat{f}(\beta)$ is the second largest coefficient, for all $\gamma \in N_\eta$ it holds that

$$\min\{|\hat{f}(\gamma)|, |\hat{f}(\eta+\gamma)|\} \geq \frac{|\hat{f}(\gamma)||\hat{f}(\eta+\gamma)|}{|\hat{f}(\beta)|},$$

so

$$\sum_\gamma |\hat{f}(\gamma)| + |\hat{f}(\eta+\gamma)| - |\hat{f}(\gamma) + \omega^\lambda \hat{f}(\eta+\gamma)| \geq \frac{1}{|\hat{f}(\beta)|} \cdot C(\pi/2) \cdot \sum_{\gamma \in N_\eta} |\hat{f}(\gamma)||\hat{f}(\eta+\gamma)|. \tag{16}$$

Taking the complex conjugate and multiplying by $\omega^{-\lambda}$, it is also clear that for all γ

$$|\hat{f}(\gamma)||\hat{f}(\eta+\gamma)| = |\hat{f}(\gamma)||\omega^{-\lambda}\hat{f}(-\eta-\gamma)| = |\hat{f}(\gamma)\omega^{-\lambda}\hat{f}(-\eta-\gamma)| \geq -\operatorname{Re}\left(\hat{f}(\gamma)\omega^{-\lambda}\hat{f}(-\eta-\gamma)\right).$$

Hence (16) and (14) imply

$$\begin{aligned}
\sum_\gamma |\hat{f}(\gamma)| + |\hat{f}(\eta+\gamma)| - |\hat{f}(\gamma) + \omega^\lambda \hat{f}(\eta+\gamma)| &\geq \frac{C(\pi/2)}{|\hat{f}(\beta)|} \cdot \sum_{\gamma \in N_\eta} -\operatorname{Re}\left(\hat{f}(\gamma)\omega^{-\lambda}\hat{f}(-\eta-\gamma)\right) \\
&\geq \frac{C(\pi/2)}{|\hat{f}(\beta)|} \operatorname{Re}\left(\omega^{-\lambda}\hat{f}(\alpha)\hat{f}(-\beta)\right).
\end{aligned}$$

And (12) now gives

$$\sum_\gamma |\hat{f}(\gamma)| + |\hat{f}(\eta+\gamma)| - |\hat{f}(\gamma) + \omega^\lambda \hat{f}(\eta+\gamma)| \geq \frac{C(\pi/2) \cdot \cos(\pi/3) \cdot |\hat{f}(\alpha)| \cdot |\hat{f}(\beta)|}{|\hat{f}(\beta)|} \geq c|\hat{f}(\alpha)|$$

where c is an absolute constant. By Lemma 4.5 the L_1 norm of the restricted function has decreased by at least $c|\hat{f}(\alpha)|/3$. □

We are now ready to prove the main lemma for functions over $\mathbb{Z}_p^n$.

Proof of Lemma 4.1. Let $\eta = \beta - \alpha$ as in Lemma 4.6. Let $\lambda \in \mathbb{Z}_p$, and consider the restriction $\chi_\eta = \omega^\lambda$. Let θ be the angle between $\hat{f}(\alpha)$ and $\omega^\lambda \hat{f}(\alpha + \eta) = \omega^\lambda \hat{f}(\beta)$. If θ is larger in absolute value than $\pi/3$, then under the restriction the coefficients $\hat{f}(\alpha)$ and $\hat{f}(\beta)$ collapse into the same coefficient, resulting by Lemma 4.3 in a $C(\pi/3) \cdot |\hat{f}(\beta)|$ loss in the L_1 norm (where $C(\cdot)$ is as stated in Lemma 4.3). If $\theta \leq \pi/3$ then Lemma 4.6 implies a loss of $c_0|\hat{f}(\alpha)|$ (which is also at least $c_0|\hat{f}(\beta)|$) in the L_1 norm where c_0 is an absolute constant. This completes Item 1 in the proof.

Furthermore, since multiplication by ω rotates $\hat{f}(\beta)$ by $2\pi/p$, there exists at least $\lfloor p/3 \rfloor$ values for $\eta \in \mathbb{Z}_p$ such that $|\theta|$ would be at most $\pi/3$, which completes Item 2 in the proof.

