2016 IEEE 57th Annual Symposium on Foundations of Computer Science

Structure of protocols for XOR functions

Hamed Hatami
School of Computer Science
McGill University
Montreal, Canada
hatami@cs.mcgill.ca

Kaave Hosseini
Computer Science and Engineering
University of California, San Diego
San Diego, USA
skhossei@ucsd.edu

Shachar Lovett
Computer Science and Engineering
University of California, San Diego
San Diego, USA
slovett@ucsd.edu

***Abstract*—Let f be a boolean function on n variables. Its associated XOR function is the two-party function F(x, y) = f(x xor y). We show that, up to polynomial factors, the deterministic communication complexity of F is equal to the parity decision tree complexity of f. This relies on a novel technique of entropy reduction for protocols, combined with existing techniques in Fourier analysis and additive combinatorics.**

***Keywords*-Communication complexity, Parity decision tree, XOR functions**

I. Introduction

Let $F : X \times Y \to \{0,1\}$ be a boolean function and suppose Alice and Bob receive $x \in X$ and $y \in Y$, respectively. A natural question capturing the essence of communication complexity is the following: How much communication between Alice and Bob is required to compute $F(x,y)$ in the worst case? One of the fundamental open problems in communication complexity, the log-rank conjecture, links this question to the rank of F as a real matrix.

Conjecture I.1. *(Log-rank conjecture [LS93]) Is it true that for every boolean function $F : X \times Y \to \{0,1\}$,*

$$D(F) \leq \mathrm{polylog}(rank(F))$$

where $D(\cdot)$ is the deterministic communication complexity.

Yet, after over 30 years of active research, we are far from settling this conjecture, directing attention towards solving the log-rank conjecture for special classes of boolean functions. A natural and important such class is the so called XOR functions.

Let $\mathbb{F}_2^n$ be the n-dimensional vector space over the field of two elements. For a given function $f : \mathbb{F}_2^n \to \{0,1\}$ define its XOR function as $f_\oplus(x,y) = f(x+y)$. This class of functions is sufficiently large to capture many interesting examples (e.g., equality and Hamming distance functions), but it is also especially attractive for it allows use of tools from discrete Fourier analysis. This is because the eigenvalues of $f_\oplus$ as a matrix are the same as the Fourier coefficients of f; therefore, the rank of $f_\oplus$ is equal to the *Fourier sparsity* of f, which is the number of non-zero Fourier coefficients of f. Moreover, if $A \times B \subseteq \mathbb{F}_2^n \times \mathbb{F}_2^n$ is a monochromatic rectangle in $f_\oplus$, then f is constant on all of $A + B$, where the sum-set $A+B$ is defined as $\{a+b : a \in A, b \in B\}$. This directly links communication complexity of XOR functions to the structure of sum-sets in additive combinatorics. We will discuss this relation in more details later.

Going back to the log-rank conjecture for XOR functions, an interesting approach to settle the conjecture is via another complexity measure, called the *parity decision tree* complexity (PDT in short), denoted $\mathrm{pdt}(\cdot)$. A parity decision tree for a boolean function f is an extension of the usual notion of decision trees. While in a regular decision tree, intermediate nodes query variables, in a parity decision tree they are allowed to query an arbitrary linear function of the inputs. A depth k parity decision tree for a boolean function f can be used to construct a $2k$-bit communication protocol for $f_\oplus(x,y)$. Indeed for every linear function L, since $L(x \oplus y) = L(x) \oplus L(y)$, Alice and Bob need to exchange only 2-bits to evaluate $L(x \oplus y)$. Hence they can simulate the PDT exchanging only $2k$-bits, and thus $D(f_\oplus) \leq 2 \cdot \mathrm{pdt}(f)$.

In the opposite direction, since Fourier characters are exponentials of linear functions and f has Fourier sparsity at most $2^{D(f_\oplus)}$, we have $\mathrm{pdt}(f) \leq 2^{D(f_\oplus)}$. Our main interest in this work is whether this direction can be made efficient. Namely, is is true that an efficient deterministic protocol for an XOR function implies a polynomial depth parity decision tree for the corresponding boolean function. Our main result is a polynomial relation between the two.

Theorem I.2 (Main theorem)**.** *For any $f : \mathbb{F}_2^n \to \{0,1\}$*

H. Hatami is supported by an NSERC grant.
K. Hosseini is supported by an NSF CAREER award 1350481.
S. Lovett is supported by an NSF CAREER award 1350481 and a Sloan fellowsip.

0272-5428/16 $31.00 © 2016 IEEE
DOI 10.1109/FOCS.2016.38

we have $\mathrm{pdt}(f) \leq O(D(f_\oplus)^6)$.

