The Harmonic Analysis of Boolean Functions

Jacob Denson

January 14, 2017

Table Of Contents

Chapter 1

Introduction to Boolean Analysis

This course is about the harmonic analysis of functions on the abelian group $\mathbf{F}_2^n$, and the surprising applications of the field to computing science. It seems that whenever one wants to deal with a combinatorial problem involving the Hamming distance, which takes two strings x and y, and counts the number of indices for which $x_i \neq y_i$,

$$\Delta(x,y) = \#\{i : x_i \neq y_i\}$$

For these problems, harmonic analysis seems to be the right tool for the job. It turns out that measuring the similarity of Fourier coeffients of a function tends to give good information about the hamming distance between two functions.

Perhaps the first reason why harmonic analysis occurs in computing science, is that the operations on $\mathbf{F}_2$ naturally correspond to boolean operations on the truth values $\{\top, \bot\}$. If we consider the correspondence between truth values and $\mathbf{F}_2$, mapping $\top$ to 1, and $\bot$ to 0, then the operation of addition on $\mathbf{F}_2$ corresponds to the xor operation $\oplus$ on T, and multiplication corresponds to the logical operation of conjunction $\wedge$. For truth values X and Y, $X \oplus Y = \top$ if and only if $X \neq Y$, so that the output of xor is intimately connected to the Hamming distance between two strings.

It shall be notationally convenient to consider a further correspondence between the additive structure of $\mathbf{F}_2$ and the multiplicative group $\mu_2 = \{-1, 1\}$ of 2nd roots of unity.

$$0 \mapsto 1 \qquad 1 \mapsto -1$$

This is initially a very strange correspondence to pick, since this means that $\top$ corresponds to -1, and $\bot$ to 1, and we normally think of 1 as the 'true' truth value. However, we note that the correspondence does give an isomorphism between the two group structures, and fits with the general convention of performing harmonic analysis over the multiplicative group $\mathbf{T}^n$ where possible. Furthermore, since harmonic analysis considers the representations of functions valued in real or complex numbers by characters, we require a representation of $\mathbf{F}_2$ in the complex numbers in order to apply harmonic analysis to functions $f : \mathbf{F}_2^n \to \mathbf{F}_2$.

On $\mathbf{T}$, the characters take the form of monomials $x_1^{m_1} \dots x_n^{m_n}$. The imbedding $\mu_2^n \to \mathbf{T}^n$ gives rise to a surjective reduction map $(\mathbf{T}^n)^* \to (\mu_2^n)^*$, so that every character on $\mathbf{F}_2$ can be represented by a monomial, and since $x_i^2 = 1$ for all $x_i \in \mathbf{F}_2$, these monomials can be chosen with $m_i \in \{0, 1\}$. It is easy to verify that these are all the identifications we can make, so that we have 2^n characters on the space. Since this is equivalent to identifying the monomial with the subset S of $[n]$ upon which m_i takes the value one, we shall often let x^S stand for this monomial. Thus we have 2^n charactes on $\{-1, 1\}^n$, and they can all be identified with a subset of $[n]$. The basic theory of discrete abelian harmonic analysis tells us that for any $f : \mathbf{F}_2^n \to \mathbf{R}$ we have a unique expansion

$$f(x) = \sum_{S \subset [n]} \widehat{f}(S) x^S$$

where $\widehat{f}(S) \in \mathbf{R}$ is the Fourier coefficient corresponding to S.

Example. *The maximum function* $\max(x, y)$ *on* μ_2^n *has a Fourier expansion*

$$\max(x, y) = \frac{1}{2}(1 + x + y - xy)$$

which corresponds to the the conjunction function $f(x, y) = x \wedge y$ *on* $\{\top, \bot\}^2$*, or the minimum function on* $\mathbf{F}_2^2$*. To obtain this expansion, and a more general expansion for the maximum function on n variables note that we can write*

$$\mathbf{I}(x_1 = -1, \dots, x_n = -1) = \mathbf{I}(x = -1) = \frac{1}{2^n} \prod_{i=1}^{n} (1 - x_i) = \frac{1}{2^n} \sum_S (-1)^{|S|} x^S$$

so that

$$\begin{aligned}\max(x_1,\ldots,x_n) &= \mathbf{I}(x \neq -1) - \mathbf{I}(x = -1)\\ &= 1 - 2\mathbf{I}(x = -1)\\ &= \left(1 - \frac{1}{2^{n-1}}\right) - \frac{1}{2^{n-1}} \sum_{S \neq \emptyset} (-1)^{|S|} x^S\end{aligned}$$

The general approach of writing a function as a linear combination of functions with well known Fourier expansion is essentially a 'power series' method of Boolean expansions.