Next, we prove Item 3. Let $C = C(\pi/p) = (1 - \cos(\pi/p))/2 = O(1/p^2)$ as in Lemma 4.3. We distinguish between two cases: The first case we consider is $\alpha \neq 0$. In this case, by the fact the f is real-valued $|\hat{f}(-\alpha)| = |\hat{f}(\alpha)|$. So $\beta = -\alpha$ and by Item 1, restricting on $\chi_\eta = \omega^\lambda$, for any $\lambda \in \mathbb{Z}_p$, yields

$$\|\widehat{f|_{\chi_\eta = \omega^\lambda}}\|_1 \leq \|\hat{f}\|_1 - c_0 \cdot |\hat{f}(\alpha)|, \tag{17}$$

which implies Item 3 for this case.

The second case is that the largest Fourier coefficient in absolute value is achieved on $\alpha = 0$. In this case $\beta \neq 0$, and $\eta = \beta$. By the assumption $\|\hat{f}\|_1 > 1$, we have $|\hat{f}(\beta)| > 0$. We define the *weight* of a pair $\{\gamma, \gamma - \beta\} \subseteq \mathbb{Z}_p^n$ to be $w(\gamma) = \min\left\{|\hat{f}(\gamma)|, |\hat{f}(\gamma - \beta)|\right\}$, and denote

$$W = \sum_{\gamma \neq 0, \beta} w(\gamma).$$

Thus By Lemma 4.4, we have

$$2|\hat{f}(0)| \leq W. \tag{18}$$

Note that when restricting f on $\chi_\beta = \omega^\lambda$, $\hat{f}(\gamma)$ and $\hat{f}(\gamma - \beta)$ collapse to the same coefficient. The new coefficient becomes $|\hat{f}(\gamma) + \omega^\lambda \hat{f}(\gamma + \beta) + \cdots + \omega^{\lambda(p-1)} \hat{f}(\gamma + (p-1)\beta)|$. We analyze only the loss in the L_1 norm obtained from the collapse of $\hat{f}(\gamma)$ and $\hat{f}(\gamma + (p-1)\beta) = \hat{f}(\gamma - \beta)$ to the same coefficient. Let θ be the angle between $\hat{f}(\gamma)$ and $\hat{f}(\gamma - \beta)$. Since multiplication by ω is equivalent to rotation by $2\pi/p$, as λ traverses over $0, 1, ..., p-1$, the angle between $\hat{f}(\gamma)$ and $\omega^{\lambda(p-1)} \hat{f}(\gamma - \beta)$ attains all possible values $\theta + 2\kappa\pi/p$ for $\kappa = 0, 1, ..., p-1$. Hence, there exists at most one choice of λ such that the angle between $\hat{f}(\gamma)$ and $\omega^{\lambda(p-1)} \hat{f}(\gamma - \beta)$ is less than π/p. We call $\lambda \in \mathbb{Z}_p$ *good* with respect to γ if the angle between $\hat{f}(\gamma)$ and $\omega^{\lambda(p-1)} \hat{f}(\gamma - \beta)$ is at least π/p. If we fix β, then for every pair there exist at least $p - 1$ good elements in $\mathbb{Z}_p$. Intuitively, each element λ which is good guarantees a loss of at least $C \cdot \min\{|\hat{f}(\gamma)|, |\hat{f}(\gamma - \beta)|\} = Cw(\gamma)$ in the spectral norm (the actual analysis, which will now follow, is a bit more delicate).

Consider now the matrix M whose rows are indexed by elements $\gamma \in \mathbb{Z}_p^n$ for all $\gamma \neq 0, \beta$, and whose columns are indexed by all elements $\lambda \in \mathbb{Z}_p$. We define:

$$M_{\gamma,\lambda} = \begin{cases} w(\gamma) & \text{if } \lambda \text{ is good with respect to } \gamma \\ 0 & \text{otherwise} \end{cases}.$$

Since for every γ there are at least $p - 1$ good elements, we have

$$\sum_{\gamma,\lambda} M_{\gamma,\lambda} \geq (p-1) \sum_{\gamma \neq 0, \beta} w(\gamma) = (p-1)W. \tag{19}$$

While for every fixed column λ_0,

$$\sum_{\gamma} M_{\gamma,\lambda_0} \leq W. \tag{20}$$

As there are p columns, (19) and (20) together imply that there is at most one column in which the total weight is less than $W/2$, i.e. for all $\lambda \in \mathbb{Z}_p$ but at most one, it holds that

$$\sum_{\gamma} M_{\gamma,\lambda} \geq W/2 \ . \tag{21}$$

Every element $\lambda \in \mathbb{Z}_p$ which satisfies (21) will be called *good.* We thus proved the existence of at least $p-1$ good elements λ.

