Electronic Colloquium on Computational Complexity, Revision 2 of Report No. 52 (2013)

Upper Bounds on Fourier Entropy

Sourav Chakraborty* Raghav Kulkarni†
Satyanarayana V. Lokam‡ Nitin Saurabh§

Abstract

Given a function $f : \{0,1\}^n \to \{+1,-1\}$, its *Fourier Entropy* is defined to be $-\sum_S \widehat{f}(S)^2 \log \widehat{f}(S)^2$, where $\hat{f}$ denotes the Fourier transform of f. In the analysis of Boolean functions, an outstanding open question is a conjecture of Friedgut and Kalai (1996), called the Fourier Entropy Influence (FEI) Conjecture, asserting that the Fourier Entropy of any Boolean function f is bounded above, up to a constant factor, by the total influence (= average sensitivity) of f.

In this paper we give several upper bounds on the Fourier Entropy. We first give upper bounds on the Fourier Entropy of Boolean functions in terms of several complexity measures that are known to be bigger than the influence. These complexity measures include, among others, the logarithm of the number of leaves and the average depth of a parity decision tree. We then show that for the class of Linear Threshold Functions (LTF), the Fourier Entropy is $O(\sqrt{n})$. It is known that the average sensitivity for the class of LTF is $\Theta(\sqrt{n})$. We also establish a bound of $O_d(n^{1-\frac{1}{4d+6}})$ for general degree-d polynomial threshold functions. Our proof is based on a new upper bound on the *derivative of noise sensitivity*. Next we proceed to show that the FEI Conjecture holds for read-once formulas that use AND, OR, XOR, and NOT gates. The last result is independent of a result due to O'Donnell and Tan [1] for read-once formulas with *arbitrary* gates of bounded fan-in, but our proof is completely elementary and very different from theirs. Finally, we give a general bound involving the first and second moments of sensitivities of a function (average sensitivity being the first moment), which holds for real-valued functions as well.

*Chennai Mathematical Institute, Chennai, India. sourav@cmi.ac.in
†Centre for Quantum Technologies, Singapore. kulraghav@gmail.com
‡Microsoft Research India, Bangalore, India. satya@microsoft.com
§The Institute of Mathematical Sciences, Chennai, India. nitin@imsc.res.in

ISSN 1433-8092

1 Introduction

Fourier transforms are extensively used in a number of fields such as engineering, mathematics, and computer science. Within theoretical computer science, Fourier analysis of Boolean functions evolved into one of the most useful and versatile tools; see the book [2] for a comprehensive survey of this area. In particular, it plays an important role in several results in complexity theory, learning theory, social choice, inapproximability, metric spaces, etc. If $\hat{f}$ denotes the Fourier transform of a Boolean function f, then $\sum_{S\subseteq[n]} \widehat{f}(S)^2 = 1$ and hence we can define the (Shannon) entropy of the distribution given by $\widehat{f}(S)^2$:

$$\mathbb{H}(f) := \sum_{S\subseteq[n]} \widehat{f}(S)^2 \log \frac{1}{\widehat{f}(S)^2} \ . \tag{1}$$

The Fourier Entropy-Influence (FEI) Conjecture, made by Friedgut and Kalai [3] in 1996, states that for every Boolean function, its Fourier entropy is bounded above by its total influence :

Fourier Entropy-Influence Conjecture There exists a universal constant C such that for all $f : \{0,1\}^n \to \{+1,-1\}$,

$$\mathbb{H}(f) \leqslant C \cdot \mathsf{Inf}(f) \ , \tag{2}$$

where $\mathsf{Inf}(f)$ is the total influence of f which is the same as the average sensitivity $\mathsf{as}(f)$ of f.

The notion of influence was studied by Ben-or and Linial [4] in the context of sharing an unbiased common random bit in the distributed setting. For a set $S \subseteq [n]$, the influence of S on f, $\mathsf{Inf}_S(f)$, is the probability that f is not constant upon setting all the variables *not* in S uniformly at random. The total influence of f, $\mathsf{Inf}(f)$, is defined as $\sum_{S:\ |S|=1} \mathsf{Inf}_S(f)$. Hence, intuitively, the total influence may be viewed as the expected number of coordinates of an input which, when flipped, will cause the value of f to be changed. For example, the Parity function on n variables has total influence n. That is, the parity function is never constant even when all but one of the variables are set. In particular, every variable has maximum possible influence of 1. Consider a $\mathsf{dictator}$ function $f(x_1, \ldots, x_i, \ldots, x_n) = x_i$. It follows that the influence of the i-th variable is 1, whereas the rest of the variables have 0 influence. Thus, exactly one variable has high influence. An interesting example is the $\mathsf{Majority}$ function, where each variable has *low* influence $\Theta(1/\sqrt{n})$. This is interesting because $\mathsf{Majority}$ is a also a balanced function, that is, $\Pr[\mathsf{Majority}(X) = 1] = \Pr[\mathsf{Majority}(X) = -1] = \frac{1}{2}$. Such functions, balanced and all variables having low influence, were the main object of study in [4].

1.1 Motivation

Resolving the FEI conjecture is one of the most important open problems in the Fourier analysis of Boolean functions. The conjecture intuitively asserts that if the Fourier coefficients of a Boolean function are "smeared out," then its influence must be large, i.e., at a typical input, the value of f changes in several different directions. The original motivation for the conjecture stems from a study of threshold phenomena in random graphs. The existence of sharp thresholds for various graph properties is one of the significant discoveries in the theory of random graphs [5]. Friedgut and Kalai [3] asked *how large can the threshold interval be for a monotone graph property?* Consider $f : \{0,1\}^n \to \{0,1\}$ representing a monotone graph property. Define $A_f(p) := \Pr[f(X_1, X_2, \ldots, X_n) = 1]$, where X_i's are independent random variables that is, 1 with probability p and 0 with probability $1-p$. Let $\delta > 0$ be a small number. By threshold interval we mean the length of the interval $[p, q]$ such that at p the probability that $\Pr_X[f(X) = 1]$ is δ, but at q the probability is $1 - \delta$. Then, the length of the threshold interval is inversely proportional to the derivative of $A_f(p)$, and by Russo's formula [6, 7], the derivative of $A_f(p)$ equals the total influence of f under the product measure where each bit is 1 with probability p and 0 otherwise. Hence, the graph property has a small threshold interval around p, that is, sharp threshold, if and only if it has large influence (under the product measure). Therefore, Friedgut and Kalai [3] asked for generic conditions that would force the influence to be large. Motivated by the Fourier-analytic formulae of the entropy and influence, they conjectured that a spread-out Fourier spectrum, i.e. large Fourier entropy, might be one such condition (cf. Eq. (2)).

The FEI conjecture has numerous applications [8]. It implies that for any n-vertex monotone graph property, the influence is at least $c(\log n)^2$. In other words, following the discussion in preceding paragraph it implies that for a monotone graph property on n vertices any threshold interval is of length at most $c'(\log n)^{-2}$. The best known upper bound, by Bourgain and Kalai [9], is $C_\epsilon(\log n)^{-2+\epsilon}$, for any $\epsilon > 0$. That is, a lower bound of $\Omega((\log n)^{2-\epsilon})$ on the influence of any n-vertex monotone graph property.

It also implies a variant of *Mansour's Conjecture* [10] stating that for a Boolean function computable by a DNF formula with m terms, most of its Fourier mass is concentrated on $\mathsf{poly}(m)$-many coefficients. A proof of Mansour's conjecture would imply a polynomial time *agnostic* learning algorithm for DNF's [11] answering a major open question in computational learning theory.

1.2 Prior Work

The first progress on the FEI conjecture was made in 2010 in [12] showing that the conjecture holds for random DNFs. O'Donnell et al. [13] proved that the conjecture holds for symmetric functions and more generally for any d-part symmetric functions for constant d. They also established the conjecture for functions computable by read-once decision trees. Keller et al. [14] studied a generalization of the conjecture to biased product measures on the Boolean cube and proved a variant of the conjecture for functions with extremely low Fourier weight on high levels. O'Donnell and Tan [1] verified the conjecture for read-once formulas using a composition theorem for the FEI conjecture. Wan et al. [15] studies the conjecture from the point of view of existence of efficient prefix-free codes for the random variable, $\mathcal{X} \sim \hat{f}^2$, that is distributed according to $\hat{f}^2$. Using this interpretation, they verify the conjecture for bounded-read decision trees. It is also relatively easy to show that the FEI conjecture holds for a random Boolean function, e.g., see [16] for a proof. By direct calculation, one can verify the conjecture for simple functions such as AND, OR, $\mathsf{Majority}$, Tribes etc.

1.3 Our results

We report here various upper bounds on Fourier entropy that may be viewed as progress toward the FEI conjecture.

Upper bounds by Complexity Measures The $\mathsf{Inf}(f)$ of a Boolean function f is used to derive lower bounds on a number of complexity parameters of f such as the number of leaves or the average depth of a decision tree computing f. Hence a natural weakening of the FEI conjecture is to prove upper bounds on the Fourier entropy in terms of such complexity measures of Boolean functions. By a relatively easy argument, we show that

$$\mathbb{H}(f) = O(\log \mathsf{L}(f)), \tag{3}$$

where $\mathsf{L}(f)$ denotes the minimum number of leaves in a decision tree that computes f. If $\mathsf{DNF}(f)$ denotes the minimum size of a DNF for the function f, note that $\mathsf{DNF}(f) \leqslant \mathsf{L}(f)$. Thus improving Eq. (3) with $O(\log \mathsf{DNF}(f))$ on the right hand side would resolve *Mansour's conjecture* – a long-standing open question about sparse Fourier approximations to DNF formulas motivated by applications to learning theory – and a special case of the FEI conjecture for DNF's. We note that Eq. (3) also holds when the queries made by the decision tree involve parities of subsets of variables, conjunctions of variables, etc. It also holds when $\mathsf{L}(f)$ is generalized to the number of subcubes in a *subcube partition* that represents f. Note that for a Boolean function

$$\mathsf{Inf}(f) \leqslant \log(L_c(f)) \leqslant \log(\mathsf{L}(f)) \leqslant D(f),$$

where $L_c(f)$ is number of subcubes in a *subcube partition* that represents f and $D(f)$ is the depth of the decision tree computing f.

We also prove the following strengthening of Eq. (3):

$$\mathbb{H}(f) = O(\bar{\mathsf{d}}(f)), \tag{4}$$

where $\bar{\mathsf{d}}(f)$ denotes the *average depth* of a decision tree computing f (observe that $\bar{\mathsf{d}}(f) \leqslant \log(\mathsf{L}(f))$). Note that the average depth of a decision tree is also a kind of entropy: it is given by the distribution induced on the leaves of a decision tree when an input is drawn uniformly at random. Thus Eq. (4) relates the two kinds of entropy, but only up to a constant factor. We further strengthen Eq. (4) by improving the right-hand side in Eq. (4) to *average depth* of a *parity* decision tree computing f, that is, queries made by the decision tree are parities of a subset of variables.

Upper bounds on the Fourier Entropy of Polynomial Threshold Functions The Fourier Entropy-Influence conjecture is known to be true for unweighted threshold functions, i.e., when $f(x) = \mathsf{sign}(x_1 + \cdots + x_n - \theta)$ for some integer $\theta \in [0..n]$. This follows from a result due to O'Donnell et al. [13] that the FEI conjecture holds for all symmetric Boolean functions. It is known that the influence for the class of linear threshold functions is $\Theta(\sqrt{n})$ (where the lower bound is witnessed by Majority [17]). Recently Harsha et al. [18] studied average sensitivity of polynomial threshold functions (see also [19]). They proved that average sensitivity of degree-d polynomial threshold functions is bounded by $O_d(n^{1-(1/4d+6)})$, where $O_d(\cdot)$ denotes that the constant depends on degree d. This suggests a natural and important weakening of the FEI conjecture: *Is Fourier Entropy of polynomial threshold functions bounded by a similar function of n as their average sensitivity?* In this paper we answer this question in the positive. An important ingredient in our proof is a bound on the derivative of noise sensitivity in terms of the noise parameter.

FEI inequality for Read-Once Formulas We also prove that the FEI conjecture holds for a special class of Boolean functions, namely Read-Once Formulas over {AND, OR and XOR}, i.e., functions computable by a tree with AND, OR and XOR gates at internal nodes and each variable (or its negation) occurring *at most once* at the leaves. Our result is independent of a concurrent (with the conference version of this paper) result by O'Donnell and Tan [1] that proves the FEI conjecture holds for read-once formulas that allow *arbitrary* gates of bounded fan-in. However, our proof is completely elementary and very different from theirs. Prior to these results, O'Donnell et al. [13] proved that the FEI conjecture holds for read-once *decision trees*. Our result for read-once formulas is a strict generalization of their result. For instance, the tribes function is computable by read-once formulas but not by read-once decision trees. Our proof for read-once formulas is a consequence of a kind of tensorizability for $\{0,1\}$-valued Boolean functions. In particular, we show that an inequality similar to the FEI inequality is preserved when functions depending on *disjoint* sets of variables are combined by AND, OR and XOR operators.