In fact, we prove a stronger statement than the one stated above. Assuming that there is a partition of $f_\oplus$ to K monochromatic rectangles, we show that $\mathrm{pdt}(f) \leq O(\log^6(K))$. The partition does not necessarily need to come from a communication protocol.

A. Proof overview

Fix $f : \mathbb{F}_2^n \to \{0,1\}$, where we assume that $f_\oplus$ has an efficient deterministic protocol. Our goal is to design a low depth PDT for f.

a) Reduction to monochromatic subspaces: Note that if f has a PDT of depth k, then in particular, the leaves of the PDT determine affine subspaces of co-dimension $\leq k$ on which f is constant. We call such subspaces *monochromatic subspaces* for f. From here onwards, we use "subspace" as a shorthand for "affine subspace".

It turns out that in order to design a PDT for f, it suffices to show that there exists a large monochromatic subspace for f. This follows from [TWXZ13] who showed (among other things) that if f is constant on a subspace V, then the Fourier sparsity of f restricted to any coset of V reduces by at least a factor of two. This is sufficient for our application, as the existence of an efficient deterministic protocol for $f_\oplus$ implies in particular that f has low Fourier sparsity. This reduces Theorem I.2 to the following question, which is the main problem we investigate in this paper.

Question I.3. *Let $f : \mathbb{F}_2^n \to \{0,1\}$ with with $D(f_\oplus) \leq k$. Find a subspace V of co-dimension $\mathrm{poly}(k)$ on which f is constant.*

In the next few paragraphs we give a brief discussion of how to find such a subspace. We first describe a natural approach, which only tries to exploit the existence of a large monochromatic rectangle for $f_\oplus$ (many techniques in communication complexity follow this approach; in the randomized settings, one needs to replace "monochromatic rectangle" with "biased rectangle"). However, as we discuss below, a direct application of this technique fails, and a more careful application requires unproven conjectures in additive combinatorics. As such, we follow a different route, which exploits the entire structure of the protocol. This is less common in communication complexity, and we view this is as a conceptual contribution of this work.

b) Using a large monochromatic rectangle, and why it fails: The existence of an efficient deterministic protocol for $f_\oplus$ implies that it is constant on a large rectangle $A \times B$, and consequently f is constant on $A + B$. As a first attempt, one may hope that if $A, B \subseteq \mathbb{F}_2^n$ are large sets, then $A+B$ must contain a large subspace. This would directly imply that f is constant on this subspace. Unfortunately this is false, as the following example of Green [Gre04] shows.

Example I.4. *Let $A = B = \mathcal{B}(n/2-\sqrt{n})$ where $\mathcal{B}(r) \subset \{0,1\}^n$ is the hamming ball of radius r. Then $|A| = |B| = \Omega(2^n)$, $A + B = \mathcal{B}(n - 2\sqrt{n})$ but the largest subspace contained in $A + B$ has co-dimension $2\sqrt{n}$. For example, such a subspace can be obtained by fixing the first $2\sqrt{n}$ bits to zero.*

The situation improves for sum-sets involving more than two sets. Sanders [San12] showed that for a set $A \subset \mathbb{F}_2^n$ with $|A| \geq \varepsilon 2^n$, $4A = A + A + A + A$ contains a subspace of co-dimension $O(\log^4(1/\varepsilon))$. As Yao showed [Yao15], it follows directly from this result that a k-bit deterministic protocol for the 4-party function $F(x, y, z, w) = f(x \oplus y \oplus z \oplus w)$ implies a parity decision tree of depth $O(k^5)$ for f.

Going back to two-fold sum-sets, we note that despite Example I.4, for our application one might still be able to use other properties of f to find a large monochromatic subspace in $A+B$. For example, since f has low Fourier sparsity, if we find a subspace V on which f is nearly constant, then f will be in fact constant on this subspace. More precisely, since the Fourier sparsity of f is at most 2^k, its Fourier coefficients are all of the form $a/2^k$ for integers a (for a proof see [GOS+11]). In particular, $\mathbb{E}[f|_V] < 2^{-k}$ implies $f|_V \equiv 0$, and $\mathbb{E}[f|_V] > 1 - 2^{-k}$ implies $f|_V \equiv 1$. Therefore, given large sets $A, B \subseteq \mathbb{F}_2^n$, rather than showing the existence of a large subspace in $A + B$, it suffices to show that $A + B$ contains most of a large subspace, and then the Fourier sparsity of f implies that f is constant on this subspace. Working out the details, it turns out that we would need the following conjecture:

Conjecture I.5. *Let $A \subset \mathbb{F}_2^n$ be of size $|A| \geq \varepsilon 2^n$. Then for any $\delta > 0$ there exists a subspace V such that $|2A \cap V| \geq (1 - \delta)|V|$, where the co-dimension of V is at most $\mathrm{polylog}(1/\varepsilon\delta)$.*

For this and related conjectures see [SS14] (in particular Section 9, the paragraph on correlations of $2A, 3A, 4A$). We note that two partial results towards Conjecture I.5 are known, both due to Sanders:

- [San10] proves the existence of a subspace with co-dimension $O((1/\varepsilon)\log(1/\delta))$.
- [San12] proves the existence of a subspace with co-dimension $O((1/\delta^2)\log^4(1/\varepsilon))$.