Example. *The majority function $Maj_3(x,y,z)$ on μ_2^3 returns the boolean value which appears most often in x,y,z. Then*

$$Maj_3(x,y,z) = \frac{1}{2}(x + y + z - xyz)$$

It has the same interpretation on the other boolean domains.

Example. *Consider the minimum function $\min(x) : \mu_2^n \to \mu_2$ on n bits, which corresponds to the disjunction operation on $\{\top,\bot\}^n$. Since we have the relationship $\min(x) = -\max(-x)$, we have $\min = -\max \circ (\pi^1,\ldots,\pi^n)$, where $\pi \in S_2$ is the permutation which swaps 1 with -1. In general, if $h = f \circ g$, where $x^S \circ g = \sum c^S_{S'} x^{S'}$ for some constants $c^S_{S'}$, and if we have expansions*

$$h(x) = \sum a_S x^S \qquad f(x) = \sum b_S x^S$$

Then

$$(f \circ g)(x) = \sum b_S c^S_{S'} x^{S'}$$

so

$$a_S = \sum b_{S'} c^{S'}_S$$

In our case, if g is the negation operation $g(x_1,\ldots,x_n) = (-x_1,\ldots,-x_n)$, then $(x^i \circ g)(x) = -x^i$, so $(x^S \circ g) = (-1)^S x^S$ and therefore

$$\min(x) = -\left[\left(1 - \frac{1}{2^{n-1}}\right) - \frac{1}{2^{n-1}} \sum_{S \neq \emptyset} x^S\right] = \left(\frac{1}{2^{n-1}} - 1\right) + \frac{1}{2^{n-1}} \sum_{S \neq \emptyset} x^S$$

Note that the L^1 distances between the functions min *and* max *and constant functions tend to zero, so that the L^∞ distance of their Fourier transforms also*

tend to zero, and therefore the non-constant Fourier coefficients of min *and* max *become very small over time (this becomes rigorous only when we view each μ_2^n as a subset of the abelian group $\mu_2^\infty = \prod_{i=1}^\infty \mu_2$, with the same notion of L_1 distance and L_∞ distance).*

Example. *In general, the indicator functions $\mathbf{I}(x = a) : \mu_2^n \to \{0,1\}$, for some $a \in \mu_2^n$ can be expressed as*

$$\mathbf{I}(x = a) = \frac{1}{2^n}\prod_{i=1}^n (1 - x_i a_i) = \frac{1}{2^n}\sum_S (-1)^{|S|}\left(\prod_{i\in S} a_i\right) x^S$$

If we only have a $X = \{i_1, \ldots, i_n\}$ subset of indices we want to specify, then

$$\mathbf{I}(x_{i_1} = a_1, \ldots, x_{i_k} = a_k) = \frac{1}{2^n}\sum_S (-1)^{|S|}\left(\prod_{i\in S\cap X} a_i\right) x^S$$

Example. *Consider n $\{-1,1\}$-valued Bernoulli distributions with means $\mu_1, \ldots, \mu_n$, and consider the corresponding product distribution of $\{-1,1\}^n$, with density function f. If X is a random variable with this distribution, then*

$$\mathbf{P}(X_i = 1) - \mathbf{P}(X_i = -1) = 2\mathbf{P}(X_i = 1) - 1 = \mu_i$$

so $\mathbf{P}(X_i = 1) = \frac{1}{2}(\mu_i + 1) = P_i \in [0,1]$. Then we can write

$$\begin{aligned}
f(x_1, \ldots, x_n) &= \prod_{i=1}^n [P_i \mathbf{I}(x_i = 1) + (1 - P_i)\mathbf{I}(x_i = 1)] \\
&= \sum_{a\in\mathbf{F}_2^n} \left(\prod_{a_i=1} P_i\right)\left(\prod_{a_i=0}(1 - P_i)\right)\mathbf{I}(x = a) \\
&= \frac{1}{2^n}\sum_S (-1)^{|S|} \sum_{a\in\mathbf{F}_2^n} \left(\prod_{i\in S} a_i\right)\left(\prod_{a_i=1} P_i\right)\left(\prod_{a_i=0}(1 - P_i)\right) x^S \\
&= \frac{1}{2^n}\sum_S (-1)^{|S|} \prod_{i\in S}(2P_i - 1) x^S = \frac{1}{2^n}\sum_S (-1)^{|S|}\left(\prod_{i\in S}\mu_i\right) x^S
\end{aligned}$$

where we obtain the last equation by repeatedly factoring out the coordinates a_i for $a_i = 1$ and $a_i = -1$.