We now fix a good element λ and consider the restriction $\chi_\beta = \omega^\lambda$. By Lemma 4.5, the loss of the spectral norm under this restriction is at least

$$1/3 \cdot \sum_{\gamma} |\hat{f}(\gamma)| + |\hat{f}(\gamma - \beta)| - |\hat{f}(\gamma) + \omega^{\lambda(p-1)}\hat{f}(\gamma - \beta)|,$$

which is, by Lemma 4.3 and the definition of $M_{\lambda,\gamma}$, at least

$$1/3 \cdot \sum_{\gamma:\lambda \text{ is good w.r.t. } \gamma} C \cdot \min\left\{|\hat{f}(\gamma)|, |\hat{f}(\gamma - \beta)|\right\} = 1/3 \cdot \sum_{\gamma} C \cdot M_{\gamma,\lambda} \geq C \cdot W/6 \geq |\hat{f}(0)| \cdot C/3,$$

where we used (21) and (18) for the penultimate and last inequalities, respectively. Letting $c_1 = C/3$ completes the proof of the lemma. □

4.3 Analogs of Theorems 1.1, 1.2, 1.3 and 1.4

Theorems 4.7, 4.8, 4.10 and 4.11 now follow as consequences of Lemma 4.1. Their proofs use the same arguments we used to deduce their $\mathbb{Z}_2^n$ counterparts from Lemma 3.1. We use the notation $O_p(\cdot)$ when the underlying constant depends on p, whereas when we use $O(\cdot)$, the underlying constant is some absolute constant.

Theorem 4.7. *Let $f : \mathbb{Z}_p^n \to \{+1, -1\}$ be a Boolean function with $\|\hat{f}\|_1 = A$. Then there exists an affine subspace $V \subseteq \mathbb{Z}_p^n$ of co-dimension at most $O(A^2)$ such that f is constant on V.*

Proof. Apply Lemma 4.1 iteratively on f. By assumption $p > 2$, so $\lfloor p/3 \rfloor \geq 1$, and then using Item 2 in the proof, after at most A^2/c_0 steps, we are left with a function g which is a restriction of f on an affine subspace defined by the restrictions so far, such that $\|\hat{g}\|_1 = 1$. By Lemma 4.2 $g = \pm 1$. □

Theorem 4.8. *Let $f : \mathbb{Z}_p^n \to \{+1, -1\}$ be a Boolean function with $\|\hat{f}\|_1 = A$. Then* $\text{size}_{\oplus_p}(f) \leq p^{O_p(A^2)} n^{O_p(A)}$.

Proof. By Lemma 4.1, there is a constant $0 < c \leq 1$ (where $c := \min\{c_0, c_1\}$ depends only on p), a linear function $\gamma \in \mathbb{Z}_p^n$ and $\lambda_1, ..., \lambda_{p-1} \in \mathbb{Z}_p$ such that $\|\widehat{f|_{\chi_\gamma = \omega^{\lambda_j}}}\|_1 \leq A - c|\hat{f}(\alpha)|$ for all $1 \leq j \leq p-1$. Furthermore, for the p-th direction λ_p, the same lemma shows that $\|\widehat{f|_{\chi_\gamma = \omega^{\lambda_p}}}\|_1 \leq A - c|\hat{f}(\beta)|$ where $\hat{f}(\beta)$ is the second largest coefficient.

As before, let

$$L(n, A) \overset{\text{def}}{=} \max_{\substack{f:\mathbb{Z}_p^n \to \{+1,-1\} \\ \|\hat{f}\|_1 \leq A}} \text{size}_{\oplus_p} f.$$

We show, by induction on n, that $L(n, A) \leq p^{2A^2/c} n^{2A/c}$. For $n = 1$ the result is trivial. Let $n > 1$ and further assume that $A > 1$ (if $A = 1$ then the claim follows from Lemma 4.2).

As in the proof of Theorem 1.2, we consider two cases:

1. $|\hat{f}(\alpha)| \geq 1/2$
2. $1/2 > |\hat{f}(\alpha)| \geq 1/A$ and $|\hat{f}(\beta)| > \frac{1-|\hat{f}(\alpha)|^2}{A} \geq \frac{3}{4A}$.