A Bound for Real-valued Functions via Second Moment Recall [20] that total influence $\mathsf{Inf}(f)$ or average sensitivity $\mathsf{as}(f)$ is related to $\hat{f}$ by the well-known identity: $\mathsf{as}(f) = \mathsf{Inf}(f) = \sum_S |S|\,\widehat{f}(S)^2$. Hence, an equivalent way to state the FEI conjecture is that there is an absolute constant C such that for all Boolean f,

$$\mathbb{H}(f) \leqslant C \cdot \sum_S |S|\,\widehat{f}(S)^2 \ . \tag{5}$$

Here, we prove that for all δ, $0 \leqslant \delta \leqslant 1$, and for all f with $\sum_S \widehat{f}(S)^2 = 1$,

$$\mathbb{H}(f) \leqslant \sum_S |S|^{1+\delta}\widehat{f}(S)^2 \ + \ (\log n)^{O(\frac{1}{\delta})} \ . \tag{6}$$

An alternative interpretation of the above theorem states

$$\mathbb{H}(f) \leqslant \mathsf{as}(f)^{1-\delta} \cdot \mathsf{as}_2(f)^{\delta} \ + \ (\log n)^{O(\frac{1}{\delta})} \ , \tag{7}$$

where $\mathsf{as}_2(f) := \sum_S |S|^2\,\widehat{f}(S)^2$. Note that $\mathsf{as}_2(f) \leqslant \mathsf{s}(f)^2$, where $\mathsf{s}(f)$ is the *maximum* sensitivity of f.

It is important to note that Eq. (6) holds for *arbitrary*, i.e., even non-Boolean, f such that (without loss of generality) $\sum_S \widehat{f}(S)^2 = 1$. On the other hand, there are examples of non-Boolean f for which the FEI conjecture Eq. (5) is *false* (see A).

Remainder of the paper We give basic definitions in Section 2. Section 3 contains upper bounds in terms of complexity measures. In Section 4 and Section 5, we consider special classes of Boolean functions namely, the *polynomial threshold functions* and *Read-Once formulas*. We then provide bounds for real valued functions in Section 6.

2 Preliminaries

We recall here some basic facts of Fourier analysis. For a detailed treatment, please refer to [21, 2]. Consider the space of all functions from $\{0,1\}^n$ to $\mathbb{R}$, endowed with the inner product $\langle f, g\rangle = 2^{-n}\sum_{x\in\{0,1\}^n} f(x)g(x)$. The character functions $\chi_S(x) := (-1)^{\sum_{i\in S} x_i}$ for $S \subseteq [n]$ form an orthonormal basis for this space of functions w.r.t. the above inner product. Thus, every function $f : \{0,1\}^n \longrightarrow \mathbb{R}$ of n Boolean variables has the *unique Fourier* expansion: $f(x) = \sum_{S\subseteq[n]} \hat{f}(S)\chi_S(x)$. The vector $\hat{f} = (\hat{f}(S))_{S\subseteq[n]}$ is called the Fourier transform of the function f. The Fourier coefficient $\hat{f}(S)$ of f at S is then given by, $\hat{f}(S) = 2^{-n}\sum_{x\in\{0,1\}^n} f(x)\chi_S(x)$. The norm of a function f is defined to be $\|f\| = \sqrt{\langle f, f\rangle}$. Orthonormality of $\{\chi_S\}$ implies *Parseval's identity*: $\|f\|^2 = \sum_S \widehat{f}(S)^2$.

We only consider *finite probability distributions* in this paper. The *entropy* of a distribution $\mathcal{D}$ is given by, $\mathbb{H}(\mathcal{D}) := \sum_{i\in\mathrm{supp}(\mathcal{D})} p_i \log\frac{1}{p_i}$. In particular, the *binary* entropy function, denoted by $\mathrm{H}(p)$, equals $-p\log p-(1-p)\log(1-p)$. All logarithms in the paper are base 2, unless otherwise stated.

We consider Boolean functions with range $\{-1,+1\}$. For an $f : \{0,1\}^n \to \{-1,+1\}$, $\|f\|$ is clearly 1 and hence Parseval's identity shows that for Boolean functions $\sum_S \widehat{f}(S)^2 = 1$. This implies that the squared Fourier coefficients can be thought of as a probability distribution and the notion of Fourier Entropy Eq. (1) is well-defined.

The *influence of f in the i-th direction*, denoted $\mathsf{Inf}_i(f)$, is the fraction of inputs at which the value of f gets flipped if we flip the i-th bit:

$$\mathsf{Inf}_i(f) = 2^{-n}|\{x \in \{0,1\}^n : f(x) \neq f(x \oplus e_i)\}| \ ,$$

where $x \oplus e_i$ is obtained from x by flipping the i-th bit of x.

The (total) *influence* of f, denoted by $\mathsf{Inf}(f)$, is $\sum_{i=1}^n \mathsf{Inf}_i(f)$. The influence of i on f can be shown, e.g., [20], to be $\mathsf{Inf}_i(f) = \sum_{S\ni i} \hat{f}(S)^2$ and hence it follows that $\mathsf{Inf}(f) = \sum_{S\subseteq[n]} |S|\hat{f}(S)^2$.

For $x \in \{0,1\}^n$, the *sensitivity of f at x*, denoted $\mathsf{s}_f(x)$, is defined to be $\mathsf{s}_f(x) := |\{i : f(x) \neq f(x\oplus e_i), 1 \leqslant i \leqslant n\}|$, i.e., the number of coordinates of x, which when flipped, will flip the value of f. The (maximum) *sensitivity of the function f*, denoted $\mathsf{s}(f)$, is defined to be the largest sensitivity of f at x over all $x \in \{0,1\}^n$: $\mathsf{s}(f) := \max\{\mathsf{s}_f(x) : x \in \{0,1\}^n\}$. The *average sensitivity of f*, denoted $\mathsf{as}(f)$, is defined to be $\mathsf{as}(f) := 2^{-n}\sum_{x\in\{0,1\}^n} s_f(x)$. It is easy to see that $\mathsf{Inf}(f) = \mathsf{as}(f)$ and hence we also have $\mathsf{as}(f) = \sum_{S\subseteq[n]} |S|\hat{f}(S)^2$.

The noise sensitivity of f at ϵ, $0 \leqslant \epsilon \leqslant 1$, denoted $\mathsf{NS}_\epsilon(f)$, is given by $\Pr_{x,y\sim_\epsilon x}[f(x) \neq f(y)]$, where x is chosen uniformly at random, and $y \sim_\epsilon$

x denotes that y is obtained by flipping each bit of x independently with probability ϵ. It follows that $\mathsf{NS}_\epsilon(f) = \frac{1}{2} - \frac{1}{2}\sum_S (1-2\epsilon)^{|S|}\hat{f}(S)^2$, see for instance, [22], or Theorem 2.49 in [2]. Hence the derivative of $\mathsf{NS}_\epsilon(f)$ with respect to ϵ, denoted $\mathsf{NS}'_\epsilon(f)$, equals $\sum_{S\neq\emptyset}|S|(1-2\epsilon)^{|S|-1}\hat{f}(S)^2$.

3 Bounding Entropy using Complexity Measures

In this section, we prove upper bounds on Fourier entropy in terms of some complexity parameters associated to decision trees and subcube partitions.

3.1 *via* leaf entropy : Average Decision Tree Depth

Let T be a decision tree computing $f : \{0,1\}^n \to \{+1,-1\}$ on variable set $X = \{x_1, \ldots, x_n\}$. If $A_1, \ldots, A_L$ are the sets (with repetitions) of variables queried along the root-to-leaf paths in the tree T, then the average depth (w.r.t. the uniform distribution on inputs) of T is defined to be $\bar{d} := \sum_{i=1}^{L} |A_i| 2^{-|A_i|}$. Note that the average depth of a decision tree is also a kind of entropy: if each leaf λ_i is chosen with the probability $p_i = 2^{-|A_i|}$ that a uniformly chosen random input reaches it, then the entropy of the distribution induced on the λ_i is $\mathbb{H}(\lambda_i) = -\sum_i p_i \log p_i = \sum_i |A_i| 2^{-|A_i|}$. Here, we will show that *the Fourier entropy is at most twice the leaf entropy of a decision tree.*

Without loss of generality, let x_1 be the variable queried by the root node of T and let T_1 and T_2 be the subtrees reached by the branches $x_1 = +1$ and $x_1 = -1$ respectively and let g_1 and g_2 be the corresponding functions computed on variable set $Y = X \setminus \{x_1\}$. Let $\bar{d}$ be the average depth of T and $\bar{d_1}$ and $\bar{d_2}$ be the average depths of T_1 and T_2 respectively. We first observe a fairly straightforward lemma relating Fourier coefficients of f to the Fourier coefficients of restrictions of f.

Lemma 3.1. *Let $S \subseteq \{2, \ldots, n\}$.*

(i) $\widehat{f}(S) = (\widehat{g_1}(S) + \widehat{g_2}(S))/2.$

(ii) $\widehat{f}(S \cup \{1\}) = (\widehat{g_1}(S) - \widehat{g_2}(S))/2.$

(iii) $\bar{d} = (\bar{d_1} + \bar{d_2})/2 + 1.$

Proof. Observe that

$$f(x_1, x_2, \ldots, x_n) = f(x_1, y) = \frac{(1 + x_1)}{2} g_1(y) + \frac{(1 - x_1)}{2} g_2(y)$$
$$= \frac{1}{2}(g_1(y) + g_2(y)) + \frac{x_1}{2}(g_1(y) - g_2(y)).$$

(i) and (ii) now follow by linearity of the Fourier transform.

Let $\{B_i\}_{i=1}^{L_1}$ be the variable sets queried along the root-to-leaf paths in T_1 and similarly let $\{C_i\}_{i=1}^{L_2}$ be the variable sets queried along the root-to-leaf paths in T_2. Then, note that the variable sets $\{A_i\}_{i=1}^{L}$, where $L = L_1 + L_2$, queried along the root-to-leaf paths in T are given by

$$\{B_i \cup \{x_1\}\}_{i=1}^{L_1} \bigcup \{C_i \cup \{x_1\}\}_{i=1}^{L_2} .$$

It thus follows that

$$\begin{aligned}
\bar{d} &:= \sum_{i=1}^{L} |A_i| 2^{-|A_i|} \\
&= \sum_{1=1}^{L_1} (|B_i| + 1) 2^{-|B_i|-1} \; + \; \sum_{1=1}^{L_2} (|C_i| + 1) 2^{-|C_i|-1} \\
&= \frac{1}{2} \sum_{1=1}^{L_1} |B_i| 2^{-|B_i|} + \frac{1}{2} \sum_{1=1}^{L_1} 2^{-|B_i|} + \frac{1}{2} \sum_{1=1}^{L_2} |C_i| 2^{-|C_i|} + \frac{1}{2} \sum_{1=1}^{L_2} 2^{-|C_i|} \\
&= \frac{1}{2}\bar{d}_1 + \frac{1}{2} + \frac{1}{2}\bar{d}_2 + \frac{1}{2},
\end{aligned}$$

where the last line follows by applying the definition of average depth to T_1 and noting that $\sum_{1=1}^{L_1} 2^{-|B_i|} = 1$ for the decision tree T_1 and similarly for T_2. This proves (iii). □

Remark 3.1. *Note that g_1 and g_2 differ on an input y if and only if f is sensitive to x_1 at (x_1, y). In particular, it is easy to see $\frac{1}{4}\|g_1 - g_2\|^2 = \mathsf{Inf}_f(1)$ and $\frac{1}{4}\|g_1 + g_2\|^2 = 1 - \mathsf{Inf}_f(1)$.*

Using Lemma 3.1 and concavity of entropy we establish the following technical lemma, which relates the entropy of f to entropies of restrictions of f.

Lemma 3.2. *Let g_1 and g_2 be defined as before in Lemma 3.1. Then,*

$$\mathbb{H}(f) \leqslant \frac{1}{2}\,\mathbb{H}(g_1) + \frac{1}{2}\,\mathbb{H}(g_2) + 2 \ . \tag{8}$$

Proof. We will use the concavity of the function $x \log \frac{1}{x}$ (for $0 \leqslant x \leqslant 1$)[1]. For simplicity of notation below, let $N' := \{2, \ldots, n\}$.

$$\begin{aligned}
\mathbb{H}(f) &= \sum_{T \subseteq [n]} \widehat{f}(T)^2 \log \frac{1}{\widehat{f}(T)^2} \\
&= \sum_{S \subseteq N'} \widehat{f}(S)^2 \log \frac{1}{\widehat{f}(S)^2} + \widehat{f}(S \cup \{1\})^2 \log \frac{1}{\widehat{f}(S \cup \{1\})^2} \\
&\leqslant \sum_{S \subseteq N'} (\widehat{f}(S)^2 + \widehat{f}(S \cup \{1\})^2) \log \frac{2}{\widehat{f}(S)^2 + \widehat{f}(S \cup \{1\})^2} \\
&= \sum_{S \subseteq N'} \frac{\widehat{g_1}(S)^2 + \widehat{g_2}(S)^2}{2} \log \frac{4}{\widehat{g_1}(S)^2 + \widehat{g_2}(S)^2} \text{ (by Lemma 3.1 (i) and (ii))} \\
&= \frac{1}{2} \sum_{S \subseteq N'} \widehat{g_1}(S)^2 \log \frac{1}{\widehat{g_1}(S)^2 + \widehat{g_2}(S)^2} + \frac{1}{2} \sum_{S \subseteq N'} \widehat{g_2}(S)^2 \log \frac{1}{\widehat{g_1}(S)^2 + \widehat{g_2}(S)^2} \\
&\qquad + \sum_{S \subseteq N'} \widehat{g_1}(S)^2 + \widehat{g_2}(S)^2 \\
&\leqslant \frac{1}{2} \sum_{S \subseteq N'} \widehat{g_1}(S)^2 \log \frac{1}{\widehat{g_1}(S)^2} + \frac{1}{2} \sum_{S \subseteq N'} \widehat{g_2}(S)^2 \log \frac{1}{\widehat{g_2}(S)^2} + 2.
\end{aligned}$$

The first inequality follows from the concavity of $x \log \frac{1}{x}$, and the last because of the monotonicity of Logarithm, and Parseval's identitiy, that is, $\sum_{S \subseteq N'} \widehat{g_1}(S)^2 = \sum_{S \subseteq N'} \widehat{g_2}(S)^2 = 1$. □

Let $\bar{\mathsf{d}}(f)$ denote the minimum *average* depth of a decision tree computing f. As a consequence of Lemma 3.2 we obtain the following theorem.