Unfortunately, neither of these two bounds is strong

enough for our application. If $f_\oplus$ has a k-bit deterministic protocol, then the largest monochromatic rectangle satisfies $|A|, |B| \geq 2^{n-k}$. We thus have $\varepsilon = 2^{-k}$. Furthermore, $f_\oplus$ has at most 2^k nonzero Fourier coefficients, which means that we need a subspace which is 2^{-k} close to being monochromatic, and thus we need to set $\delta < 2^{-k}$. Hence to achieve our goal of finding a subspace of co-dimension $\mathrm{poly}(k)$, we need poly-logarithmic dependency on both ε and δ.

c) Our approach: utilizing the entire protocol: We circumvent the need to use unproven conjectures by devising an alternative route based on information theory, which exploits the entire structure of the protocol. Fix a deterministic protocol for $f_\oplus$ which sends k bits, and let $K = 2^k$. Let $A_i \times B_i$ for $i \in [K]$ be the partition of $\mathbb{F}_2^n \times \mathbb{F}_2^n$ induced by the protocol. For an input (x, y), let $\Pi_{xy} \in [K]$ denote the index of the unique rectangle that contains (x, y). By our assumption $f_\oplus$ is constant on each $A_i \times B_i$ (or equivalently the value of $f_\oplus(x, y)$ is determined by Π_{xy}), which means that f is constant on each $A_i + B_i$.

Let $\mu = \mathbb{E}[f]$ be the average of f on the entire space, and assume without loss of generality that $\mu \geq 1/2$. We may use the existence of a large monochromatic rectangle to find a large subspace V on which the average of f is far from the global average. Concretely, let $A \times B$ be the largest rectangle on which f equals to zero. It can be shown that $|A|, |B| \geq 2^{n-2k}$. The result of [San12] implies the existence of a subspace V such that $|V \cap (A + B)| \geq (3/4)|V|$, where the co-dimension of V is $O(k^4)$. This implies that $\mathbb{E}[f|_V] \leq 1/4$. For $x \in \mathbb{F}_2^n$, let $\tilde{x}$ be the unique element in $\mathbb{F}_2^n / V$ satisfying $x \in V + \tilde{x}$. Note that $x + y \in V$ if and only if $\tilde{x} = \tilde{y}$. Hence for (X, Y) uniformly sampled from $\mathbb{F}_2^n \times \mathbb{F}_2^n$, we have

$$\mathbb{E}[f(X+Y)] - \mathbb{E}[f(X+Y)|\tilde{X} = \tilde{Y}] \geq \frac{1}{2} - \frac{1}{4} = \frac{1}{4}. \quad (1)$$

This shows that Π_{XY} is not independent from $\tilde{X}\tilde{Y}$. However, we need to quantify this, and to this end, in Lemma III.4 we show that (1) implies that the mutual information between Π_{XY} and $\tilde{X}\tilde{Y}$ is large:

$$I(\Pi_{XY}; \tilde{X}\tilde{Y}) = H(\Pi_{XY}) - H(\Pi_{XY}|\tilde{X}\tilde{Y}) \geq 2^{-8}.$$

In other words, knowing which shifts of V, X and Y belong to, decreases the entropy of Π_{XY} significantly on average. In particular, there exists a coset $(V + w_1) \times (V + w_2)$ on which the entropy decreases by at least 2^{-8}. We may now iterate this process. As originally we have $H(\Pi_{XY}) \leq k$ (since the partition $\mathcal{P}$ is to $K = 2^k$ rectangles), after $O(k)$ iterations we will reach a constant function on a subspace of co-dimension $O(k^5)$.

Note that $\tilde{X} = \tilde{Y}$ can be a very small probability event (this is the case when V is a small subspace), and thus in the first glance it might be surprising that it is possible to use (1) to obtain an absolute lower bound for $I(\Pi_{XY}; \tilde{X}\tilde{Y})$, independent of the size of V. Indeed Lemma III.4 exploits the assumption that Π_{XY} is defined by a partition into combinatorial rectangles and as the following example shows this is not true for partitions into generic sets.