We again Consider the tree whose first query is the linear function χ_γ. In the first case, by the choice of γ we obtain the following recursion:

$$L(n, A) \leq (p-1)L(n-1, A-c/2) + L(n-1, A).$$

The induction hypothesis then implies (using the assumption that $A > 1$)

$$\begin{aligned}
L(n, A) &\leq (p-1)p^{2(A-c/2)^2/c}(n-1)^{2(A-c/2)/c} + p^{2A^2/c}(n-1)^{2A/c} \\
&\leq (p-1)p^{2(A^2/c)-1}(n-1)^{2A/c-1} + p^{2A^2/c}(n-1)^{2A/c} \\
&\leq p^{2A^2/c}(n-1)^{2A/c-1}\left(1+(n-1)\right) \\
&\leq p^{2A^2/c}n^{2A/c}.
\end{aligned}$$

While in the second case, we have the recurrence

$$L(n, A) \leq (p-1)L(n-1, A-c/A) + L(n-1, A-3c/(4A)) \leq p \cdot L(n-1, A-3c/(4A))$$

Again, the induction hypothesis implies (using the assumption that $A > 1$)

$$\begin{aligned}
L(n, A) &\leq p \cdot p^{2(A-3c/(4A))^2/c}(n-1)^{2(A-(3c/(4A)))/c} \\
&\leq n^{2A/c} \cdot p^{1+2A^2/c-3+18c/(16A^2)} \\
&\leq n^{2A/c} \cdot p^{2A^2/c}\ .
\end{aligned}$$

□

As an immediate corollary, we get:

Corollary 4.9. *Let $f : \mathbb{Z}_p^n \to \{+1, -1\}$ be a Boolean function with $\|\hat{f}\|_1 = A$. Then $f = \sum_{i=1}^{p^{O_p(A^2)}n^{O_p(A)}} \pm \mathbb{1}_{V_i}$, where each V_i is an affine subspace of $\mathbb{Z}_p^n$.*

Theorem 4.10. *Let $f : \mathbb{Z}_p^n \to \{+1, -1\}$ be such that $\|\hat{f}\|_1 = A$ and $|\{\alpha \mid \hat{f}(\alpha) \neq 0\}| = s$. Then f can be computed by a $\oplus_p$-DT of depth $O(A^2 \log s)$.*

Proof. By Theorem 4.7, there exist $K = O(A^2)$ linear functions $\alpha_1, \ldots, \alpha_K$ which can be fixed to values $\omega^{\lambda_1}, \ldots, \omega^{\lambda_K}$ where $\lambda_j \in \mathbb{Z}_p$ for $1 \leq j \leq K$, such that f restricted to the subspace $\{x \mid \chi_{\alpha_j}(x) = \omega^{\lambda_j}\ ,\ \forall 1 \leq j \leq K\}$ is constant. Once again, this implies that for any non-zero coefficient $\hat{f}(\beta)$ there exists at least one other non-zero coefficient $\hat{f}(\beta+\gamma)$ for $\gamma \in \mathrm{span}\{\alpha_1, \ldots, \alpha_K\}$, since if no such coefficient exists then the restriction $f|_{\chi_{\alpha_1}(x)=\omega^{\lambda_1}, \ldots, \chi_{\alpha_K}=\omega^{\lambda_k}}$ will have the non-constant term $\hat{f}(\beta) \cdot \chi_\beta$. Therefore, for any other fixing of $\chi_{\alpha_1}, \ldots, \chi_{\alpha_K}$, both $\hat{f}(\beta)\chi_\beta$ and $\hat{f}(\beta+\gamma)\chi_{\beta+\gamma}$ collapse to the same (perhaps non-zero) linear function, which implies that $\mathrm{spar}(f|_{\chi_{\alpha_1}=\omega^{\lambda'_1}, \ldots, \chi_{\alpha_K}=\omega^{\lambda'_K}}) \leq \mathrm{spar}(f)/2$ for any choice of $\lambda'_1, \ldots, \lambda'_K$. Thus, we can continue by induction until all the functions in the leaves are constant. □

Theorem 4.11. *Let $f : \mathbb{Z}_p^n \to \{+1, -1\}$ be such that $\|\hat{f}\|_1 = A$. Then for every $\delta, \epsilon > 0$ there is a randomized algorithm that given a query oracle to f outputs (with probability at least $1-\delta$) a $\oplus_p$-DT of depth $O(A^2 + \log(1/\epsilon))$ and size $\min\{p^{O(A^2+\log(1/\epsilon))}, p^{O_p(A^2)} \cdot O(A^2+\log(1/\epsilon))^{O_p(A)}\}$ that computes a Boolean function g_ϵ such that $dist(f, g_\epsilon) \leq \epsilon$. The algorithm runs in polynomial time in $n, \exp(A^2), 1/\epsilon$ and $\log(1/\delta)$.*