Theorem 3.3. *For every Boolean function f, $\mathbb{H}(f) \leqslant 2 \cdot \bar{\mathsf{d}}(f)$.*

Proof. The proof is by induction on the number of variables of f.

$$\begin{aligned}
\mathbb{H}(f) &\leqslant \frac{1}{2} \mathbb{H}(g_1) + \frac{1}{2} \mathbb{H}(g_2) + 2 && \text{(by Lemma 3.2)} \\
&\leqslant \bar{d}_1 + \bar{d}_2 + 2 && \text{(by induction, } \mathbb{H}(g_i) \leqslant 2\bar{d}_i \text{ for } i = 1, 2\text{)} \\
&= 2\bar{d} && \text{(by Lemma 3.1 (iii)).}
\end{aligned}$$

□

Remark 3.2. *The constant* 2 *in the bound of Theorem 3.3 cannot be replaced by* 1. *Indeed, let* $f(x, y) = x_1 y_1 + \cdots + x_{n/2} y_{n/2} \bmod 2$ *be the inner product*

[1] That is, $x \log \frac{1}{x} + y \log \frac{1}{y} \leqslant (x + y) \log \frac{2}{x+y}$.

mod 2 function. Then because $\widehat{f}(S)^2 = 2^{-n}$ for all $S \subseteq [n]$, $\mathbb{H}(f) = n$. On the other hand, it can be shown that $\bar{\mathsf{d}}(f) = \frac{3}{4}n - o(n)$. Hence, the constant must be at least $4/3$. We will see later (Theorem 3.5 and Remark 3.3) that the above proof technique cannot yield a constant factor better than 2.

3.1.1 Average Parity Decision Tree Depth

Applying a linear transformation L on a Boolean function f, we obtain another Boolean function Lf which is defined as $Lf(x) := f(Lx)$, for all $x \in \{0,1\}^n$. We begin with a useful observation.

Proposition 3.4. *Let $f : \{0,1\}^n \to \{+1,-1\}$ be a Boolean function. For an invertible linear transformation $L \in \mathsf{GL}_n(\mathbb{F}_2)$, $\mathbb{H}(f) = \mathbb{H}(Lf)$.*

Proof. The porposition follows if we show that L permutes the Fourier-spectrum of f. Let us consider the Fourier coefficients of Lf. Let a row vector $y \in \{0,1\}^n$ denote a subset $S \subseteq [n]$, that is, $y_i = 1$ iff $i \in S$. Then,

$$\begin{aligned} \widehat{Lf}(y) &= \sum_{x \in \{0,1\}^n} Lf(x) \cdot (-1)^{\langle y,x \rangle} = \sum_{x \in \{0,1\}^n} f(Lx) \cdot (-1)^{\langle yL^{-1}, Lx \rangle} \\ &= \sum_{z \in \{0,1\}^n} f(z) \cdot (-1)^{\langle yL^{-1}, z \rangle} = \widehat{f}(yL^{-1}). \end{aligned}$$

□

Let T be a parity decision tree computing $f : \{0,1\}^n \to \{+1,-1\}$ on variable set $X = \{x_1, \ldots, x_n\}$. Also, let L be an invertible linear transformation. Note that a parity decision tree computing f also computes Lf and vice versa. This implies that, after applying a linear transformation, we can always assume that a variable is queried at the root node of T. Let us denote the new variable set, after applying the linear transformation, by $Y = \{y_1, \ldots, y_n\}$. Without loss of generality, let y_1 be the variable queried at the root. Let T_1 and T_2 be the subtrees reached by the branches $y_1 = 0$ and $y_1 = 1$ respectively, and let g_1 and g_2 be the corresponding functions computed on variable set $Y \setminus \{y_1\}$. Using Proposition 3.4, we see that the proofs of Lemma 3.1 and Lemma 3.2 go through in the setting of parity decision trees too. Hence, we get the following strengthening of Theorem 3.3.

Theorem 3.5. *For every Boolean function f, $\mathbb{H}(f) \leqslant 2 \cdot \oplus\text{-}\bar{\mathsf{d}}(f)$, where $\oplus\text{-}\bar{\mathsf{d}}(f)$ denotes the minimum average depth of a parity decision tree computing f.*

Remark 3.3. *The constant 2 in the bound of Theorem 3.5 is optimal, that is, it cannot be replaced by a smaller number. As before, we consider the inner product mod 2 function. It's Fourier entropy is n, but* $\left(\frac{n}{2}+1\right) \geqslant \oplus\text{-}\bar{\mathsf{d}}(f) \geqslant \mathsf{Inf}(f) = \frac{1}{2^n}\sum_{k=1}^{n} k \cdot \binom{n}{k} = \frac{n}{2}$.

3.2 *via* L_1-norm : Decision Trees and Subcube Partitions

Note that a decision tree computing a Boolean function f induces a partition of the cube $\{0,1\}^n$ into monochromatic subcubes, i.e., f has the same value on all points in a given subcube, with one subcube corresponding to each leaf. But there exist monochromatic subcube partitions that are not induced by any decision tree. Consider any subcube partition $\mathcal{C}$ computing f (see Definition 3.7). There is a natural way to associate a probability distribution with $\mathcal{C}$: C_i has probability mass $2^{-(\text{number of co-ordinates fixed by } C_i)}$. Let us call the entropy associated with this probability distribution *partition entropy.* Based on the results of the previous subsection, a natural direction would be to prove that the Fourier entropy is bounded by the partition entropy. Unfortunately we were not quite able to show that but, interestingly, there is a very simple proof to see that the Fourier entropy is bounded by the logarithm of the number of partitions in $\mathcal{C}$. In fact, the proof gives a slightly better upper bound of the logarithm of the spectral-norm of f. For completeness sake, we note this observation but we remark that it should be considered folklore.

Lemma 3.6. *(Folklore) Let $f : \{0,1\}^n \to \mathbb{R}$ be such that $\sum_S \hat{f}(S)^2 \leqslant 1$. Let $L_1(f) := \sum_S |\hat{f}(S)|$ be the L_1-norm of the Fourier transform of f. Assume, further, that $L_1(f) \geqslant 1$. Then, $\mathbb{H}(f) \leqslant O(\log L_1(f))$.*

Proof. There are many ways to prove this, but we present one particular proof.

Let $L := L_1(f)$, and $\theta := 1/(16L^2)$. Define $\mathcal{G} := \{S : |\hat{f}(S)| \geqslant \theta\}$. Note that for $x \geqslant 16$, $\log x \leqslant \sqrt{x}$. We thus have $\log \frac{1}{|\hat{f}(S)|} \leqslant \frac{1}{\sqrt{|\hat{f}(S)|}}$, for $S \notin \mathcal{G}$. Therefore,

$$\begin{aligned}
\mathbb{H}(f) = \sum_S \widehat{f}(S)^2 \log \frac{1}{\widehat{f}(S)^2} &\leqslant \sum_{S\in\mathcal{G}} \hat{f}(S)^2 \log \frac{1}{\hat{f}(S)^2} + 2\sum_{S\notin\mathcal{G}} \hat{f}(S)^2 \frac{1}{\sqrt{|\hat{f}(S)|}} \\
&\leqslant \log\frac{1}{\theta^2} \sum_{S\in\mathcal{G}} \hat{f}(S)^2 + 2\max_{S\notin\mathcal{G}} \sqrt{|\hat{f}(S)|} \sum_{S\notin\mathcal{G}} |\hat{f}(S)| \\
&\leqslant \log(256L^4) + 2\cdot\frac{1}{4L}\cdot L = 4\log L + 8.5.
\end{aligned}$$

$\square$

It is well-known [23, 24] and easy to prove[2] that the L_1-norm of a function, f, is at most the *minimum* number of leaves, $L(f)$, in a decision tree computing f. In fact, even if we allow the queries at each internal node of a decision tree to be parities or conjunctions of subsets of variables (or more generally any function with bounded L_1-norm), then also we have $L_1(f) = O(L(f))$. Let us consider how other complexity measures associated with a Boolean function compare with L_1-norm. We start with some definitions.

The decision tree *depth* $D(f)$ of a function f is the minimum depth (length of a longest root-to-leaf path) of a decision tree computing f. The *degree* $\deg(f)$ is the degree of the (unique) multilinear polynomial over $\mathbb{R}$ that represents f. The *block sensitivity* $\mathsf{bs}_f(x)$ on an input x is the maximum number of disjoint subsets $B_1, \ldots, B_t$ of $[n]$ such that for all j, $f(x) \neq f(x \oplus e_{B_j})$, where e_{B_j} is the characteristic vector of the set B_j. The *block sensitivity* $\mathsf{bs}(f)$ is $\max_x \mathsf{bs}_f(x)$. The *certificate complexity* $\mathsf{C}(f)$ measures how many of the variables have to be given a value in order to fix the value of f. More precisely, an f-certificate of an input x is a subset S of $[n]$ with an assignment $\alpha \in \{0,1\}^{|S|}$ such that $x|_S = \alpha$, and for all input y such that $y|_S = x|_S$, $f(x) = f(y)$. The *size* of a certificate is the cardinality of the subset S. The *certificate complexity* $\mathsf{C}_f(x)$ on an input x is the size of a smallest f-certificate for x. The *certificate complexity* $\mathsf{C}(f)$ of a function is $\max_x \mathsf{C}_f(x)$.

Fig. 1 illustrates known relationship between the measures; $a \to b$ implies that $a = O(b)$. It is easy to see that the relationships in the figure follows more or less from their definitions. Thus, using Lemma 3.6 with Fig. 1, we could bound Fourier entropy by combinatorial measures. In particular, we immediately have

$$\mathbb{H}(f) = O(\log L(f)), \mathbb{H}(f) = O(D(f)), \text{ and } \mathbb{H}(f) = O(\deg(f)). \tag{9}$$

Remark 3.4. *A natural question to ask is how important Boolean-ness of functions is in the entropy upper bounds. While Lemma 3.6 holds for real-valued functions as well, we note that the inequalities in Eq.* (9) *hold* only for Boolean-valued *functions. In fact, we give examples in A to show that these bounds fail for non-Boolean functions.*

3.3 *via* Concentration : Subcube Partitions

As established before, we can generalize the bound $\mathbb{H}(f) = O(\log L(f))$ in Eq. (9) to subcube partitions, that is, $\mathbb{H}(f) = O(\log L_c(f))$. Nevertheless, we

[2] See also Lemma 3.8.

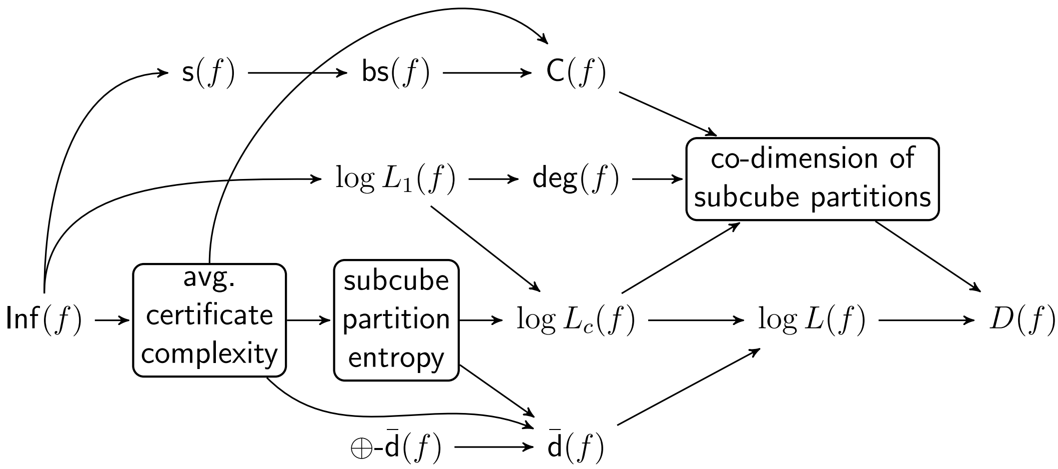

Figure 1: Relationship among complexity measures.

present the generalization to subcube partitions here. Our goal is to illustrate a different approach. The approach uses the concentration property of the Fourier transform and uses a general, potentially powerful, technique. One way to do this is to use a result due to Bourgain and Kalai (Theorem 3.2 in [14]). However, we give a more direct proof for the special case of subcube partitions.