Example I.6. *Let $V = \{x \in \mathbb{F}_2^n : x_1 = 0\}$ so that $\tilde{x} = (x_2, \ldots, x_n)$ for $x = (x_1, \ldots, x_n) \in \mathbb{F}_2^n$. Consider the following partition of $\mathbb{F}_2^n \times \mathbb{F}_2^n$ into three sets*

$$\Pi_{xy} = \begin{cases} 1 & \tilde{x} = \tilde{y} \\ 2 & \tilde{x} \neq \tilde{y}, x_1 = 0 \\ 3 & \tilde{x} \neq \tilde{y}, x_1 = 1 \end{cases},$$

and let $f(x, y) = 1$ if $\Pi_{xy} = 2$ and $f(x, y) = 0$ otherwise. Then $\mathbb{E}[f] \approx \frac{1}{2}$ while $\mathbb{E}[f|\tilde{X} = \tilde{Y}] = 0$. However $I(\Pi_{XY}; \tilde{X}\tilde{Y}) = o(1)$. Similarly it is easy to construct examples showing that it is essential that X and Y are independent.

Paper organization: We give some preliminary definitions in Section II. We establish the key steps required in the proof of our main result in Sections III-A, III-B, III-C, and we apply them in Section III-D to prove our main result, Theorem I.2. We discuss some open problems in Section IV.

II. Preliminaries

Combinatorial rectangles and Partitions: The Cartesian product of two sets $A, B \subseteq \mathbb{F}_2^n$ is called a combinatorial rectangle. It is well-known that the inputs that lead to a particular leaf in a deterministic communication protocol form a combinatorial rectangle, and thus every such protocol provides a partition of $\mathbb{F}_2^n \times \mathbb{F}_2^n$ into combinatorial rectangles.

We will use functions $\Pi : \mathbb{F}_2^n \times \mathbb{F}_2^n \to [K]$ to denote partitions of $\mathbb{F}_2^n \times \mathbb{F}_2^n$. Here Π maps every input to the index of the unique rectangle that contains it. For every vector space V over $\mathbb{F}_2$ we extend these definitions to $V \times V$ by identifying $V \cong \mathbb{F}_2^n$ for $n = \dim(V)$.

Entropy, Mutual Information, and Divergence: The *entropy* of a discrete random variable X is defined as

$$H(X) = \sum_{a \in \mathrm{supp}(X)} \Pr[X = a] \log \frac{1}{\Pr[X = a]},$$

where here and throughout the paper, logarithms are in base two. The entropy of X conditioned on a random

variable Y is defined as

$$H(X|Y) = \sum_y \Pr[Y=y] H(X|Y=y)$$
$$= H(XY) - H(Y),$$

and corresponds to the amount of information that is left in X after knowing Y. Here and throughout the paper, as it is customary in information theory, we use XY to denote (X,Y).

The *mutual information* between X and Y is defined as

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$
$$= H(XY) - H(X) - H(Y).$$

Mutual information is symmetric, it is always non-negative, and it measures the amount of the information shared between two random variables. Let μ and ν be two probability distributions on the same space. The *Kullback-Leibler divergence* (or KL-divergence, or simply divergence) of ν from μ is defined as

$$D(\mu\|\nu) = \mathbb{E}_{a\sim\mu}\left[\log\frac{\mu(a)}{\nu(a)}\right].$$

The divergence $D(\mu\|\nu)$ is non-negative, and it is not symmetric in μ and ν. It is equal to $+\infty$ if $\mathrm{supp}(\mu) \not\subseteq \mathrm{supp}(\nu)$. The so called *Pinsker's inequality* states that divergence can be used to bound the distance between the two probability measures:

$$\sum_a |\mu(a) - \nu(a)| \le \sqrt{2D(\mu\|\nu)}. \tag{2}$$

Mutual information can be expressed using divergence. Indeed if $p(x,y)$ denotes the joint distribution of (X,Y), then

$$I(X;Y) = D(p(x,y)\|p_1(x)p_2(y)), \tag{3}$$

where $p_1(x)$ is the marginal distribution of X and $p_2(y)$ is the marginal distribution of Y.