The proof of Theorem 4.11 follows the same outline as the proof of Theorem 1.4. A *functional $\oplus_p$-DT* is defined as a $\oplus_p$-DT where we allow every leaf to be labeled by a Boolean function on $\mathbb{Z}_p^n$, and the bias of a function $f : \mathbb{Z}_p^n \to \{+1, -1\}$ is defined as in the binary case. We again show there exists a low depth $\oplus - DT$ which computes a function g such that $\Pr_x[f(x) \neq g(x)] \leq \epsilon$ (where x is drawn from the uniform distribution over $\mathbb{Z}_p^n$.

Lemma 4.12. *Let $f : \mathbb{Z}_2^n \to \{+1, -1\}$ be a Boolean function with $\|\hat{f}\|_1 \leq A$. Then there exists a functional $\oplus$-DT of depth at most $O(A^2 + \log(1/\epsilon))$ and size $\min\{p^{O(A^2+\log(1/\epsilon))}, p^{O_p(A^2)} \cdot O(A^2 + \log(1/\epsilon))^{O_p(A)}\}$, computing a function g such that $\Pr_x[f(x) \neq g(x)] \leq \epsilon$.*

Proof. The proof is essentially the same as the proof of Lemma 3.6, taking $K = \frac{20}{c_0}(A^2 + \log(1/\epsilon))$, where c_0 is as in Lemma 4.1. Note that in this case, if $|\hat{f}(\alpha)| > 1 - \epsilon$, then since $|\hat{f}(\alpha)| = |\hat{f}(-\alpha)|$, by Parseval's identity, if $\epsilon < (1 - 1/\sqrt{2})$ then this can only happen if $\alpha = 0$, hence f is already highly biased. Furthermore, for a random $x \in \mathbb{Z}_p^n$ and fixed $\gamma \in \mathbb{Z}_p^n$, Lemma 4.1 implies a node labeled χ_γ is norm reducing (by an absolute constant c_0/A) for x with probability $\frac{\lfloor p/3 \rfloor}{p} \geq 1/5$, hence a similar argument to the one used in Lemma 3.6 shows that a random input x arrives at an unbiased leaf with probability at most ϵ.

The bound on the tree size, which is $\min\{p^K, 2^{O_p(A^2)} \cdot K^{O_p(A)}\}$, also follows in the same way as in the proof of Lemma 3.6, using a similar recursion formula whose solution is similar to the formula on Theorem 4.8. □

Finally, we note that although this result is not stated in [KM93], the algorithm from Lemma 3.7 can be modified in the straightforward way to work equally well for functions $f : \mathbb{Z}_p^n \to \{+1, -1\}$, with virtually the same proof.

Proof of Theorem 4.11. We use the algorithm from Lemma 3.7 to find f's largest Fourier coefficient in absolute value, $\hat{f}(\alpha)$. Whenever $|\hat{f}(\alpha)| \leq 1 - \epsilon$, the same algorithm can be used to find the second largest coefficient, $\hat{f}(\beta)$, in polynomial time (in n, $1/\epsilon$ and $\log(1/\delta)$). We use Lemma 4.12 to construct a functional $\oplus_p$-DT, and replace every function-labeled leaf with the constant it is biased towards.

We again mention, as in the proof of Theorem 1.4, that we do not need to calculate $\hat{f}(\alpha)$ and $\hat{f}(\beta)$ exactly, but only to within an error of, say, $\epsilon/(2pA)$, which can be guaranteed (with high probability) by the algorithm of Lemma 3.7. □

5 Conclusions and open problems

In this work we obtained structural results for Boolean functions over $\mathbb{Z}_p^n$, for prime p. Our results provide a more refined structure than the one given in the works of Green and Sanders [GS08a, GS08b]. For a certain range of parameters we also obtain improved results in the setting of the works [GS08a, GS08b].

We were also able to achieve new results in the field of computational learning theory by showing that such functions can be learned with $\oplus$-DTs as the class of hypotheses.

There are still many intriguing open problems related to the structure of Boolean functions with small spectral norm. Most of these are related to the tightness of our results (as well as to the tightness of the results of Green and Sanders [GS08a]).