Definition 3.7. *A* subcube *C of the cube $B_n := \{0,1\}^n$ is given by a mapping (partial assignment) $\alpha : [n] \to \{-1,+1,*\}$ and is defined to be the set of all vectors in B_n that agree with α on coordinates* fixed*, i.e., assigned a non-$*$ value, by α: $C := C_\alpha := \{x \in B_n \ : \ \alpha(i) \neq * \implies x_i = \alpha(i)\}$. We use $A := \{i \in [n] : \alpha(i) \neq *\}$ to denote the set of fixed coordinates of α and denote the cube C also by the pair (A,α). The cardinality of the set A is called the* co-dimension *of C.*

For a function $f : \{0,1\}^n \to \{+1,-1\}$, a partition $\mathcal{C} = \{C_1,\ldots,C_m\}$ of B_n into subcubes C_i such that f is constant on each C_i is called a (monochromatic) subcube partition *with respect to f. If $\mathcal{C}$ is a subcube partition monochromatic w.r.t. f, we also say $\mathcal{C}$* computes *f. We define the* co-dimension *of a subcube partition $\mathcal{C}$ as,* $\max_i$ *co-dimension(C_i).*

We denote by $L_c(f)$ the minimum number of subcubes in a subcube partition that computes f.

The most natural subcube partitions w.r.t. a function f are the ones induced by decision trees computing f: the set of all inputs reaching a leaf of the decision tree is given by a subcube C_α, where α denotes the partial assignment defined by the path from the root to that leaf. But there exist subcube partitions that are not induced by any decision tree.

Suppose f is computed by a subcube partition $\mathcal{C} = \{C_1, \ldots, C_L\}$, where $C_i = (A_i, \alpha_i)$. Let $\phi_i : \{0,1\}^n \to \{0,1\}$ be the characteristic function of the subcube C_i : $\phi_i(x) = 1$ if $x \in C_i$ and $\phi_i(x) = 0$ otherwise. Let $\beta_i \in \{-1, +1\}$ be the value of f on C_i. Then, clearly

$$f(x) = \sum_{i=1}^{L} \beta_i \phi_i(x).$$

By linearity of the Fourier transform, it follows that $\hat{f}(S) = \sum_{i=1}^{L} \beta_i \hat{\phi}_i(S)$. A simple calculation shows that, for the characteristic function ϕ of a subcube $C = (A, \alpha)$, the Fourier transform is given by $\hat{\phi}(S) = 2^{-|A|}\chi_S(\alpha)$ if $S \subseteq A$ and $\hat{\phi}(S) = 0$ otherwise. It follows that

$$\hat{f}(S) = \sum_{i\,:\,S \subseteq A_i} 2^{-|A_i|} \cdot \beta_i \chi_S(\alpha_i). \tag{10}$$

In particular, $\hat{f}(S) \neq 0 \implies \exists i\ S \subseteq A_i$.

The following lemma directly follows from the above observations.

Lemma 3.8 ([25]). *Let f be computed by the subcube partition $\mathcal{C} = \{C_1, \ldots, C_L\}$, where $C_i = (A_i, \alpha_i)$. Then,*

(i) $\sum_S |\hat{f}(S)| \leqslant L$, *and*

(ii) *For any integer $t \geqslant 0$, $\sum_{|S| \geqslant t} \hat{f}^2(S) \leqslant \sum_{|A_i| \geqslant t} 2^{-|A_i|}$.*

Proof. Using Eq. (10),

$$\sum_S |\hat{f}(S)| = \sum_S \Big| \sum_{i\,:\,S \subseteq A_i} \beta_i \chi_S(\alpha_i) 2^{-|A_i|} \Big| \leqslant \sum_S \sum_{i\,:\,S \subseteq A_i} 2^{-|A_i|} = L.$$

By Eq. (10), if $|S| \geqslant t$, the contribution to $\hat{f}(S)$ comes from only the C_i such that $|A_i| \geqslant t$. Let $g \equiv \sum_{|A_i| \geqslant t} \beta_i \phi_i$ be the restriction of f to subcubes with codimension $\geqslant t$. It is then clear that

$$\sum_{|S| \geqslant t} \widehat{f}(S)^2 = \sum_{|S| \geqslant t} \widehat{g}(S)^2 \leqslant \sum_S \widehat{g}(S)^2 = 2^{-n} \sum_{|A_i| \geqslant t} |C_i| = \sum_{|A_i| \geqslant t} 2^{-|A_i|}.$$

This proves (ii). $\square$

Combining Lemma 3.8(i) and Lemma 3.6, it immediately follows that $\mathbb{H}(f) = O(\log L_c(f))$.

However, we give here a different approach to prove essentially the same result. The approach uses the concentration property of the Fourier transform and illustrates a general, potentially powerful, technique. One way to do this is to combine Lemma 3.8(ii) with the following result due to Bourgain and Kalai:

Theorem 3.9 (Bourgain-Kalai, cited in [14])**.** *For $f : \{0,1\}^n \to \{+1, -1\}$, suppose that there exist $c_0 > 0$, $0 < a < 1/2$, and integer k, such that for all t,*

$$\sum_{S:|S|>t} \widehat{f}(S)^2 \leqslant e^{c_0 k} \cdot e^{-at}.$$

Then, for any $\alpha > 1$, there exists a set $\mathcal{B}_\alpha$ such that

(i) $\log |\mathcal{B}_\alpha| \leqslant C \cdot \alpha k$*, where C depends only on a and c_0, and*

(ii) $\sum_{S \notin \mathcal{B}_\alpha} \widehat{f}(S)^2 \leqslant n^{-\alpha}$.

However, we give a more direct proof here that nevertheless derives statements analogous to (i) and (ii) of Theorem 3.9, but for the special case of subcube partitions.

Theorem 3.10. *Let $f : \{0,1\}^n \to \{+1, -1\}$ be computed by a subcube partition $\mathcal{C}$ of size L. Then,*

$$\mathbb{H}(f) \leqslant 2 \log L(f) + 2 \log n + 2.$$

Corollary 3.11. *Let $f : \{0,1\}^n \to \{+1, -1\}$ depend on all its variables and be computed by a subcube partition $\mathcal{C}$ of size $L(f)$. Then, for some absolute constant $c > 1$,*

$$\mathbb{H}(f) \leqslant c \cdot \log L(f).$$

Proof. We use the observation that a function that depends on all its n variables requires a subcube partition of size at least $n+1$ to compute it. This is proved in Lemma 3.12 below. It follows that $\log n \leqslant \log L(f)$ and using this in Theorem 3.10, we get the corollary. □

Proof. (of Theorem 3.10) To bound entropy via concentration, we use the following simple idea. Suppose $\mathcal{E}$ is a subset of Fourier coefficients of a Boolean function f such that $\sum_{S \in \mathcal{E}} \widehat{f}(S)^2 = \epsilon$. For a subset of coefficients $\mathcal{B}$, let $\mathbb{H}(\mathcal{B})$ denote the Fourier entropy restricted to that set $\mathcal{B}$, appropriately normalized. Then a simple manipulation shows

$$\sum_S \widehat{f}(S)^2 \log \frac{1}{\widehat{f}(S)^2} = (1-\epsilon)\,\mathbb{H}(\overline{\mathcal{E}}) + \epsilon\,\mathbb{H}(\mathcal{E}) + \mathrm{H}(\epsilon), \tag{11}$$

where $\mathrm{H}(p) := p \log \frac{1}{p} + (1-p)\log\frac{1}{1-p}$ is the binary entropy function.

Now, let

$$\mathcal{B}_t := \{S \;:\; \exists\, i\, |A_i| \leqslant t \;\text{ such that }\; S \subseteq A_i\}.$$

Note that if $S \notin \mathcal{B}_t$, then every set A_i that contains S must have size larger than t. Hence, using Eq. (10), only sets of size larger than t contribute to such $\hat{f}(S)$. We now argue as in the proof of Lemma 3.8(ii). Let $g \equiv \sum_{|A_i|>t} \beta_i \phi_i$ be the restriction of f to subcubes with co-dimension $> t$. It is then clear that

$$\sum_{S\notin\mathcal{B}_t} \widehat{f}(S)^2 = \sum_{S\notin\mathcal{B}_t} \widehat{g}(S)^2 \leqslant \sum_S \widehat{g}(S)^2 = 2^{-n} \sum_{|A_i|>t} |C_i| = \sum_{|A_i|>t} 2^{-|A_i|} < 2^{-t}L. \tag{12}$$

Since $\sum_i 2^{-|A_i|} = 1$, we have that $|\{i \;:\; |A_i| \leqslant t\}| \leqslant 2^t$. Since every $S \in \mathcal{B}_t$ is a subset of some A_i with $|A_i| \leqslant t$, it follows

$$|\mathcal{B}_t| \leqslant \sum_{|A_i|\leqslant t} 2^{|A_i|} \leqslant 2^t \cdot |\{i \;:\; |A_i| \leqslant t\}| \leqslant 2^{2t}. \tag{13}$$

Fix $t := \log(Ln)$. We can now estimate the Fourier entropy of a subcube partition:

$$\begin{aligned}
\mathbb{H}(f) &= \sum_S \hat{f}^2(S) \log \frac{1}{\hat{f}^2(S)} \\
&= (1-1/n)\,\mathbb{H}(\hat{f}^2(S) \;:\; S \in \mathcal{B}_t) + (1/n)\,\mathbb{H}(\hat{f}^2(S) \;:\; S \notin \mathcal{B}_t) + \mathrm{H}(1/n) \\
&\leqslant (1-1/n)\log|\mathcal{B}_t| + 1/n \cdot n + \mathrm{H}(1/n) \\
&\leqslant 2t + 1 + \mathrm{H}(1/n) \\
&\leqslant 2\log L + 2\log n + 2.
\end{aligned}$$

The second equality follows from using Eq. (11) and Eq. (12), and the next inequality follows from Eq. (13). □

Lemma 3.12. *Suppose $f : \{0,1\}^n \to \{+1,-1\}$ depends on all its variables. Then any subcube partition that computes f must have at least $n+1$ subcubes in it.*

Proof. To prove this lemma, we will use the following theorem proved in [26].

Theorem: Suppose $f : \{0,1\}^n \to \{+1,-1\}$ depends on all its variables. Then there must exist an index $i \in [n]$ such that at least one of the restrictions $f_{\upharpoonright x_i=0}$ or $f_{\upharpoonright x_i=1}$ must depend on all the remaining variables in $[n] \setminus \{i\}$.

We can now prove the lemma by induction on n. For $n = 1$, the claim is trivial since if the function depends on a variable, the variable and its complement must be in different (single point) subcubes. For $n > 1$, we note that since the function depends on all its variables, for *every* variable x_j, there must be at least one subcube fixing $x_j = 0$ and at least one subcube fixing $x_j = 1$. Now, let x_i be a variable given by the above theorem such that, say, $f_{\upharpoonright x_i=0}$ depends on all its $n-1$ variables. By induction, we must have at least n subcubes in the restricted partition computing $f_{\upharpoonright x_i=0}$, where the restricted partition is obtained by restricting each of the subcubes in the original partition computing f to $x_i = 0$ half-cube. In the $x_i = 1$ half-cube, we must have at least one subcube, namely the one that restricts $x_i = 1$ in the original partition. All the n subcubes previously counted are disjoint from this since they either restricted $x_i = 0$ in the original partition or they didn't restrict x_i at all. So, all together we must have $n + 1$ subcubes in the original partition computing f. $\square$

4 Upper bound on Fourier Entropy of Threshold Functions

In this section, we establish a better upper bound on the Fourier entropy of polynomial threshold functions. We show that the Fourier entropy of a linear threshold function is $O(\sqrt{n})$, and we also show that for a degree-d threshold function it is $O_d(n^{1-\frac{1}{4d+6}})$. We remark that the bound is significant because the average sensitivity of a linear threshold function on n variables is $O(\sqrt{n})$. Moreover, Majority over n bits Maj_n is a linear threshold function such that both $\mathsf{Inf}(\mathsf{Maj}_n)$ and $\mathbb{H}(\mathsf{Maj}_n)$ are $\Omega(\sqrt{n})$. Also our upper bound on the Fourier entropy of degree-d threshold functions is of the same order as the best known upper bound on their average sensitivity [18, 19].

For $f : \{0,1\}^n \to \{+1,-1\}$, let $W^k[f] := \sum_{|S|=k} \widehat{f}(S)^2$ and $W^{\geqslant k}[f] := \sum_{|S|\geqslant k} \widehat{f}(S)^2$. We first note a simple inequality.

Proposition 4.1. *[17] For any $f : \{0,1\}^n \to \{+1,-1\}$, $\epsilon \in (0, \frac{1}{2}]$,*

$$\sum_{S:|S|\geqslant 1/\epsilon} \hat{f}(S)^2 \leqslant \frac{2}{1-e^{-2}} \mathsf{NS}_\epsilon(f).$$

Using Proposition 4.1 we prove our main technical lemma which translates a bound on noise sensitivity to a bound on the derivative of noise sensitivity.