III. Main theorem

As we have discussed in the introduction, the proof of Theorem I.2 can be divided into the following three steps:

- *Step I:* Applying Sanders's result [San12] together with Fourier sparsity of f to find a large subspace V such that
$$|\mathbb{E}[f] - \mathbb{E}[f|V]| \ge \frac{1}{4}.$$
- *Step II:* Applying information theoretic techniques to deduce from Step I that there exist $w', w'' \in \mathbb{F}_2^n$ with
$$H(\Pi|XY \in (V+w')\times(V+w'')) \le H(\Pi) - 2^{-8}.$$
Repeated application of Steps I and II will show the existence of a large subspace V such that $f|_V$ is constant; this will answer Question I.3.
- *Step III:* Using Fourier sparsity of f to deduce from Step II that f can be computed by a parity decision tree of low depth.

Next we will show how these three steps can be carried out.

A. Step I: A large subspace on which the average changes significantly

We use the following result of Sanders [San12] (see also [CS10] and [CŁS13]).

Theorem III.1. *Let $A, B \subseteq \mathbb{F}_2^n$ be sets of size $|A|, |B| \ge 2^n/K$. For any $\eta > 0$, there exists an affine subspace V of co-dimension $d \le O(\log^4(K)/\eta^2)$ such that*

$$|(A+B)\cap V| \ge (1-\eta)|V|.$$

We also need the following lemma from [GOS+11], which shows that Fourier sparse boolean functions cannot be too close to constant without actually being constant.

Lemma III.2 (Theorem 12 in [GOS+11])**.** *Let $f : \mathbb{F}_2^n \to \{0,1\}$ be a function which has at most 2^s nonzero Fourier coefficients. Then all the Fourier coefficients of f are of the form $\frac{a}{2^s}$ where $a \in \mathbb{Z}$. In particular, if $\mathbb{E}[f] < 2^{-s}$ then $f \equiv 0$, and if $\mathbb{E}[f] > 1 - 2^{-s}$ then $f \equiv 1$.*

The following corollary establishes Step I of the proof.

Corollary III.3. *Let $\Pi : \mathbb{F}_2^n \times \mathbb{F}_2^n \to [2^k]$ be a partition into $f_\oplus$-monochromatic rectangles. There exists a subspace $V \subseteq \mathbb{F}_2^n$ of co-dimension $O(k^4)$ such that*

$$|\mathbb{E}[f] - \mathbb{E}[f|V]| \ge \frac{1}{4}.$$

Proof: Assume without loss of generality that $\mathbb{E}[f] \ge 1/2$ (otherwise replace f with $1-f$). By Lemma III.2, we have $\mathbb{E}[f_\oplus] = \mathbb{E}[f] \le 1 - 2^{-k}$. Considering all the 0-rectangles in the partition, there must exist a rectangle $A \times B$ in the partition such that $f(A+B) = 0$ and $|A\times B| \ge 2^{2n-2k}$. In particular, $|A|, |B| \ge 2^{n-2k}$. Applying Theorem III.1 to A, B with $K = 2^{2k}, \eta = 1/4$, we deduce the existence of an affine subspace V of co-dimension $O(k^4)$ such that $|(A+B)\cap V| \ge (3/4)|V|$. In particular, $\mathbb{E}[f|V] \le 1/4$. ∎

B. Step II: Decreasing the entropy of the partition

Consider $f : \mathbb{F}_2^n \to \{0,1\}$, a partition $\Pi : \mathbb{F}_2^n \times \mathbb{F}_2^n \to [K]$ into rectangles such that $f_\oplus$ is constant on each rectangle, and a subspace V of $\mathbb{F}_2^n$. For $x \in \mathbb{F}_2^n$, let $\tilde{x}$ be the unique element in $\mathbb{F}_2^n/V$ satisfying $x \in V + \tilde{x}$.

Lemma III.4. *If $|\mathbb{E}[f] - \mathbb{E}[f|V]| \geq \epsilon$ and (X,Y) takes values in $\mathbb{F}_2^n \times \mathbb{F}_2^n$ uniformly at random, then for $\Pi = \Pi(X,Y)$, we have*

$$I(\Pi; \tilde{X}\tilde{Y}) \geq \epsilon^2/16.$$

Proof: Denote $W = \mathbb{F}_2^n/V$, and for every $t \in [K]$ and $w \in W$, let $p_t = \Pr[\Pi = t]$ and $p_{t|w,w} = \Pr[\Pi = t|\tilde{X} = \tilde{Y} = w]$. It follows from the assumption

$$\begin{aligned}|\mathbb{E}[f] - \mathbb{E}[f|V]| \\ = |\mathbb{E}[f(X+Y)] - \mathbb{E}[f(X+Y)|\tilde{X} = \tilde{Y}]| \geq \epsilon\end{aligned}$$

that $\sum_t |p_t - \mathbb{E}_{w \in W}\left[p_{t|w,w}|\right] \geq \epsilon$. In particular

$$\sum_t \mathbb{E}_{w \in W}\left[|p_t - p_{t|w,w}|\right] \geq \epsilon.$$

Since Π is a partition into *rectangles*, for every $w \in W$ and $t \in [K]$, we have

$$\begin{aligned}p_{t|w,w} &= \frac{\Pr[\tilde{X} = \tilde{Y} = w|\Pi = t] \times p_t}{\Pr[\tilde{X} = \tilde{Y} = w]} \\ &= p_t \times \frac{\Pr[\tilde{X} = w|\Pi = t]}{\Pr[\tilde{X} = w]} \times \frac{\Pr[\tilde{Y} = w|\Pi = t]}{\Pr[\tilde{Y} = w]}.\end{aligned}$$