Example. *Consider the 'inner product mod 2' function f on $\mathbf{F}_2^n \times \mathbf{F}_2^n$ defined by*

$$f(x,y) = (-1)^{\langle x,y\rangle} = \prod_{i=1}^{n} (-1)^{x_i y_i}$$

Then $f(x,y)$ measures the parity of the number of x_i which are equal to y_i. For $n = 1$, we have

$$f(x,y) = (-1)^{xy} = \begin{cases} -1 & x = y = 1 \\ 1 & \textit{otherwise} \end{cases}$$

Hence, if we view f as a function on $\{-1,1\}^2$, the induced function is just the max *function on two variables, and has a Fourier representation*

$$f(x,y) = \frac{1}{2}(1 + x + y - xy)$$

In general, any set S corresponding to a set of indices on the functions of f can be associated with a unique pair of $S_1, S_2 \subset [n]$, corresponding to the indices relating to x on the left side of the function, and the indices relating to y on the right side of the equation. Let $\alpha(S)$ be the cardinality of $S_1 \cap S_2$. Then we have

$$\begin{aligned} f(x,y) &= \frac{1}{2^n} \prod_{i=1}^{n} (1 + x_i + y_i - x_i y_i) \\ &= \frac{1}{2^n} \sum_{S \subset [n]} (-1)^{|S|} \left(\prod_{i \in S} x_i y_i \right) \left(\prod_{i \notin S} (1 + x_i + y_i) \right) \\ &= \frac{1}{2^n} \sum_{S \subset [n]} (-1)^{|S|} \sum_{S' \subset S^c} \left(\prod_{i \in S} x_i y_i \right) \left(\prod_{i \in S'} (x_i + y_i) \right) \\ &= \frac{1}{2^n} \sum_{S} (-1)^{\alpha(S)} x^S \end{aligned}$$

1.1 Testing Linearity

A function $f : \mathbf{F}_2^n \to \mathbf{F}_2$ is linear if and only if $f(x+y) = f(x) + f(y)$ for all $x, y \in \mathbf{F}_2^n$, or equivalently, if there is $a \in \mathbf{F}_2^n$ such that $f(x) = \langle a, x \rangle$. Given a general function f, there are effectively only two ways to explicitly check that the function is linear, we either check that $f(x+y) = f(x) + f(y)$ for all

choices of the arguments x and y, or check that $f(x) = \langle a, x \rangle$ holds for some choice of a, over all choices of x. Even assuming that f can be evaluated in constant time, both methods take exponential time in n to compute. This is effectively guaranteed, because the description of f is always exponential in n, and if an algorithm determining linearity does not check the entire description of a linear function f, we can modify f to a non-linear function g which is not linear, yet indistinguishable to f int he algorithm. The problem of testing linearity is a method of family of problesm in the field of **property testing**, which attempts to design efficient algorithms to determine whether a particular boolean-valued function has a certain property.

We might not be able to come up with a polynomial time algorithm to verify linearity, but we can make headway by considering the possiblity of coming up with an appropriate *randomized* algorithm which can verify linearity with high probability. The likely solution to the problem, given some function f, would to perform the linearity test for $f(X+Y) = f(X)+f(Y)$ for a certain set of randomly chosen inputs X and Y. If $f(X+Y) \neq f(X)+f(Y)$ for some particular input, we can guarantee the function is non-linear. Otherwise, we have a high probability that the function is linear, in a manner which becomes more likely as we test more random inputs.