Lemma 4.2. *Let* $f : \{0,1\}^n \to \{+1,-1\}$ *be such that* $\mathsf{NS}_\epsilon(f) \leqslant \alpha \cdot \epsilon^\beta$*, where* α *is independent of* ϵ *and* $\beta < 1$*. Then,*

$$\mathsf{NS}'_\epsilon(f) \leqslant \frac{5}{1-e^{-2}} \cdot \frac{\alpha}{1-\beta} \cdot (1/\epsilon)^{1-\beta} \ .$$

Proof. We start with the formula for the derivative of noise sensitivity in terms of the Fourier weights.

$$\begin{aligned}
\mathsf{NS}'_\epsilon(f) &= \sum_{k=1}^{n} W^k[f] \cdot k \cdot (1-2\epsilon)^{k-1} \\
&= \sum_{k=1}^{t} W^k[f] \cdot k \cdot (1-2\epsilon)^{k-1} + \sum_{k=t+1}^{n} W^k[f] \cdot k \cdot (1-2\epsilon)^{k-1}, \quad (t = \lfloor 1/\epsilon \rfloor) \\
&\leqslant \sum_{k=1}^{t} W^k[f] \cdot k + \sum_{k=t}^{n} W^k[f] \cdot k \cdot (1-2\epsilon)^{k-1}. \qquad (14)
\end{aligned}$$

Let $T_1 := \sum_{k=1}^{t} W^k[f] \cdot k$, and $T_2 := \sum_{k=t}^{n} W^k[f] \cdot k \cdot (1-2\epsilon)^{k-1}$. We will bound these sums individually using Proposition 4.1.

$$\begin{aligned}
T_1 = \sum_{k=1}^{t} W^k[f] \cdot k \leqslant \sum_{k=1}^{t} W^{\geqslant k}[f] &\leqslant \frac{2}{1-e^{-2}} \sum_{k=1}^{t} \mathsf{NS}_{\frac{1}{k}}(f) \\
&\leqslant \frac{2}{1-e^{-2}} \sum_{k=1}^{t} \alpha \cdot k^{-\beta} \simeq \frac{2}{1-e^{-2}} \cdot \alpha \cdot \frac{t^{1-\beta}}{1-\beta} \\
&\leqslant \frac{2}{1-e^{-2}} \cdot \frac{\alpha}{1-\beta} \cdot (1/\epsilon)^{1-\beta}. \qquad (15)
\end{aligned}$$

$$
\begin{aligned}
T_2 &= \sum_{k=t}^{n} W^k[f] \cdot k \cdot (1-2\epsilon)^{k-1} \\
&\leqslant t \cdot W^{\geqslant t}[f] \cdot (1-2\epsilon)^{t-1} + \sum_{k \geqslant t+1} (1-2\epsilon)^{k-1} W^{\geqslant k}[f] \\
&\leqslant \frac{2}{1-e^{-2}} \left[t \cdot \mathsf{NS}_{\frac{1}{\mathsf{t}}}(f) + \sum_{k \geqslant t+1} (1-2\epsilon)^{k-1} \mathsf{NS}_{\frac{1}{\mathsf{k}}}(f) \right] \\
&\leqslant \frac{2}{1-e^{-2}} \left[t \cdot \alpha \cdot t^{-\beta} + \sum_{k \geqslant t+1} (1-2\epsilon)^{k-1} \cdot \alpha \cdot k^{-\beta} \right] \\
&\leqslant \frac{2}{1-e^{-2}} \left[\alpha \cdot t^{1-\beta} + \alpha \cdot (t+1)^{-\beta} \sum_{k \geqslant t+1} (1-2\epsilon)^{k-1} \right] \\
&\leqslant \frac{2}{1-e^{-2}} \left[\alpha \cdot t^{1-\beta} + \alpha \cdot (t+1)^{-\beta} \cdot \frac{(1-2\epsilon)^t}{2\epsilon} \right] \\
&\leqslant \frac{3}{1-e^{-2}} \cdot \alpha \cdot (1/\epsilon)^{1-\beta}. \qquad (16)
\end{aligned}
$$

Using Eq. (15) and Eq. (16), in Eq. (14), we obtain the claimed bound in the lemma. □

From [13] we have the following bound on entropy.

Lemma 4.3. *[13] Let $f : \{0,1\}^n \to \{+1,-1\}$ be a Boolean function. Then,*

$$\mathbb{H}(f) \leqslant (3 + \log_2 e) \cdot \mathsf{Inf}[f] + \log_2 e \cdot \sum_{k=1}^{n} W^k[f] k \ln \frac{n}{k}.$$

This lemma suggests that one way to prove a non-trivial upper bound on Fourier entropy is to bound the second summand on the right in a general way. Using Lemma 4.2, we prove another technical lemma that provides a bound on $\sum_{k=1}^{n} W^k[f] k \ln \frac{n}{k}$. The first few steps in the proof below are the same as in [13].

Lemma 4.4. *Let $f : \{0,1\}^n \to \{+1,-1\}$ be a Boolean function. Then,*

$$\sum_{k=1}^{n} W^k[f] k \ln \frac{n}{k} \leqslant exp(1/2) \cdot \frac{5}{1-e^{-2}} \cdot \frac{\alpha}{(1-\beta)^2} \cdot (4n)^{1-\beta} \quad .$$

Proof.

$$\begin{aligned}
\sum_{k=1}^{n} W^k[f]k \ln \frac{n}{k} &\leqslant \sum_{k=1}^{n} W^k[f]k \cdot \sum_{j=k}^{n} \frac{1}{j} \leqslant \sum_{j=1}^{n} \frac{1}{j} \sum_{k=1}^{j} W^k[f]k \\
&\leqslant \sum_{j=1}^{n} \frac{1}{j} \sum_{k=1}^{j} W^k[f]k \cdot \exp(1/2)(1-\frac{1}{2j})^{k-1}, \\
&\qquad \left[\text{since } \exp(1/2)(1-\tfrac{1}{2j})^m \geqslant 1, \forall m \leqslant (j-1)\right] \\
&\leqslant \sum_{j=1}^{n} \frac{1}{j} \cdot \exp(1/2) \cdot \mathsf{NS}'_{\frac{1}{4j}}(f) \\
&\leqslant exp(1/2) \cdot \frac{5}{1-e^{-2}} \cdot \frac{\alpha}{1-\beta} \cdot \sum_{j=1}^{n} \frac{1}{j} \cdot (4j)^{1-\beta} \\
&\leqslant exp(1/2) \cdot \frac{5}{1-e^{-2}} \cdot \frac{\alpha}{1-\beta} \cdot 4^{1-\beta} \cdot \sum_{j=1}^{n} j^{-\beta} \\
&\leqslant exp(1/2) \cdot \frac{5}{1-e^{-2}} \cdot \frac{\alpha}{(1-\beta)^2} \cdot 4^{1-\beta} \cdot n^{1-\beta}. \qquad (17)
\end{aligned}$$

□

Using Lemma 4.4 and Lemma 4.3, we obtain the following theorem which bounds the Fourier entropy of a Boolean function.

Theorem 4.5. *Let $f : \{0,1\}^n \to \{+1,-1\}$ be a Boolean function such that $\mathsf{NS}_\epsilon(f) \leqslant \alpha \cdot \epsilon^\beta$. Then*

$$\mathbb{H}(f) \leqslant C \cdot \left(\mathsf{Inf}[f] + \frac{\alpha}{(1-\beta)^2} \cdot (4n)^{1-\beta}\right) \quad ,$$

where C is a universal constant.

In particular, for polynomial threshold functions, there exist non-trivial bounds on their noise sensitivity.

Theorem 4.6 (Peres's Theorem)**.** *[17] Let $f : \{0,1\}^n \to \{+1,-1\}$ be a linear threshold function. Then $\mathsf{NS}_\epsilon(f) \leqslant O(\sqrt{\epsilon})$.*

Theorem 4.7. *[18] For any degree-d polynomial threshold function $f : \{0,1\}^n \to \{+1,-1\}$ and $0 < \epsilon < 1$, $\mathsf{NS}_\epsilon(f) \leqslant 2^{O(d)} \cdot \epsilon^{1/(4d+6)}$.*

As corollaries of Theorem 4.5, using Theorem 4.6 and Theorem 4.7, we obtain the following bounds on the Fourier entropy of polynomial threshold functions.

Corollary 4.8. *Let $f : \{0,1\}^n \to \{+1,-1\}$ be a linear threshold function. Then, $\mathbb{H}(f) \leqslant C \cdot \sqrt{n}$, where C is a universal constant.*

Corollary 4.9. *Let $f : \{0,1\}^n \to \{+1,-1\}$ be a degree-d polynomial threshold function. Then, $\mathbb{H}(f) \leqslant C \cdot 2^{O(d)} \cdot n^{1-\frac{1}{4d+6}}$, where C is a universal constant.*

5 Entropy-Influence Inequality for Read-Once Formulas

In this section, we will prove the Fourier Entropy-Influence conjecture for read-once formulas using AND, OR, XOR, and NOT gates. We note that a concurrent (with the conference version of this paper) and independent result of O'Donnell and Tan [1] proves the conjecture for read-once formulas with *arbitrary* gates of bounded fan-in. But since our proof is completely elementary and very different from theirs, we choose to present it here.

It is well-known that both Fourier entropy and average sensitivity add up when two functions on disjoint sets of variables are added modulo 2.

Fact 5.1. *Let $f = g_1 \oplus g_2$ for $g_i : \{0,1\}^{V_i} \to \{-1,+1\}$, where $V_1 \cap V_2 = \emptyset$. Let $V = V_1 \cup V_2$. Then,*

1. $\mathbb{H}(f) = \mathbb{H}(g_1) + \mathbb{H}(g_2)$
2. $\mathsf{as}(f) = \mathsf{as}(g_1) + \mathsf{as}(g_2)$.

Our main result here is to show that somewhat analogous "tensorizability" properties hold when composing functions on disjoint sets of variables using AND and OR operations.

For $f : \{0,1\}^n \to \{+1,-1\}$, let $f_{\mathbb{B}}$ denote its 0-1 counterpart: $f_{\mathbb{B}} \equiv \frac{1-f}{2}$. Let's define the following 0-1 variant of $\mathbb{H}$:

$$\mathbf{H}(f_{\mathbb{B}}) := \sum_S \widehat{f_{\mathbb{B}}}(S)^2 \log \frac{1}{\widehat{f_{\mathbb{B}}}(S)^2}. \tag{18}$$

An easy relation enables translation between $\mathbb{H}(f)$ and $\mathbf{H}(f_{\mathbb{B}})$:

Lemma 5.2. *Let $p = \Pr[f_{\mathbb{B}} = 1] = \widehat{f_{\mathbb{B}}}(\emptyset) = \sum_S \widehat{f_{\mathbb{B}}}(S)^2$ and $q := 1 - p$. Then,*

$$\mathbb{H}(f) = 4 \cdot \mathbf{H}(f_{\mathbb{B}}) + \varphi(p), \quad \textit{where} \tag{19}$$

$$\varphi(p) := \mathrm{H}(4pq) - 4p(\mathrm{H}(p) - \log p). \tag{20}$$

Now, let $f = \mathsf{AND}(g_1, g_2)$ for $g_i : \{0,1\}^{V_i} \to \{-1,+1\}$, where $V_1 \cap V_2 = \emptyset$. Let $V = V_1 \cup V_2$. Let $g_{i\mathbb{B}} \equiv \frac{1-g_i}{2}$ and $p_i = \widehat{g_{i\mathbb{B}}}(\emptyset)$. It is then obvious that $f_{\mathbb{B}} \equiv g_{1\mathbb{B}} \cdot g_{2\mathbb{B}}$.

Lemma 5.3. *With the above notations, the following identities hold:*

1. *For all $S \subseteq V$, $\widehat{f_{\mathbb{B}}}(S) = \widehat{g_{1\mathbb{B}}}(S \cap V_1) \cdot \widehat{g_{2\mathbb{B}}}(S \cap V_2)$*
2. $\mathbf{H}(f_{\mathbb{B}}) = p_2 \cdot \mathbf{H}(g_{1\mathbb{B}}) + p_1 \cdot \mathbf{H}(g_{2\mathbb{B}})$
3. $\mathsf{as}(f) = p_2 \cdot \mathsf{as}(g_1) + p_1 \cdot \mathsf{as}(g_2)$.

Proof of this lemma follows by direct computation and is omitted.

$$\text{For } 0 \leqslant p \leqslant 1, \text{ let's also define:} \quad \psi(p) := p^2 \log \frac{1}{p^2} - 2\,\mathrm{H}(p). \tag{21}$$

▶ **Intuition:** Before going on, we pause to give some intuition about the choice of the function ψ and the function κ below in Eq. (24). In the FEI conjecture (Eq. (2)), the right hand side, $\mathsf{Inf}(f)$, does not depend on whether we take the range of f to be $\{-1,+1\}$ or $\{0,1\}$. In contrast, the left hand side, $\mathbb{H}(f)$, depends on the range being $\{-1,+1\}$. Just as the usual entropy-influence inequality composes w.r.t. the parity operation (Fact 5.1) with $\{-1,+1\}$ range, we expect a corresponding composition with $\{0,1\}$ range to hold for the AND operation (and by symmetry for the OR operation). However, Lemma 5.2 shows the translation to $\{0,1\}$-valued functions results in the annoying additive "error" term $\varphi(p)$. Such additive terms that depend on p create technical difficulties in the inductive proofs below and we need to choose the appropriate functions of p carefully.

For example, we know $4\,\mathbf{H}(f_{\mathbb{B}}) + \varphi(p) = \mathbb{H}(f) = 4\,\mathbf{H}(1 - f_{\mathbb{B}}) + \varphi(q)$ from Lemma 5.2. If the conjectured inequality for the $\{0,1\}$-valued entropy-influence inequality has an additive error term $\psi(p)$ (see Eq. (22) below), then we must have $\mathbf{H}(f_{\mathbb{B}}) - \mathbf{H}(1 - f_{\mathbb{B}}) = \psi(p) - \psi(q) = (\varphi(q) - \varphi(p))/4 = p^2 \log \frac{1}{p^2} - q^2 \log \frac{1}{q^2}$, using Eq. (20). Hence, we may conjecture that $\psi(p) = p^2 \log \frac{1}{p^2} +$ (an additive term symmetric w.r.t. p and q). Given this and the other required properties, e.g., Lemma 5.4 below, for the composition to go through, we are lead to the definition of ψ in Eq. (21). Similar considerations w.r.t. composition by parity operation (in addition to those by AND, OR, and NOT) leads us to the definition of κ in Eq. (24). ◀

Let us define the **FEI01 Inequality** (the 0-1 version of FEI) as follows:

$$\mathbf{H}(f_{\mathbb{B}}) \leqslant c \cdot \mathsf{as}(f) + \psi(p), \tag{22}$$

where $p = \widehat{f_{\mathbb{B}}}(\emptyset) = \Pr_x[f_{\mathbb{B}}(x) = 1]$ and c is a constant to be fixed later.