Consequently, using $\max(0, 1-ab) \leq |1-a| + |1-b|$ and Pinsker's inequality (2) we have

$$\begin{aligned}\epsilon &\leq \sum_t \mathbb{E}_{w \in W}\left[|p_t - p_{t|w,w}|\right] \\ &= 2\sum_t \mathbb{E}_{w \in W}\left[\max(0, p_t - p_{t|w,w})\right] \\ &= 2\mathbb{E}_{t \sim \Pi}\mathbb{E}_{w \in W}\left[\max\left(0, 1 - \frac{\Pr[\tilde{X} = w|\Pi = t]}{\Pr[\tilde{X} = w]} \times \right.\right. \\ &\left.\left. \frac{\Pr[\tilde{Y} = w|\Pi = t]}{\Pr[\tilde{Y} = w]}\right)\right] \\ &\leq 2\mathbb{E}_{t \sim \Pi}\mathbb{E}_{w \in W}\left[\left|1 - \frac{\Pr[\tilde{X} = w|\Pi = t]}{\Pr[\tilde{X} = w]}\right| + \right. \\ &\left. \left|1 - \frac{\Pr[\tilde{Y} = w|\Pi = t]}{\Pr[\tilde{Y} = w]}\right|\right] \\ &\leq 2\sqrt{2I(\Pi; \tilde{X})} + \sqrt{2I(\Pi; \tilde{Y})} \\ &\leq 4\sqrt{I(\Pi; \tilde{X}) + I(\Pi; \tilde{Y})}.\end{aligned}$$

where we used (3) to show that $I(\Pi; \tilde{X}) = D(p_{t,w} \parallel p_t q_w)$ with $p_{t,w} = \Pr[\Pi = t, \tilde{X} = w]$ and $q_w = \Pr[\tilde{X} = w]$, and the similar identity for $I(\Pi; \tilde{Y})$. Finally since $\tilde{X}$ and $\tilde{Y}$ are independent (even after conditioning on Π), we have $I(\Pi; \tilde{X}\tilde{Y}) = I(\Pi; \tilde{X}) + I(\Pi; \tilde{Y})$. ∎

Remark III.5. *Note that the proof of Lemma III.4 shows that the following general statement is true. Let μ and ν be two distributions on $\mathbb{F}_2^n$, and let A and B be two functions on $\mathbb{F}_2^n$ such that $A(X)$ and $B(Y)$ have the same distribution if $(X,Y) \sim \mu \times \nu$. If $\Pi : \mathbb{F}_2^n \times \mathbb{F}_2^n \to [K]$ is a partition into rectangles, and $g : \mathbb{F}_2^n \times \mathbb{F}_2^n \to \{0,1\}$ is constant on each rectangle, then*

$$\begin{aligned}&|\mathbb{E}[g(X,Y)] - \mathbb{E}[g(X,Y)|A(X) = B(Y)]| \\ &\leq 4\sqrt{I(\Pi(X,Y); A(X)B(Y))}.\end{aligned}$$

C. Step III: Constructing the PDT

Tsang et al. [TWXZ13] showed that in order to design a parity decision tree, it suffices to find a large subspace on which the function is constant; and then recurse. For completeness, we reproduce their argument. Let $\mathrm{rank}(f)$ denote the rank of the real matrix $M_{x,y} = f(x+y)$. It equals the number of nonzero Fourier coefficients of f. Note that $\log \mathrm{rank}(f) \leq \mathrm{D}^\oplus(f)$.