We do not believe that the bound given in Equation (2) is tight. Perhaps it is even true that one could represent f as a sum of polynomially (in A) many characteristic functions of subspaces (note that this is not true for functions over general abelian groups. See [GS08b]). Similarly, we do not believe that the bounds we obtain in Theorems 1.2 and 4.8 are tight. It seems more reasonable to believe that the true bound should be $\mathsf{poly}(n, A)$.

Recall that [ZS10, MO09] conjectured that Boolean functions with sparse Fourier spectrum can be computed by a $\oplus$-DT of depth $\mathsf{poly}(\log \operatorname{spar} f)$. Theorems 1.3 and 4.10 give an affirmative answer only for the case that f also has a small spectral norm. Thus, the general case is still open.

Finally, Theorems 1.4 and 4.11 give shallow $\oplus_p$-DTs approximating functions with small spectral norm. These results too do not seem tight. In particular, it is interesting to understand whether something better can be obtained if we assume in addition that f can be computed exactly by a small $\oplus_p$-DT. Namely, can one output a shallow $\oplus_p$-DT approximating f over the uniform distribution using polynomially many membership queries (i.e. oracle calls) to f, assuming that f can be exactly computed by such a $\oplus_p$-DT (and has a small spectral norm).

References

[ABF$^+$08] M. Alekhnovich, M. Braverman, V. Feldman, A. R. Klivans, and T. Pitassi. The complexity of properly learning simple concept classes. *J. of Computer and System Sciences*, 74(1):16–34, 2008. 1, 2

[BCH$^+$96] M. Bellare, D. Coppersmith, J. Håstad, M. A. Kiwi, and M. Sudan. Linearity testing in characteristic two. *IEEE Transactions on Information Theory*, 42(6):1781–1795, 1996. 1

[BdW02] H. Buhrman and R. de Wolf. Complexity measures and decision tree complexity: a survey. *Theoretical Computer Science*, 288(1):21–43, 2002. 5

[BLR93] M. Blum, M. Luby, and R. Rubinfeld. Selftesting/correcting with applications to numerical problems. *J. of Computer and System Sciences*, 47(3):549–595, 1993. 1

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

[GL89] O. Goldreich and L. A. Levin. A hardcore predicate for all one-way functions. In *Proceedings of the 21st STOC*, pages 25–32, 1989. 13

[GOS$^+$11] P. Gopalan, R. O'Donnell, R. Servedio, A. Shpilka, and K. Wimmer. Testing fourier dimensionality and sparsity. *SIAM Journal on Computing*, 40(4):1075–1100, 2011. 1

[Gro97] V. Grolmusz. On the power of circuits with gates of low L_1 norms. *Theoretical computer science*, 188(1):117–128, 1997. 1

[GS08a] B. Green and T. Sanders. Boolean functions with small spectral norm. *GAFA*, 18:144–162, 2008. 1, 2, 3, 4, 21

[GS08b] B. Green and T. Sanders. A quantitative version of the idempotent theorem in harmonic analysis. *Annals. of Math.*, 168(3):1025–1054, 2008. 3, 4, 21, 22

[Hås01] J. Håstad. Some optimal inapproximability results. *J. ACM*, 48(4):798–859, 2001. 1

[Kal02] G. Kalai. A Fourier-theoretic perspective on the Condorcet paradox and Arrow's theorem. *Advances in Applied Mathematics*, 29(3):412–426, 2002. 1

[KKL88] J. Kahn, G. Kalai, and N. Linial. The influence of variables on Boolean functions. In *Proceedings of the 29th annual FOCS*, pages 68–80, 1988. 1

[KM93] E. Kushilevitz and Y. Mansour. Learning Decision Trees Using the Fourier Spectrum. *SIAM J. Comput.*, 22(6):1331–1348, 1993. 1, 5, 13, 21

[LMN93] N. Linial, Y. Mansour, and N. Nisan. Constant depth circuits, Fourier transform and learnability. *J. ACM*, 40(3):607–620, 1993. 1

[Man94] Y. Mansour. Learning Boolean functions via the Fourier transform. In *Theoretical advances in neural computation and learning*, pages 391–424, 1994. 1

[MO09] A. Montanaro and T. Osborne. On the communication complexity of XOR functions. *CoRR*, abs/0909.3392, 2009. 1, 2, 4, 22

[O'D12] R. O'Donnell. Analysis of Boolean Functions. `http://www.analysisofbooleanfunctions.org/`, 2012. 1, 7

[TWXZ13] H. Y. Tsang, C. H. Wong, N. Xie, and S. Zhang. Fourier sparsity, spectral norm, and the log-rank conjecture. *CoRR*, abs/1304.1245, 2013. 4, 10

[ZS10] Z. Zhang and Y. Shi. On the parity complexity measures of Boolean functions. *Theoretical Computer Science*, 411(26-28):2612 – 2618, 2010. 1, 2, 4, 22

A Proof of Lemma 2.5

The proof of Lemma 2.5 relies upon the following even simpler lemma.