To proceed along these lines, we must introduce the probabilistic interpretation of the Fourier expansion. The Fourier coefficients of $f : \mathbf{F}_2^n \to \mathbf{R}$ can, of course, be calculated as

$$\widehat{f}(S) = \frac{1}{2^n} \sum_x f(x) x^S$$

where the 2^n factor is included so that characters are normalized. We can also see this as a probabilistic statement, if we consider the uniform distribution on μ_2^n, and consider the corresponding random variable X. Then we can write

$$\widehat{f}(S) = \mathbf{E}[f(X)X^S]$$

So that the Fourier coeffient measures the average value of f, relative to the parity over S. Note that this implies that a boolean-function $f : \mu_2^n \to \mu_2^n$ is *unbiased*, that is, it has an equal chance of taking -1 and 1, if and only if $\widehat{f}(\varnothing) = 0$. Of course, we have Parseval's equality

$$\mathbf{E}[f^2(X)] = \sum \widehat{f}^2(S)$$

and so

$$\mathbf{V}[f(X)] = \mathbf{E}[f^2(X)] - \mathbf{E}[f(X)]^2 = \sum_{S \neq \emptyset} \widehat{f}^2(S)$$

and similarily,

$$\mathrm{Cov}[f(X), g(X)] = \sum_{S \neq \emptyset} \widehat{f}(S)\widehat{g}(S)$$

So that the Fourier coefficients are efficient measures of probabilistic quantities.

If f and g are *boolean-valued* maps $\mathbf{F}_2^n \to \{-1, 1\}$, then

$$\begin{aligned}\mathbf{E}[f(X)g(X)] &= \mathbf{P}(f(X) = g(X)) - \mathbf{P}(f(X) \neq g(X)) \\ &= 1 - 2\mathbf{P}(f(X) \neq g(X))\end{aligned}$$

Note that $\mathbf{P}(f(X) \neq g(X))$ differs from the Hamming distance of f and g by a constant factor. We define $\mathbf{P}(f(X) \neq g(X))$ to be the relative hamming distance between f and g, denoted $d(f, g)$. Since $f^2 = 1$, this implies $\|f\|_2 = 1$, and therefore

$$\begin{aligned}\mathbf{V}(f) &= \mathbf{E}[f(X)^2] - \mathbf{E}[f(X)]^2 \\ &= 1 - \mathbf{E}[f(X)]^2 \\ &= 1 - (\mathbf{P}(f(X) = 1) - \mathbf{P}(f(X) = -1))^2 \\ &= 4\mathbf{P}(f(X) = -1)\mathbf{P}(f(X) = 1) \in [0, 1]\end{aligned}$$

so the variance of a boolean-valued function is very closely related to the degree to which the function is constant.

Theorem 1.1. *If* $\varepsilon = \min[d(f, 1), d(f, -1)]$, *then* $2\varepsilon \leqslant \mathbf{V}(f) \leqslant 4\varepsilon$.

Proof. We are effectively proving that for any $x \in [0, 1]$, $2\min(x, 1 - x) \leqslant 4x(1 - x) \leqslant 4\min(x, 1 - x)$. Since $4x(1 - x) = 4\max(x, 1 - x)\min(x, 1 - x)$, we may divide by $4\min(x, 1 - x)$ to restate the inequality as

$$1/2 \leqslant \max(x, 1 - x) \leqslant 1$$

which is true, because $x, 1 - x \leqslant 1$, and also

$$\max(x, 1 - x) \geqslant \frac{1}{2}[x + (1 - x)] = 1/2$$

This shows that if we have a sequence of functions $f_i : \{-1,1\}^n \to \{-1,1\}$, then $\varepsilon_i \sim x_i$ holds if and only if $\mathbf{V}(f) \sim x_i$. The minimum hamming distance to a constant value is asymptotically the same as the variation of the function. □

Since a function $f : \{-1,1\}^n \to \{-1,1\}$ always has square-norm 1, this implies that $\sum \widehat{f}(S)^2 = 1$ if we sum over all the Fourier coeffients, so that a boolean-valued function on n variables gives rise to a probability distribution over $[n]$. We call this the spectral sample of f.

Often times, the fourier coeffients of sets over some particular cardinality are more than enough to determine some result. We define the weight of a function $f : \{-1,1\}^n \to \mathbf{R}$ of degree k to be

$$W^k(f) = \sum_{|S|=k} \widehat{f}^2(S)$$

If f is boolean-valued, then we could have also defined $W^k(f)$ as the probability that some set drawn from the spectral sample of f has cardinality k. This is also the square norm of the function

$$f^k(x) = \sum_{|S|=k} \widehat{f}(S) x^S$$

which can be seen as the projection of f onto a subspace of the characters in $(\mathbf{F}_2^n)^*$.