The following technical lemma gives us the crucial property of ψ:

Lemma 5.4. *For ψ as in Eq.* (21) *and $p_1, p_2 \in [0,1]$, $p_1\psi(p_2) + p_2\psi(p_1) \leqslant \psi(p_1p_2)$.*

Since the proof of the lemma is somewhat technical, we move the proof to B. Given this lemma, an inductive proof yields our theorem for read-once formulas over the complete basis of $\{\mathsf{AND}, \mathsf{OR}, \mathsf{NOT}\}$. We now complete the proof.

Lemma 5.5. *Suppose $f_{\mathbb{B}} = \mathsf{AND}(g_{1\mathbb{B}}, g_{2\mathbb{B}})$, where the g_i depend on* disjoint *sets of variables. If each of the g_i satisfies the FEI01 Inequality* (22)*, then so does f.*

Proof.

$$\begin{aligned}
\mathbf{H}(f_{\mathbb{B}}) &= p_2\,\mathbf{H}(g_{1\mathbb{B}}) + p_1\,\mathbf{H}(g_{2\mathbb{B}}) \quad \text{by Lemma 5.3(2)} \\
&\leqslant p_2(c\cdot \mathsf{as}(g_1) + \psi(p_1)) + p_1(c\cdot \mathsf{as}(g_2) + \psi(p_2)) \quad \text{since } g_i \text{ satisfy Eq. (22)} \\
&= c\cdot(p_2\mathsf{as}(g_1) + p_1\mathsf{as}(g_2)) + (p_2\psi(p_1) + p_1\psi(p_2) \\
&\leqslant c\cdot \mathsf{as}(f) + \psi(p) \quad \text{by Lemma 5.3(3) and Lemma 5.4}
\end{aligned}$$

□

Lemma 5.6. *If f satisfies FEI01 inequality* (22)*, then so does its negation, i.e., $1-f$.*

Proof. Note that $\mathbf{H}(1-f) = \mathbf{H}(f) - p^2\log\frac{1}{p^2} + q^2\log\frac{1}{q^2}$ and because $\mathrm{H}(p) = \mathrm{H}(q)$, $\psi(p) - \psi(q) = p^2\log\frac{1}{p^2} - q^2\log\frac{1}{q^2}$. □

Corollary 5.7. *Suppose $f_{\mathbb{B}} = \mathsf{OR}(g_{1\mathbb{B}}, g_{2\mathbb{B}})$, where the g_i depend on* disjoint *sets of variables. If each of the g_i satisfies the FEI01 Inequality* (22)*, then so does f.*

Proof. Note that $1 - f_{\mathbb{B}} = (1 - g_{1\mathbb{B}}) \cdot (1 - g_{2\mathbb{B}})$ and apply lemmas 5.5 and 5.6. □

Theorem 5.8. *The FEI01 inequality* (22) *holds for all read-once Boolean formulas using* AND*,* OR*, and* NOT *gates, with constant $c = 5/2$.*

Proof. Let f be computed by a read-once Boolean formula. We assume without loss of generality that negations only appear at the bottom with leaves. We proceed by induction on the underlying tree. At the leaves f is a literal associated with a single variable, say x_1. Then, since $f_{\mathbb{B}}(\emptyset) = 1/2$ and $f_{\mathbb{B}}(\{1\}) = -1/2$, we calculate $\mathbf{H}(f_{\mathbb{B}}) = \frac{1}{4}\log 4 + \frac{1}{4}\log 4 = 1$, $\mathsf{as}(f) = 1$, $p = 1/2$, and $\psi(1/2) = -3/2$. Thus with $c = 5/2$, Eq. (22) is satisfied.

Now, Lemma 5.5 and Corollary 5.7 imply that at every AND gate and OR gate, the inequality (22) is preserved, i.e., if it holds at both the inputs, it also holds at the output. □

We now proceed to show that the above result can be extended to read-once formulas that include XOR gates as well. To switch to the usual FEI inequality (in the $\{-1,+1\}$ notation), we combine Eq. (22) and Eq. (19) to obtain

$$\mathbb{H}(f) \leqslant 10 \cdot \mathsf{as}(f) + \kappa(p), \quad \text{where} \tag{23}$$
$$\kappa(p) := 4\psi(p) + \varphi(p) = -8\,\mathrm{H}(p) - 8pq - (1-4pq)\log(1-4pq). \tag{24}$$

Since it uses the $\{-1,+1\}$ range, we expect that Eq. (23) should be preserved by parity composition of functions. The only technical detail is to show that the function κ also behaves well w.r.t. parity composition. We show that this indeed happens. Consider $f \equiv g_1 \oplus g_2$. Since parity is a simple product over $\{-1,+1\}$ range we have $f = g_1 \cdot g_2$, and therefore, $p = p_1q_2 + p_2q_1$. Thus we only need to show

Lemma 5.9. *For κ as defined by Eq.* (24)*, $\kappa(p_1) + \kappa(p_2) \leqslant \kappa(p_1q_2 + p_2q_1)$.*

Again, due to the technical nature of the proof, we move it to B.

We can now prove the following composition lemma which leads us to the main theorem of this section.

Lemma 5.10. *Suppose $f = g_1 \cdot g_2$, where the g_i depend on* disjoint *sets of variables. If each of the g_i satisfies the entropy-influence inequality* (23)*, then so does f.*

Proof.

$$\begin{aligned}
\mathbb{H}(f) &= \mathbb{H}(g_1) + \mathbb{H}(g_2) && \text{by Fact 5.1(i)}\\
&\leqslant 10 \cdot \mathsf{as}(g_1) + \kappa(p_1) + 10 \cdot \mathsf{as}(g_2) + \kappa(g_2) && \text{since } g_i \text{ satisfy Eq. (23)}\\
&= 10 \cdot \mathsf{as}(f) + \kappa(p_1) + \kappa(p_2) && \text{by Fact 5.1(ii)}\\
&\leqslant 10 \cdot \mathsf{as}(f) + \kappa(p) && \text{by Lemma 5.9.}
\end{aligned}$$

□

Theorem 5.11. *If f is computed by a read-once formula using* AND*,* OR*,* XOR*, and* NOT *gates, then $\mathbb{H}(f) \leqslant 10\,\mathsf{Inf}(f) + \kappa(p)$.*

Proof. We use induction on the tree given by the formula computing f to prove Eq. (23). Without loss of generality we assume that negations are only at the bottom with leaves. So the leaves are input variables or their negations and the claim that they satisfy Eq. (23) can be verified by direct calculation. At any internal node, its two inputs are given by subformulas depending on disjoint sets of variables by the read-once property of the formula. When the internal node is an AND or OR gate, the claim follows from Eq. (19), Lemma 5.5, Corollary 5.7, and Eq. (24). When the internal node is an XOR gate, the claim follows from Lemma 5.10. Thus Eq. (23) holds at the root of the tree and hence for f. $\square$

Remark 5.1. *The parity function on n variables shows that the bound in Theorem 5.11 is tight; it is not tight without the additive term $\kappa(p)$. It is easy to verify that $-10 \leqslant \kappa(p) \leqslant 0$ for $p \in [0,1]$. Hence the theorem implies $\mathbb{H}(f) \leqslant 10\,\mathsf{Inf}(f)$ for all read-once formulas f using* AND*,* OR*,* XOR*, and* NOT *gates.*

6 A Bound for Real valued Functions via Second Moment

An equivalent way to state the Fourier-entropy Influence conjecture is: *there exists a universal constant C such that for all $f : \{0,1\}^n \to \{+1,-1\}$,*

$$\mathbb{H}(f) \leqslant C \cdot \sum_{S \subseteq [n]} |S| \widehat{f}(S)^2.$$

In this section, we relax the Boolean-ness condition on f, and consider *real* valued functions $f : \{0,1\}^n \to \mathbb{R}$ defined over Boolean hypercube. We obtain a nontrivial bound on the entropy (suitably defined) of such functions. In particular, we show that for all f such that $\sum_S \widehat{f}(S)^2 = 1$, and for all $\delta \in (0,1]$,

$$\mathbb{H}(f) \leqslant \sum_S |S|^{1+\delta} \widehat{f}(S)^2 + (\log n)^{O(\frac{1}{\delta})}.$$

Our proof below uses the following lemma:

Lemma 6.1. *For any t, let $\mathcal{T} \subseteq \{S \mid |\widehat{f}(S)| \leqslant 1/t\}$. Suppose $|\mathcal{T}| \leqslant t$. Then,*

$$\sum_{S \in \mathcal{T}} \widehat{f}(S)^2 \log\left(\frac{1}{\widehat{f}(S)^2}\right) \leqslant 2.$$

Furthermore, for any k,

$$\sum_{S : |S| \leqslant k} \widehat{f}(S)^2 \log\left(\frac{1}{\widehat{f}(S)^2}\right) \leqslant 2 + 2k \log n.$$

Proof. We will prove the second part of the lemma since that includes proof of the first part. First note that the number of summands in the second part is at most n^k. Let $S_k := \{S \mid |\widehat{f}(S)| < 1/n^k\}$, then

$$\sum_{S\in S_k} \widehat{f}(S)^2 \log\left(\frac{1}{\widehat{f}(S)^2}\right) \leqslant \frac{2}{n^k}\sum_{S\in S_k} |\widehat{f}(S)| \log\left(\frac{1}{|\widehat{f}(S)|}\right) \leqslant 2,$$

where the last inequality follows from the fact that $|\widehat{f}(S)|\log(1/|\widehat{f}(S)|) < 1$, since $x\log(1/x) < 1$, for all $0 \leqslant x \leqslant 1$.

Now for all S such that $|S| \leqslant k$ and $S \notin S_k$, $\log(1/|\hat{f}(S)|) \leqslant k\log n$. Hence,

$$\sum_{S:|S|\leqslant k \text{ and } S\notin S_k} \widehat{f}(S)^2 \log\left(\frac{1}{\widehat{f}(S)^2}\right) \leqslant 2k\log n.$$

□

We now state and prove the main theorem of this section.

Theorem 6.2. *If $f = \sum_{S\subseteq[n]} \widehat{f}(S)\chi_S$ is a real-valued function on the domain $\{0,1\}^n$ such that $\sum_S \widehat{f}(S)^2 = 1$, then, for any $\delta \in (0,1]$,*

$$\sum_{S\subseteq[n]} \widehat{f}(S)^2 \log\left(\frac{1}{\widehat{f}(S)^2}\right) \leqslant \sum_{S\subseteq[n]} |S|^{1+\delta}\widehat{f}(S)^2 + 2(2\log n)^{\frac{1+\delta}{\delta}} + O(\log\log n/\log(1+\delta)) \ .$$

Proof. Since the proof consists of careful counting, we highlight our proof strategy first. We partition the Fourier coefficients into suitable parts and then upper bound each part. We start with suitably chosen sets $A_0, B_0 \subseteq 2^{[n]}$ and then inductively construct the sets $A_1, B_1, \ldots, A_k, B_k$. The A_i's represent the new Fourier coefficients whose total entropy we are able to upper bound. The B_i's represent the Fourier coefficients that are *not yet accounted for.* Our construction yields that as k increases B_k only consists of those $\widehat{f}(S)$ for which $|S| < \psi(k, n, \delta)$, where ψ is a suitable function of k, n and δ. Finally an appropriate choice of k gives us the desired inequality.

Following this strategy, we start by describing the sets A_i and B_i.

Let A_0 be the set of all $S \subseteq [n]$ for which $|S|^{1+\delta}$ is at least $\log\left(\frac{1}{\widehat{f}(S)^2}\right)$. That is,

$$A_0 := \{S \mid \widehat{f}(S)^2 \geqslant 1/2^{|S|^{1+\delta}}\}.$$

Clearly,

$$\sum_{S\in A_0} \widehat{f}(S)^2 \log\left(\frac{1}{\widehat{f}(S)^2}\right) \leqslant \sum_{S\in A_0} |S|^{1+\delta}\widehat{f}(S)^2. \tag{25}$$

Now, let A_1 be all the $S \subseteq [n]$ for which $|\widehat{f}(S)| < 2^{-n}$. Since $|A_1|$ is clearly at most 2^n, Lemma 6.1 above applies and we conclude that

$$\sum_{S \in A_1} \widehat{f}(S)^2 \log\left(\frac{1}{\widehat{f}(S)^2}\right) \leqslant 2. \tag{26}$$

Further let $B_1 = \{0,1\}^n \setminus (A_0 \cup A_1)$. By the definition of A_0 and A_1,

$$B_1 \subseteq \left\{ S \mid \frac{1}{2^{2n}} \leqslant \widehat{f}(S)^2 \leqslant \frac{1}{2^{|S|^{1+\delta}}} \right\}.$$

It follows that $B_1 \subseteq \{S \mid |S| \leqslant (2n)^{1/(1+\delta)}\}$. Let $r_1 := (2n)^{1/(1+\delta)}$. Thus, $|B_1| \leqslant \sum_{i=0}^{r_1} \binom{n}{i} < n^{r_1}$.