Lemma A.1. *Let $V \subseteq \mathbb{Z}_2^n$ be an affine subspace of co-dimension k, and let $\mathbb{1}_V : \mathbb{Z}_2^n \to \{0,1\}$ be its characteristic function. Then* $\mathrm{spar}(\mathbb{1}_V) = 2^k$ *and* $\|\widehat{\mathbb{1}_V}\|_1 = 1$.

Proof. Denote $V = \alpha + U$ where U is a subspace of co-dimension k. There are k vectors $\gamma_1, \dots, \gamma_k \in \mathbb{Z}_2^n$ (a basis for $U^\perp$) and $b_1, \dots, b_k \in \{+1, -1\}$ such that $\mathbb{1}_V(x) = 1$ if and only if $\chi_{\gamma_i}(x) = b_i$ for all $1 \le i \le k$. Therefore

$$\mathbb{1}_V(x) = \prod_{i=1}^{k} \left(\frac{\chi_{\gamma_i}(x) + b_i}{2} \right).$$

Using the relation $\chi_\beta \chi_\gamma = \chi_{\beta+\gamma}$, and the fact that $\mathrm{span}\{\gamma_1, \dots, \gamma_k\} = U^\perp$, we get

$$\mathbb{1}_V(x) = \sum_{\gamma \in U^\perp} \pm 2^{-k} \chi_\gamma(x).$$

Since $|U^\perp| = 2^k$, both statements follow. □

Proof of Lemma 2.5. Let L be the set of leaves of T, and for every $\ell \in L$ let b_ℓ be its label, and $\mathbb{1}_\ell : \mathbb{Z}_2^n \to \{0,1\}$ be the characteristic function of the set of inputs x such that computation upon x arrives at the leaf ℓ. Since T computes f, we may represent f as:

$$f = \sum_{\ell \in L} b_\ell \mathbb{1}_\ell(x).$$

Now note that if ℓ's depth is t, then $\mathbb{1}_\ell$ is a characteristic function of an affine subspace of co-dimension t. The maximal depth of T is k, hence for every $\ell \in L$ we have, by Lemma A.1, $\text{spar}(\mathbb{1}_\ell) \le 2^k$ and $\|\widehat{\mathbb{1}_\ell}\|_1 = 1$. Finally, since $|L| = m$, we get

$$\text{spar}(f) \le \sum_{\ell \in L} \text{spar}(\mathbb{1}_\ell) \le m2^k,$$

and since $|b_\ell| = 1$, the triangle inequality implies

$$\|\hat{f}\|_1 \le \sum_{\ell \in L} \|\widehat{\mathbb{1}_\ell}\|_1 \le m.$$

□

B Proof of Lemma 4.3

Proof. Suppose without the loss of generality (by applying a suitable rotation and reflection if needed) that $z_1 = R$ is a positive real number, and that the angle is exactly $\theta \le \pi$ (i.e. $z_2 = re^{i\theta}$).

Note that $|z_1| + |z_2| = R + r$ and $z_1 + z_2 = (R + r\cos(\theta)) + ir\sin(\theta)$. Hence,

$$|z_1 + z_2| = \sqrt{(R + r\cos(\theta))^2 + (r\sin(\theta))^2} = \sqrt{R^2 + r^2 + 2Rr\cos(\theta)}\ .$$

It remains to be shown that

$$R + r - \sqrt{R^2 + r^2 + 2Rr\cos(\theta)} \ge \frac{1 - \cos(\theta)}{2} r.$$

This is equivalent to

$$\left(R + r - \frac{1 - \cos(\theta)}{2} r\right)^2 - R^2 - r^2 - 2Rr\cos(\theta) \ge 0.$$

Rearranging and factoring out $r \ge 0$, we get a linear function in r which is non-negative on both $r = 0$ and $r = R$, which implies the inequality holds for all $0 \le r \le R$.