The simplest version of linearity testing runs using one random query as a test – it just takes one pair of inputs (X, Y), and tests the linearity of this function against these inputs. This is known as the Blum-Luby-Rosenfeld algorithm, or BLR for short. It turns out that the sucess of this method is directly related to how linear a function is. Define a function to be **approximately linear** if

- $f(x+y) = f(x) + f(y)$ for a large majority of inputs x and y.
- There is a such that $f(x) = \langle a, x \rangle$ for a large number of inputs x.

It is clear that the second property implies the first, but not that the first implies the second in a precise relationship. To be more precise, we say that f is ε-close to being linear if there is a linear function g with $d(f,g) \leqslant \varepsilon$. What the analysis of BLR shows is that $f(X+Y) = f(X) + f(Y)$ holds

with probability ε, then f is ε close to a linear function. Since we are applying Hamming distances in measuring how linear a function is, we can guess that the Harmonic analysis of boolean functions will come in handy.

Theorem 1.2. *Given a function $f : \mathbf{F}_2^n \to \mathbf{F}_2$, the BLR algorithm is correct with probability $1 - \varepsilon$ if and only if f is ε-close to being linear.*

Proof. We switch to multiplicative notation, so that $f : \{-1, 1\}^n \to \{-1, 1\}$. Note that the probability that $f(XY) = f(X)f(Y)$ is the same as the probability that

$$\frac{1}{2}[1 + f(X)f(Y)f(XY)] = 1$$

and if $f(XY) \neq f(X)f(Y)$, then

$$\frac{1}{2}[1 + f(X)f(Y)f(XY)] = 0$$

so that the function $g(x, y) = (1/2)[1 + f(x)f(y)f(xy)]$ is 0-1 valued, and

$$\begin{aligned} 1 - \varepsilon &= \mathbf{P}[\text{BLR accepts } f] \\ &= \mathbf{E}[g(X, Y)] \\ &= \frac{1}{2} + \frac{1}{2}\mathbf{E}_X[f(X)\mathbf{E}_Y[f(Y)f(XY)]] \\ &= \frac{1}{2} + \frac{1}{2}\mathbf{E}_X[f(X)(f * f)(X)] \\ &= \frac{1}{2} + \frac{1}{2}\sum \widehat{f}(S)^3 \end{aligned}$$

and therefore

$$1 - 2\varepsilon = \sum \widehat{f}(S)^3 \leqslant \|\widehat{f}\|_\infty \|f\|_2 = \|\widehat{f}\|_\infty$$

Now $\widehat{f}(S) = 1 - 2d(f, x^S)$, and since there is S' with $\widehat{f}(S') \geqslant 1 - 2\varepsilon$, this implies $d(f, x^{S'}) \leqslant \varepsilon$, and therefore f is ε-close to the linear function $x^{S'}$. □

Note that this algorithm, with only three evaluations of the function f, we can determine with high accuracy that f is not linear, or, if we are wrong with high probability, then f is very similar to a linear function in the first case. Yet given f, we cannot use this algorithm to determine the

linear function x^S which f is similar to. Of course, for most x, $f(x) = x^S$. Yet we cannot use this estimate to obtain accurate estimates for a fixed x, in the sense that there is no measure of the likelihood of the estimate being equal to the actual value. Note, however, that $x^S(x) = x^S(x+y)x^S(y)$, and if we let y be a random quantity, then $x^S(x+y)$ and $x^S(y)$ will likely be equal to $f(x+y)$ and $f(y)$ with a probability that is feasible to determine.

Theorem 1.3. *If $f : \mathbf{F}_2^n \to \{-1, 1\}$ is ε-close to x^S, then for any y,*

$$\mathbf{P}_X[f(X+y)f(X) = x^S(y)] \geqslant 1 - 2\varepsilon$$

Proof. The probability that $f(X+y) \neq x^S(X+y)$ is less than ε, as is the probability that $f(X) \neq x^S(X)$. By the union bound, this implies that the probability of either occuring is less than or equal to 2ε. But then the probability that both do not occur is greater than or equal to $1 - 2\varepsilon$, and when this occurs, we can guarantee that $f(X+y)f(X) = x^S(y)$, thus our estimate $f(X+y)f(X)$ is equal to $x^S(y)$ with probability $\geqslant 1 - 2\varepsilon$. □