Lemma III.6. *Let $T : \mathbb{N} \to \mathbb{N}$ be a function for which the following holds. For any function $f : \mathbb{F}_2^n \to \{0,1\}$, if $\mathrm{D}^\oplus(f) = k$ then there exists an affine subspace V of co-dimension $T(k)$ on which f is constant. Then for any function $f : \mathbb{F}_2^n \to \{0,1\}$, $\mathrm{pdt}(f) \leq T(\mathrm{D}^\oplus(f)) \cdot (\mathrm{D}^\oplus(f) + 1)$.*

Proof: The main idea is that if f is constant on V, then its rank on any coset of V reduces by at least a factor of two, which then allows for induction. To see that, assume that $\mathrm{rank}(f) = r$. Then

$$f(x) = \sum_{i=1}^{r} \hat{f}(\alpha_i)(-1)^{\langle x, \alpha_i \rangle},$$

for some $\alpha_1, \ldots, \alpha_r \in \mathbb{F}_2^n$. We know by assumption that f is constant on an affine subspace V of co-dimension $t = T(\mathrm{D}^\oplus(f))$. We may assume that V is linear subspace, by replacing $f(x)$ with $f(x+v)$ for some $v \in V$ (note that this does not change $\mathrm{D}^\oplus$ or $\mathrm{rank}(f)$). Let W be the quotient subspace $\mathbb{F}_2^n/V$ so that $\dim(W) = t$ and $\mathbb{F}_2^n = V + W$. Note that any $x \in \mathbb{F}_2^n$ can be uniquely decomposed as $x = v + w$ with $v \in V, w \in W$. Let $\pi_V : \mathbb{F}_2^n \to V$ and $\pi_W : \mathbb{F}_2^n \to W$ be the projection maps to V and W, respectively, mapping

$x = v + w$ to $\pi_V(x) = v$ and $\pi_W(x) = w$. Then

$$f|_V(v) = \sum_{i=1}^{r} \hat{f}(\alpha_i)(-1)^{\langle v, \pi_V(\alpha_i)\rangle},$$

In particular, as f is constant on V, it must be the case that for every non-zero α_i there exists some α_j such that $\pi_V(\alpha_i) = \pi_V(\alpha_j)$, or equivalently $\alpha_i + \alpha_j \in W$. Thus

$$|\{\pi_V(\alpha_i) : i \in [r]\}| \leq \frac{r+1}{2}.$$

Let $V + w$ be any coset of V. Then

$$f|_{V+w}(v+w) = \sum_{i=1}^{r} \hat{f}(\alpha_i)(-1)^{\langle w, \pi_W(\alpha_i)\rangle}(-1)^{\langle v, \pi_V(\alpha_i)\rangle}.$$

In particular, $\mathrm{rank}(f|_{V+w}) \leq |\{\pi_V(\alpha_i) : i \in [r]\}| \leq \frac{\mathrm{rank}(f)+1}{2}$.

We now construct the parity decision tree for f. We first query $w = \pi_W(x)$, which requires depth $\dim(W) = T(\mathrm{D}^{\oplus}(f))$. Each restricted function $f|_{V+w}$ has $\mathrm{D}^{\oplus}(f|_{V+w}) \leq \mathrm{D}^{\oplus}(f)$ and $\mathrm{rank}(f|_{V+w}) \leq \frac{\mathrm{rank}(f)+1}{2}$, and hence by induction can be computed by a parity decision tree of depth at most $T(\mathrm{D}^{\oplus}(f)) \cdot (\log(\mathrm{rank}(f)) + 1) \leq T(\mathrm{D}^{\oplus}(f)) \cdot (\mathrm{D}^{\oplus}(f) + 1)$. The lemma follows. ∎

D. Proof of Theorem I.2

Let $f : \mathbb{F}_2^n \to \{0,1\}$ be a boolean function. The associated XOR function is $f_{\oplus}(x,y) = f(x+y)$. Let $\mathrm{D}^{\oplus}(f)$ denote the minimum complexity of a deterministic protocol which computes $f_{\oplus}$. We restate Theorem I.2, which we prove in this section, for the convenience of the reader.

Theorem I.2 (Main theorem). *For any $f : \mathbb{F}_2^n \to \{0,1\}$ we have* $\mathrm{pdt}(f) \leq O(\mathrm{D}^{\oplus}(f)^6)$.

Proof: Let $k = \mathrm{D}^{\oplus}(f)$. By Corollary III.3 there exists an affine subspace V of co-dimension $O(k^4)$ such that

$$|\mathbb{E}[f] - \mathbb{E}[f|V]| \geq \frac{1}{4}.$$

Let $W = \mathbb{F}_2^n/V$ so that $\mathbb{F}_2^n = V + W$. Applying Lemma III.4, we obtain

$$I(\Pi; \tilde{X}\tilde{Y}) \geq 2^{-8}.$$

In particular, there exists a choice of $w', w'' \in W$ such that

$$H(\Pi|\tilde{X} = w', \tilde{Y} = w'') \leq H(\Pi) - 2^{-8}.$$

Note that by restricting the rectangles of Π to $(V + w') \times (V + w'')$, we obtain a partition $\Pi|_{(V+w')\times(V+w'')}$ of $(V+w') \times (V+w'')$ into $f|_{V+w'+w''}$-monochromatic rectangles with

$$H(\Pi|_{(V+w')\times(V+w'')}) = H(\Pi|\tilde{X} = w', \tilde{Y} = w'') \leq H(\Pi) - 2^{-8}.$$