□

C Proof of Lemma 4.5

Proof. It is enough to show that on every coset $\gamma \in \mathbb{Z}_p^n / \langle \eta \rangle$:

$$\begin{aligned} & 3 \cdot \left(\sum_{k=0}^{p-1} \left| \hat{f}(\gamma + k\eta) \right| - \left| \sum_{k=0}^{p-1} \hat{f}(\gamma + k\eta)\omega^{\lambda k} \right| \right) \\ & \ge \sum_{k=0}^{p-1} |\hat{f}(\gamma + k\eta)| + |\hat{f}(\gamma + (k+1)\eta)| - |\hat{f}(\gamma + k\eta) + \hat{f}(\gamma + (k+1)\eta)\omega^{\lambda}| \end{aligned} \tag{22}$$

Fix a coset γ. For $k = 0, \ldots, p-1$ denote by $z_k \overset{\text{def}}{=} \hat{f}(\gamma + k\eta) \cdot \omega^{\lambda k}$. Rewriting Equation (22) under this notation gives

$$3 \cdot \left(\sum_{k=0}^{p-1} |z_k \cdot \omega^{-\lambda k}| - \left| \sum_{k=0}^{p-1} z_k \right| \right) \ge \sum_{k=0}^{p-1} |z_k \cdot \omega^{-\lambda k}| + |z_{k+1} \cdot \omega^{-\lambda(k+1)}| - |z_k \cdot \omega^{-\lambda k} + z_{k+1} \cdot \omega^{-\lambda k}|\ .$$

Since multiplying by $\omega^{-\lambda k}$ does not change the norm, this is equivalent to

$$3 \cdot \left(\sum_{k=0}^{p-1} |z_k| - \left|\sum_{k=0}^{p-1} z_k\right|\right) \geq \sum_{k=0}^{p-1} |z_k| + |z_{k+1}| - |z_k + z_{k+1}| \; . \tag{23}$$

We break the right hand side of Equation (23) into 3 sums:

1. $\sum_{k \in \{0,2,\ldots,p-3\}} |z_k| + |z_{k+1}| - |z_k + z_{k+1}|$
2. $\sum_{k \in \{1,3,\ldots,p-2\}} |z_k| + |z_{k+1}| - |z_k + z_{k+1}|$
3. $\sum_{k \in \{p-1\}} |z_k| + |z_{k+1}| - |z_k + z_{k+1}|$

Each sum goes over a disjoint set of pairs $(k, k+1)$. Next, we show that each sum is at most $\sum_{k=0}^{p-1} |z_k| - \left|\sum_{k=0}^{p-1} z_k\right|$, completing the proof. We claim in general that if $A \subseteq \{0, \ldots, p-1\}$ is a subset such that $1 + A = \{1 + a \mod p : a \in A\}$ is disjoint of A, then

$$\sum_{k \in A} |z_k| + |z_{k+1}| - |z_k + z_{k+1}| \leq \sum_{k=0}^{p-1} |z_k| - \left|\sum_{k=0}^{p-1} z_k\right| \; .$$

Let $B := \{0, 1, \ldots, p-1\} \setminus (A \cup (1 + A))$, then $A \cup (1 + A) \cup B$ is a disjoint union of $\{0, \ldots, p-1\}$ and we have:

$$\begin{aligned}
&\sum_{k \in A} |z_k| + |z_{k+1}| - |z_k + z_{k+1}| \\
&\quad = \sum_{k \in A} |z_k| + |z_{k+1}| + \sum_{k \in B} |z_k| - \sum_{k \in B} |z_k| - \sum_{k \in A} |z_k + z_{k+1}| \\
&\quad = \left(\sum_{k \in A} |z_k| + \sum_{k \in 1+A} |z_k| + \sum_{k \in B} |z_k|\right) - \sum_{k \in B} |z_k| - \sum_{k \in A} |z_k + z_{k+1}| \\
&\quad \leq \sum_{k=0}^{p-1} |z_k| - \left|\sum_{k \in B} z_k + \sum_{k \in A} z_k + z_{k+1}\right| \\
&\quad = \sum_{k=0}^{p-1} |z_k| - \left|\sum_{k \in B} z_k + \sum_{k \in A} z_k + \sum_{k \in 1+A} z_k\right| \\
&\quad = \sum_{k=0}^{p-1} |z_k| - \left|\sum_{k=0}^{p-1} z_k\right| \; ,
\end{aligned}$$

where in the inequality we used the triangle inequality. □