Next, let $A_2 := \{S \in B_1 : |\widehat{f}(S)| \leqslant 1/n^{r_1}\}$ and $B_2 := B_1 \setminus A_2$.

First, note that, since $A_2 \subseteq B_1$ and $|B_1| \leqslant n^{r_1}$, Lemma 6.1 can be applied to A_2 and hence the contribution of coefficients from A_2 is at most 2.

We also have,

$$B_2 \subseteq \left\{ S \mid \frac{1}{n^{2r_1}} \leqslant \widehat{f}(S)^2 \leqslant \frac{1}{2^{|S|^{1+\delta}}} \right\}.$$

Let $r_2 = (\log(n^{2r_1}))^{1/(1+\delta)} = (2r_1 \log n)^{1/(1+\delta)}$. It is then clear that for $S \in B_2$, we must have $|S| \leqslant r_2$ and thus $|B_2| \leqslant n^{r_2}$.

Continuing this way, we define

$$\begin{aligned}
r_{k+1} &:= (2r_k \log n)^{1/(1+\delta)}, \\
A_{k+1} &:= \{S \in B_k \mid |\widehat{f}(S)| \leqslant 1/n^{r_k}\}, \text{ and} \\
B_{k+1} &:= B_k \setminus A_{k+1}.
\end{aligned}$$

In general, then,

$$B_{k+1} \subseteq \left\{ S \mid \frac{1}{n^{2r_k}} \leqslant \widehat{f}(S)^2 \leqslant \frac{1}{2^{|S|^{1+\delta}}} \right\}.$$

Thus $B_{k+1} \subseteq \{S \mid |S| \leqslant r_{k+1}\}$, and so, $|B_{k+1}| \leqslant n^{r_{k+1}}$. Since $A_{k+1} \subseteq B_k$, $|A_{k+1}| \leqslant n^{r_k}$ and Lemma 6.1 can be applied to A_{k+1}.

It is easy to see by induction that for $k \geqslant 1$,

$$r_k = (2\log n)^{\frac{1}{\delta}\left(1-(1+\delta)^{-k+1}\right)} \cdot (2n)^{(1+\delta)^{-k}}.$$

Thus, $r_k \leqslant (2\log n)^{\frac{1}{\delta}} \cdot (2n)^{(1+\delta)^{-k}}$.

By taking $k^* := \log\log 2n/\log(1+\delta)$, we get $r_{k^*} \leqslant 2(2\log n)^{\frac{1}{\delta}}$.

We repeat the above process up to k^* times. For each $k \leqslant k^*$, the coefficients from A_k contribute at most 2 to the entropy by the first part of the lemma. Note that for all sets $S \in B_{k^*}$, $|S| \leqslant r_{k^*}$. For $k = k^*$, we apply the second part of proof of Lemma 6.1 and conclude that coefficients from B_{k^*} contribute at most $2r_{k^*} \log n \leqslant 2 \cdot (2 \log n)^{1+\frac{1}{\delta}}$. Moreover, note that $A_0 \cup A_1 \cup \cdots \cup A_{k^*} \cup B_{k^*}$ is a cover of $2^{[n]}$. Hence, we accounted for contributions to the entropy from all coefficients.

Altogether, we get the total entropy to be at most

$$\sum_{S\subseteq[n]} \widehat{f}(S)^2 \log\left(\frac{1}{\widehat{f}(S)^2}\right) \leqslant \sum_{S\subseteq[n]} |S|^{1+\delta}\widehat{f}(S)^2 + 2\log\log 2n/\log(1+\delta) + 2(2\log n)^{1+\frac{1}{\delta}}.$$

□

As a corollary to Theorem 6.2, we obtain an upper bound (cf. Eq. (7)) on the Fourier Entropy of a real-valued function in terms of the first and second moments of the sensitivities of the function.

Corollary 6.3. *If $f = \sum_{S\subseteq[n]} \widehat{f}(S)\chi_S$ is a real-valued function on the domain $\{0,1\}^n$ such that $\sum_S |\widehat{f}(S)^2| = 1$, then, for any $\delta \in (0,1]$,*

$$\sum_{S\subseteq[n]} \widehat{f}(S)^2 \log\left(\frac{1}{\widehat{f}(S)^2}\right) = \mathsf{as}(f)^{1-\delta}\mathsf{as}_2(f)^{\delta} + 2(2\log n)^{\frac{1+\delta}{\delta}} + O(\log\log n/\log(1+\delta)),$$

where $\mathsf{as}_2(f) := \sum_S |S|^2 \widehat{f}(S)^2$.

Note that in the above statements $\mathsf{as}(f)$ is defined via its Fourier expansion, that is, $\mathsf{as}(f) := \sum_S |S|\widehat{f}(S)^2$. Similarly, $\mathsf{as}_2(f)$, in spite of having a combinatorial definition (see Lemma 6.5), is defined to be $\sum_S |S|^2\widehat{f}(S)^2$.

The proof of Corollary 6.3 is straightforward from the following lemma.

Lemma 6.4. *Let $f : \{0,1\}^n \to \mathbb{R}$, and $0 \leqslant \delta \leqslant 1$. Then*

$$\mathsf{as}(f)^{1-\delta}\mathsf{as}_2(f)^{\delta} \geqslant \sum_{S\subseteq[n]} |S|^{1+\delta}\widehat{f}(S)^2.$$

But for proving Lemma 6.4, we require Lemma 6.5 below. Lemma 6.5 seems to be well-known; see for instance, [27, Eq. 2.11] or [2, Ex. 2.20]. For completeness, we give here a proof by Alex Samorodnitsky (personal communication).

Lemma 6.5. *For* $f : \{0,1\}^n \to \{+1,-1\}$,

$$\frac{1}{2^n}\sum_x \mathsf{s}_f(x)^2 = \sum_{S\subseteq[n]} |S|^2\widehat{f}(S)^2 = \mathsf{as}_2(f).$$

Proof. Consider the following function $L : \{0,1\}^n \to \mathbb{R}$,

$$L(x) = \begin{cases} n & \text{for} \quad |x| = 0. \\ -1 & \text{for} \quad |x| = 1. \\ 0 & \text{for} \quad |x| > 1. \end{cases}$$

Consider the convolution $L * f(x) := \frac{1}{2^n}\sum_z f(x \oplus z)L(z)$. It is easy to see that

$$L * f(x) = \frac{2\mathsf{s}_f(x)f(x)}{2^n}.$$

Using Parseval's identity we obtain:

$$\frac{1}{2^n}\sum_x (2\mathsf{s}_f(x)/2^n)^2 = \sum_{S\subseteq[n]} \widehat{L * f}(S)^2 = \sum_{S\subseteq[n]} \widehat{L}(S)^2\widehat{f}(S)^2.$$

It is also easy to see that for any $S \subseteq [n]$, $\widehat{L}(S) = 2|S|/2^n$. So we obtain

$$\frac{1}{2^n}\sum_x 4\mathsf{s}_f(x)^2 = \sum_{S\subseteq[n]} 4|S|^2\widehat{f}(S)^2.$$

This completes the proof. □

We now prove Lemma 6.4.

Proof. (of Lemma 6.4) For $\delta = 0$, this is the Fourier expression for average sensitivity. For $\delta = 1$, this is Lemma 6.5. We next prove it for $\delta = 1/2$. We treat $\widehat{f}(S)^2$ as the probability associated to the set S and use the following version of the Cauchy-Schwartz inequality: for any two random variables $X, Y : \Omega \to \mathbb{R}$, we have $\sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)} \geqslant \mathbb{E}(XY)$. Choosing $X(S) = \sqrt{|S|}$ and $Y(S) = |S|$ immediately yields the desired inequality for the value of $\delta = \frac{1}{2}$ in light of Lemma 6.5.

In general, we can show the following: if the desired inequality holds for $\delta = \alpha$ and $\delta = \beta$ then the inequality must also hold for $\delta = \frac{\alpha+\beta}{2}$. To show this, one may apply the Cauchy-Schwartz inequality with $X(S) = |S|^{(1+\alpha)/2}$ and $Y(S) = |S|^{(1+\beta)/2}$.

Hence, by continuity, the desired inequality holds for any $\delta \in [0,1]$. □

Acknowledgements. N. Saurabh thanks Ryan O'Donnell for helpful discussions on Lemma 4.2.

References

References

[1] R. O'Donnell, L.-Y. Tan, A composition theorem for the Fourier entropy-influence conjecture, in: Proceedings of Automata, Languages and Programming - 40th International Colloquium, 2013, pp. 780–791.

[2] R. O'Donnell, Analysis of Boolean Functions, Cambridge University Press, 2014.

[3] E. Friedgut, G. Kalai, Every monotone graph property has a sharp threshold, Proceedings of the American Mathematical Society 124 (10) (1996) 2993–3002.

[4] M. Ben-Or, N. Linial, Collective coin flipping, robust voting schemes and minima of Banzhaf values, in: Proceedings of the 26th Annual Symposium on Foundations of Computer Science, 1985, pp. 408–416.

[5] P. Erdös, A. Rényi, On the Evolution of Random Graphs, Publ. Math. Inst. Hung. Acad. Sci. 5 (1960) 17–61.

[6] L. Russo, On the critical percolation probabilities, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 56 (2) (1981) 229–237.

[7] G. A. Margulis, Probabilistic characteristics of graphs with large connectivity, Probl. Peredachi Inf. 10 (1974) 101–108.

[8] G. Kalai, The entropy/influence conjecture, Terence Tao's blog. URL `https://terrytao.wordpress.com/2007/08/16/gil-kalai-the-entropyinfluence-conjecture/`

[9] J. Bourgain, G. Kalai, Influences of variables and threshold intervals under group symmetries, Geometric and Functional Analysis (GAFA) 7 (3) (1997) 438–461.

[10] Y. Mansour, An $\mathcal{O}(\mathrm{n}^{\log \log n})$ learning algorithm for DNF under the uniform distribution, Journal of Computer and System Sciences 50 (3) (1995) 543–550.

[11] P. Gopalan, A. T. Kalai, A. Klivans, Agnostically learning decision trees, in: Proceedings of the 40th annual ACM symposium on Theory of computing, STOC '08, 2008, pp. 527–536.

[12] A. Klivans, H. Lee, A. Wan, Mansour's conjecture is true for random DNF formulas, in: Proceedings of the 23rd Conference on Learning Theory, 2010, pp. 368–380.

[13] R. O'Donnell, J. Wright, Y. Zhou, The Fourier entropy-influence conjecture for certain classes of Boolean functions, in: Proceedings of Automata, Languages and Programming - 38th International Colloquium, 2011, pp. 330–341.

[14] N. Keller, E. Mossel, T. Schlank, A note on the entropy/influence conjecture, Tech. rep., arXiv:1105.2651 (2011).

[15] A. Wan, J. Wright, C. Wu, Decision trees, protocols and the entropy-influence conjecture, in: Innovations in Theoretical Computer Science, ITCS'14, 2014, pp. 67–80.

[16] B. Das, M. Pal, V. Visavaliya, The entropy influence conjecture revisited, Tech. rep., arXiv:1110.4301 (2011).

[17] R. O'Donnell, Computational applications of noise sensitivity, Ph.D. thesis, MIT (2003).

[18] P. Harsha, A. Klivans, R. Meka, Bounding the sensitivity of polynomial threshold functions, Theory of Computing 10 (1) (2014) 1–26.

[19] I. Diakonikolas, P. Raghavendra, R. Servedio, L.-Y. Tan, Average sensitivity and noise sensitivity of polynomial threshold functions, SIAM Journal on Computing 43 (1) (2014) 231–253.

[20] J. Kahn, G. Kalai, N. Linial, The influence of variables on Boolean functions, in: Proceedings of the 29th Annual IEEE Symposium on Foundations of Computer Science, 1988, pp. 68–80.

[21] R. de Wolf, A brief introduction to Fourier analysis on the Boolean cube, Theory of Computing, Graduate Surveys 1 (2008) 1–20.

[22] I. Benjamini, G. Kalai, O. Schramm, Noise sensitivity of Boolean functions and applications to percolation, Publications Mathematiques de l'Institut des Hautes Etudes Scientifiques 90 (1) (1999) 5–43.

[23] E. Kushilevitz, Y. Mansour, Learning decision trees using the Fourier spectrum, in: Proceedings of the 23rd Annual ACM Symposium on Theory of Computing, 1991, pp. 455–464.

[24] Y. Mansour, Learning Boolean functions via the Fourier transform, in: V. Roychowdhury, K.-Y. Siu, A. Orlitsky (Eds.), Theoretical Advances in Neural Computation and Learning, Springer US, 1994, pp. 391–424.

[25] Y. Brandman, A. Orlitsky, J. Hennessy, A spectral lower bound technique for the size of decision trees and two-level and/or circuits, IEEE Transactions of Computers 39 (2) (1990) 282–287.

[26] C.-J. Lee, S. V. Lokam, S.-C. Tsai, M.-C. Yang, On restrictions of non-degenerate Boolean functions and lower bounds by the polynomial method, IEEE International Symposium on Information Theory, ISIT (2015) 501–505.

[27] C. Garban, G. Pete, O. Schramm, The Fourier spectrum of critical percolation, Acta Mathematica 205 (2010) 19–104. `arXiv:http://rd.springer.com/article/10.1007\%2Fs11511-010-0051-x`.