Since $H(\Pi) \leq k$, iterating this procedure at most $2^8 k$ times, we find an affine subspace V such that $f|_V$ is constant. Furthermore since each iteration increases the co-dimension by at most $O(k^4)$, the subspace V will have co-dimension $O(k^5)$. Finally, we can apply Lemma III.6 to conclude the theorem. ∎

IV. Open problems

There are two natural open problems which stem directly from our work. The first is whether our result can be extended to randomized protocols vs randomized parity decision trees. Some partial results follow directly from our technique (concretely, a parity decision tree which approximates the function under a product distribution) but the general result still seems to be elusive.

Problem IV.1. *Let $f : \mathbb{F}_2^n \to \{0,1\}$ be a function. Assume that $f_{\oplus}$ has a randomized protocol with complexity k. Does there exist a randomized parity decision tree of depth* $\mathrm{poly}(k)$ *which computes f?*

The second question asks about what happens if we replace XOR with other gadgets. Sherstov [She11] showed that for many gadgets, including some natural 2-bit gadgets, efficient protocols imply low-degree approximating polynomials, which by the work of Nisan and Szegedy [NS94] imply efficient (standard) decision trees. This however does not hold for 1-bit gadgets. Except for XOR functions, the other class of gadgets that can be considered are AND gadgets (any other 1-bit gadget is either trivial or equivalent to either XOR or AND).

That is, for a boolean function $f : \{0,1\}^n \to \{0,1\}$ define its corresponding AND function as $f_{\wedge}(x,y) = f(x \wedge y)$, where $\wedge$ is bitwise AND function. An example of an AND function is disjointness. The analog class of decision trees are AND decision trees, where each internal node may query the AND of a subset of the inputs or their negations.

Problem IV.2. *Let $f : \mathbb{F}_2^n \to \{0,1\}$ be a function. Assume that $f_{\wedge}$ has a deterministic / randomized protocol with complexity k. Does there exist a deterministic / randomized AND decision tree of depth* $\mathrm{poly}(k)$ *which computes f?*

Acknowledgements: We thank Pooya Hatami for early stage discussions about this project.

REFERENCES

[CŁS13] Ernie Croot, Izabella Łaba, and Olof Sisask. Arithmetic progressions in sumsets and L_p-almost-periodicity. *Combinatorics, Probability and Computing*, 22(03):351–365, 2013.

[CS10] Ernie Croot and Olof Sisask. A probabilistic technique for finding almost-periods of convolutions. *Geometric and functional analysis*, 20(6):1367–1396, 2010.

[GOS+11] Parikshit Gopalan, Ryan O'Donnell, Rocco A Servedio, Amir Shpilka, and Karl Wimmer. Testing fourier dimensionality and sparsity. *SIAM Journal on Computing*, 40(4):1075–1100, 2011.

[Gre04] Ben Green. Spectral structure of sets of integers. In *Fourier analysis and convexity*, pages 83–96. Springer, 2004.

[LS93] László Lovăsz and Michael Saks. Communication complexity and combinatorial lattice theory. *Journal of Computer and System Sciences*, 47(2):322–349, 1993.

[NS94] Noam Nisan and Mario Szegedy. On the degree of boolean functions as real polynomials. *Computational complexity*, 4(4):301–313, 1994.

[San10] Tom Sanders. Green's sumset problem at density one half. *arXiv preprint arXiv:1003.5649*, 2010.

[San12] Tom Sanders. On the Bogolyubov–Ruzsa lemma. *Analysis & PDE*, 5(3):627–655, 2012.

[She11] Alexander A Sherstov. The pattern matrix method. *SIAM Journal on Computing*, 40(6):1969–2000, 2011.

[SS14] Tomasz Schoen and Olof Sisask. Roth's theorem for four variables and additive structures in sums of sparse sets. *arXiv preprint arXiv:1408.2568*, 2014.

[TWXZ13] Hing Yin Tsang, Chung Hoi Wong, Ning Xie, and Shengyu Zhang. Fourier sparsity, spectral norm, and the log-rank conjecture. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 658–667. IEEE, 2013.

[Yao15] Penghui Yao. Parity decision tree complexity and 4-party communication complexity of xor-functions are polynomially equivalent. *arXiv preprint arXiv:1506.02936*, 2015.