A Non-Boolean functions with large Fourier Entropy

A decision tree for a non-Boolean, say $\mathbb{R}$-valued, function f can be defined by a natural generalization of the one for a Boolean-valued function. It queries the (Boolean) input variables as in the usual decision tree, but produces a value in $\mathbb{R}$ at each leaf. It must guarantee that on all inputs that reach a leaf the function value must be constant and equal to the value produced at that leaf.

Our next example shows that Fourier entropy cannot be upper bounded by log(number of leaves) for non-Boolean f in contrast to Inequality (9) for Boolean functions. In fact, there is an exponential gap:

Lemma A.1. *There exists a function $f : \{0,1\}^n \to \mathbb{R}$ satisfying $\sum_S \widehat{f}(S)^2 = 1$ such that*

$$\sum_{S\subseteq[n]} \widehat{f}(S)^2 \log\left(\frac{1}{\widehat{f}(S)^2}\right) = \Omega(n), \quad \text{but} \quad \log L(f) = O(\log n).$$

Proof. Consider the following function:

$$f(x) = \sqrt{\frac{2^{d(x)}}{n+2}},$$

where $d(x) = n+1$, if $x = 0^n$, else it is the first index in x that is 1. Note that this function has a decision tree same as the OR function and thus have only $n+1$ leaves. Now to see that $\sum_{S\subseteq[n]} \widehat{f}(S)^2 = 1$ consider the following:

$$\sum_x f(x)^2 = \sum_{i\in[n+1]} \sum_{x:d(x)=i} f(x)^2 = \sum_{i\in[n]} 2^{n-i}\frac{2^i}{n+2} + \frac{2^{n+1}}{n+2} = 2^n,$$

and thus from Parseval's identity we have $\sum_{S\subseteq[n]} \widehat{f}(S)^2 = 1$.

It is easy to check that for any set $S \subseteq [n]$ if k is the largest index in S then

$$|\widehat{f}(S)| = \frac{1}{2^n}\left(2^{n-k}\sqrt{\frac{2^k}{n+2}} - \sum_{i=k+1}^{n} 2^{n-i}\sqrt{\frac{2^i}{n+2}} - \sqrt{\frac{2^{n+1}}{n+2}}\right) \approx \frac{1}{\sqrt{n2^k}}.$$

And from this it follows that the entropy for the Fourier coefficient squares is around $n/2 + \log n$ whereas $\log(L(f)) = \log(n)$. $\square$

Our next example shows that Fourier entropy can be logarithmically larger than the degree for non-Boolean functions in contrast to Inequality (9) for Boolean functions.

Lemma A.2. *There exists a function* $f : \{0,1\}^n \to \mathbb{R}$ *of degree* d *satisfying* $\sum_S \widehat{f}(S)^2 = 1$ *such that*

$$\sum_{S\subseteq[n]} \widehat{f}(S)^2 \log\left(\frac{1}{\widehat{f}(S)^2}\right) = \Omega(d \log n).$$

Proof. Consider the following function $f = \sum_{S\subseteq[n]} \widehat{f}(S)\chi_S$, where $\widehat{f}(S) = 1/\sqrt{\binom{n}{2}}$ if $|S| = 2$, and $\widehat{f}(S) = 0$ otherwise. It is easy to see that the $\mathbb{H}(f) = \log\binom{n}{2}$, whereas $\mathsf{Inf}(f) = \sum\limits_{S\subseteq[n]} |S|\widehat{f}(S)^2 = 2$.

So now if we put uniform weights on k-sized sets, that is, $\widehat{f}(S) = 1/\sqrt{\binom{n}{k}}$ if $|S| = k$, and $\widehat{f}(S) = 0$ if $|S| \neq k$, we will get $\mathsf{Inf}(f) = k$ and $\mathbb{H}(f) = \log\binom{n}{k} \geqslant k\log n - k\log k$. Choosing $k = \sqrt{n}$, we will have $\mathbb{H}(f) = \Omega(\sqrt{n}\log n)$ and $\mathsf{Inf}(f) = \sqrt{n}$. Since the degree of the function is $d = \sqrt{n}$, we get $\mathbb{H}(f) = \Omega(d \cdot \log n)$. $\square$

B Proofs from Section 5

Lemma 5.4 restated: For ψ as defined by Eq. (21) and $p_1, p_2 \in [0,1]$,

$$p_1 \cdot \psi(p_2) + p_2 \cdot \psi(p_1) \leqslant \psi(p_1 p_2).$$

Proof. We need to prove that $p_1\psi(p_2) + p_2\psi(p_1) - \psi(p_1p_2) \leqslant 0$.

Let's begin by manipulating the l.h.s.:

$$
\begin{aligned}
&p_1\psi(p_2) + p_2\psi(p_1) - \psi(p_1p_2)\\
&= p_1\left(p_2^2\log\frac{1}{p_2^2} - 2\,\mathrm{H}(p_2)\right) + p_2\left(p_1^2\log\frac{1}{p_1^2} - 2\,\mathrm{H}(p_1)\right) - (p_1p_2)^2\log\frac{1}{(p_1p_2)^2} + 2\,\mathrm{H}(p_1p_2)\\
&= 2p_1p_2\left(-p_2\log p_2 - p_1\log p_1 + p_1p_2\log p_2 + p_1p_2\log p_1\right) + 2\left(\mathrm{H}(p_1p_2) - p_2\,\mathrm{H}(p_1) - p_1\,\mathrm{H}(p_2)\right)\\
&= 2p_1p_2\left(-p_2q_1\log p_2 - p_1q_2\log p_1\right) + 2\left(-(1-p_1p_2)\log(1-p_1p_2) + p_2q_1\log q_1 + p_1q_2\log q_2\right)\\
&= 2p_1q_2\left(-p_1p_2\log p_1 + \log q_2\right) + 2p_2q_1\left(-p_1p_2\log p_2 + \log q_1\right) - 2(1-p_1p_2)\log(1-p_1p_2)\\
&\leqslant 2(1-p_1p_2)\left(-p_1p_2\log(p_1p_2) + \log\frac{q_1q_2}{1-p_1p_2}\right) \quad \text{since } p_1q_2, p_2q_1 \leqslant (1-p_1p_2)\\
&\leqslant 2(1-p_1p_2)\left(-p_1p_2\log(p_1p_2) + \log\frac{(1-\sqrt{p_1p_2})^2}{(1-p_1p_2)}\right)\\
&\qquad \text{since } q_1q_2 = (1-p_1)(1-p_2) \leqslant (1-\sqrt{p_1p_2})^2,\\
&\qquad \text{e.g., by the AM-GM inequality } p_1+p_2 \geqslant 2\sqrt{p_1p_2}.
\end{aligned}
$$

Since $p_1p_2 \in [0,1]$, it suffices to show the (univariate) inequality $\tau(x) := -x\ln x + \ln\frac{(1-\sqrt{x})^2}{1-x} \leqslant 0$ for $x \in [0,1]$. Since the boundary cases are easy to verify, it suffices to prove the that $\tau(x) \leqslant 0$ for $x \in (0,1)$. Note that $\tau(0) = 0$ and hence it suffices to prove that $\tau'(x) < 0$ for $x \in (0,1)$. But

$$
\begin{aligned}
\tau'(x) &= -1 + \ln\frac{1}{x} - \frac{1}{\sqrt{x}(1-x)}\\
&\leqslant -1 + \sqrt{\frac{1}{x}} - \frac{1}{\sqrt{x}(1-x)} \quad \text{since } \ln y \leqslant \sqrt{y}\\
&= -1 - \frac{\sqrt{x}}{1-x}\\
&< 0 \quad \text{for } x \in (0,1).
\end{aligned}
$$

□

Lemma 5.9 restated: For κ as defined by Eq. (24) and $p_1, p_2 \in [0,1]$,

$$\kappa(p_1) + \kappa(p_2) \leqslant \kappa(p_1q_2 + p_2q_1).$$

Proof. In the following, we will let $p = p_1q_2 + p_2q_1$, and $q = 1-p = p_1p_2 + q_1q_2$.

To begin with, we observe that $(1-4pq) = (p-q)^2$ and that $(p-q) = (p_1-q_1)(p_2-q_2)$, i.e., parity operation on independent Boolean variables results in multiplying their *biases*, and hence $(1-4pq) = (1-4p_1q_1)(1-4p_2q_2)$.

Using this, we relate the third terms on either side of the inequality to be proved.

$$\begin{aligned}(1-4pq)\log(1-4pq) &= (1-4p_1q_1)(1-4p_2q_2)\log((1-4p_1q_1)(1-4p_2q_2)) \\ &= (1-4p_2q_2)\left((1-4p_1q_1)\log(1-4p_1q_1)\right) \\ &\quad + (1-4p_1q_1)\left((1-4p_2q_2)\log(1-4p_2q_2)\right) \\ &\leqslant (1-4p_1q_1)\log(1-4p_1q_1) + (1-4p_2q_2)\log(1-4p_2q_2) \\ &\quad + 64p_1q_1p_2q_2,\end{aligned}$$

The last inequality follows from the fact $-(1-4p_iq_i)\log(1-4p_iq_i) \leqslant 8p_iq_i$, which in turn follows from the inequality $x\log\frac{1}{x} \leqslant$ [3] $2(1-x)$ for $x \in [0,1]$. Thus, we have

$$\begin{aligned}-(1-4p_1q_1)\log(1-4p_1q_1) - (1-4p_2q_2)\log(1-4p_2q_2)& \\ + (1-4pq)\log(1-4pq) \leqslant 64p_1q_1p_2q_2.& \qquad (27)\end{aligned}$$

Next, we simplify the second terms:

$$\begin{aligned}pq &= (p_1q_2+p_2q_1)(p_1p_2+q_1q_2) = p_1q_1(p_2^2+q_2^2) + p_2q_2(p_1^2+q_1^2) \\ &= p_1q_1(1-2p_2q_2) + p_2q_2(1-2p_1q_1) \\ &= p_1q_1 + p_2q_2 - 4p_1q_1p_2q_2.\end{aligned}$$

Hence, we have

$$-8p_1q_1 - 8p_2q_2 + 8pq = -32p_1q_1p_2q_2. \qquad (28)$$

[3] Any constant $c \geqslant \frac{1}{\ln 2}$ can be used instead of 2.

Finally, the first terms:

$$
\begin{aligned}
\mathrm{H}(p) &= \mathrm{H}(p_1q_2 + p_2q_1) \\
&= (p_1q_2 + p_2q_1)\log\frac{1}{(p_1q_2 + p_2q_1)} + (p_1p_2 + q_1q_2)\log\frac{1}{(p_1p_2 + q_1q_2)} \\
&= p_1q_2\log\frac{1}{p_1q_2} + p_1q_2\log\frac{p_1q_2}{(p_1q_2 + p_2q_1)} + p_2q_1\log\frac{1}{p_2q_1} + p_2q_1\log\frac{p_2q_1}{(p_1q_2 + p_2q_1)} \\
&\quad + \text{ similar terms for the second summand} \\
&= q_2(-p_1\log p_1) + p_1(-q_2\log q_2) + p_2(-q_1\log q_1) + q_1(-p_2\log p_2) \\
&\quad + p_1q_2\log\frac{p_1q_2}{(p_1q_2 + p_2q_1)} + p_2q_1\log\frac{p_2q_1}{(p_1q_2 + p_2q_1)} \\
&\quad + \text{ similar terms from the second half} \\
&= -p_1\log p_1(q_2 + p_2) - q_1\log q_1(p_2 + q_2) - p_2\log p_2(q_1 + p_1) - q_2\log q_2(p_1 + q_1) \\
&\quad + p_1q_2\log\frac{p_1q_2}{(p_1q_2 + p_2q_1)} + p_2q_1\log\frac{p_2q_1}{(p_1q_2 + p_2q_1)} + p_1p_2\log\frac{p_1p_2}{(p_1p_2 + q_1q_2)} \\
&\quad + q_1q_2\log\frac{q_1q_2}{(p_1p_2 + q_1q_2)} \\
&= \mathrm{H}(p_1) + \mathrm{H}(p_2) - (p_1q_2 + p_2q_1)\,\mathrm{H}\left(\frac{p_1q_2}{(p_1q_2 + p_2q_1)}\right) - (p_1p_2 + q_1q_2)\,\mathrm{H}\left(\frac{p_1p_2}{(p_1p_2 + q_1q_2)}\right) \\
&\leqslant \mathrm{H}(p_1) + \mathrm{H}(p_2) - 2\min\{p_1q_2, p_2q_1\} - 2\min\{p_1p_2, q_1, q_2\} \quad \text{using } \mathrm{H}(p) \geqslant 2\min\{p, q\} \\
&\leqslant \mathrm{H}(p_1) + \mathrm{H}(p_2) - 2p_1q_2p_2q_1 - 2p_1p_2q_1q_2 \quad \text{since } \min\{p, q\} \geqslant pq \text{ for } 0 \leqslant p, q \leqslant 1 \\
&= \mathrm{H}(p_1) + \mathrm{H}(p_2) - 4p_1q_1p_2q_2.
\end{aligned}
$$

Hence, we have

$$-8\,\mathrm{H}(p_1) - 8\,\mathrm{H}(p_2) + 8\,\mathrm{H}(p) \leqslant -32p_1q_1p_2q_2. \tag{29}$$

Combing Eq. (27), Eq. (28), Eq. (29), and the definition of κ Eq. (24), we obtain

$$\kappa(p_1) + \kappa(p_2) - \kappa(p) \leqslant 0$$

and this concludes the proof. □

ECCC ISSN 1433-8092

https://eccc.weizmann.ac.il