On Learning Monotone Boolean Functions under the Uniform Distribution

Kazuyuki Amano *, Akira Maruoka

Graduate School of Information Sciences, Tohoku University
Aoba 05, Aramaki, Sendai 980-8579 JAPAN

Abstract

In this paper, we prove two general theorems on monotone Boolean functions which are useful for constructing a learning algorithm for monotone Boolean functions under the uniform distribution.

A monotone Boolean function is called fair if it takes the value 1 on exactly half of its inputs. The first result proved in this paper is that a single variable function $f(x) = x_i$ has the minimum correlation with the majority function among all fair monotone functions. This proves the conjecture by Blum, Burch and Langford (1998, Proc. 39th FOCS, pp.408–415) and improves the performance guarantee of the best known learning algorithm for monotone Boolean functions under the uniform distribution they proposed.

Our second result is on the relationship between the influences and the average sensitivity of a monotone Boolean function. The influence of variable x_i on f is defined as the probability that $f(x)$ differs from $f(x \oplus e_i)$ where x is chosen uniformly from $\{0,1\}^n$ and $x \oplus e_i$ means x with its i-th bit flipped. The average sensitivity of f is defined as the sum of the influences over all variables x_i. We prove that a somewhat unintuitive result which says if the influence of every variable on a monotone Boolean function is small, i.e., $O(1/n^c)$ for some constant $c > 0$, then the average sensitivity of the function must be large, i.e., $\Omega(\log n)$. We also discuss how to apply this result to the construction of a new learning algorithm for monotone Boolean functions.

Key words: PAC Learning, monotone Boolean functions, Harmonic analysis, majority function

* Corresponding author.
Email addresses: ama@ecei.tohoku.ac.jp (Kazuyuki Amano), maruoka@ecei.tohoku.ac.jp (Akira Maruoka).

1 Introduction

A *monotone* Boolean function f maps $\{0,1\}^n$ to $\{0,1\}$ in such a way that $f(x) \leq f(y)$ whenever $x \leq y$. The problem of learning monotone Boolean functions has been widely investigated because they contain a very broad class of Boolean expressions.

For a target function f and a hypothesis function h, we define the error of h for f, denoted by $\mathrm{err}(f, h)$, as the fraction of points x such that $f(x) \neq h(x)$; that is $\Pr[f(x) \neq h(x)]$ where x is chosen uniformly from $\{0,1\}^n$. We sometimes use the closely related concept of *correlation*, which is defined to be $1 - 2\Pr[f(x) \neq h(x)]$.

The algorithms we discuss in this paper are for learning monotone Boolean functions under the uniform distribution. The algorithm is assumed to be given access to an example oracle EXAMPLE for an unknown target function f: When the oracle EXAMPLE is invoked, it produces an example $(x, f(x))$ for a vector x chosen uniformly from $\{0,1\}^n$. The goal of a learning algorithm is to produce a good approximation to a target f in polynomial time by calling oracle EXAMPLE a polynomial number of times.

The first algorithm for learning monotone Boolean functions under the uniform distribution was proposed by Kearns et al.[9]. Their algorithm produces one of the constant-zero, the constant-one or a single variable function that has the highest correlation on the data obtained by calling oracle EXAMPLE. The output was shown to have an error of $1/2 - \Omega(1/n)$ at most. Subsequently, Bshouty and Tamon [4] gave another algorithm which outputs a linear threshold function as a hypothesis and guarantees an error of $1/2 - \Omega((\log n)^2/n)$ at most. Recently, Blum, Burch and Langford [3] improved on this guarantee by giving a simple learning algorithm which achieves an error of $1/2 - \Omega(1/\sqrt{n})$: Draw $O(n)$ examples labeled by a target f. If at least a $1/2 + 0.05/\sqrt{n}$ fraction of examples are labeled by 1 then the algorithm outputs the constant-one function. Likewise, if at least a $1/2 + 0.05/\sqrt{n}$ fraction of examples are labeled by 0, then the algorithm outputs the constant-zero function. Otherwise, the algorithm blindly outputs as a hypothesis the majority function over all the variables. Remarkably, their simple algorithm is nearly the best possible even if the algorithm is allowed to access an oracle more powerful than EXAMPLE, i.e., the membership oracle MEMBER that allows the algorithm to query f at arbitrary points of its choosing. Furthermore they proved that no algorithm can guarantee an error of $1/2 - \omega(\log n/\sqrt{n})$ provided that the algorithm is allowed to access to oracle MEMBER only a polynomial number of times.

The performance guarantee of their algorithm is obtained by proving that for every *fair* monotone Boolean function f, the error of the majority function for f is at most $1/2 - \Omega(1/\sqrt{n})$. Here, a Boolean function is called *fair* if it takes the value 1 on exactly half of its inputs. In [3], they conjectured that a single variable function has the minimum correlation with the majority function among all fair monotone functions. The first result proved in the present paper is that the conjecture is true, i.e., the single variable function is the "farthest" from the majority function among all fair monotone functions. Although our improvement on the error of $1/2-\Omega(1/\sqrt{n})$ is only on the hidden constant in the Ω notation, we believe that the result is interesting in its own right.

Applying this result to the learning algorithm described above, we can see that the "worst" case occurs when a target function is a single variable function. We can easily determine, however, whether the target function is well correlated with a single variable. This consideration leads to a new algorithm that outputs the function in $\{0, 1, x_1, \ldots, x_n, \text{majority}\}$ that has the highest observed correlation on examples.

In the second part of this paper, we analyze the performance of this algorithm by applying ideas from Harmonic analysis and obtain a general theorem on the relationship between the influences and the average sensitivity of a monotone Boolean function. The influence of variable x_i on a Boolean function f is defined as the probability that $f(x)$ differs from $f(x \oplus e_i)$ where x is chosen uniformly from $\{0,1\}^n$ and $x \oplus e_i$ means x with its i-th bit flipped. The average sensitivity of f is defined as the sum of the influences of all the variables. The influences and the average sensitivity of Boolean functions and their application to learning have been widely investigated (e.g. [8,4]).

It is easily seen that for every monotone Boolean function f, the influence of x_i on f is identical to the correlation between f and x_i. If a single variable function x_i has a high correlation with a target function f on examples then x_i is a good approximation to f, so we are done. We can therefore assume that the target function has only a small influence for every variable.

The second result we establish in this paper is a somewhat unintuitive statement on the relationship between the influences and the average sensitivity of a monotone Boolean function. Our theorem says that if the influence of x_i on a monotone Boolean function f is small, i.e., $O(1/n^c)$ for some constant $c > 0$, for every variable x_i then the average sensitivity of f must be large, i.e., $\Omega(\log n)$. We also observe that a monotone Boolean function having large average sensitivity is, in some sense, a good approximation of the majority function. We conjecture that the error of the majority function approximating a fair monotone Boolean function with average sensitivity $s(f)$ is at most $1/2-\Omega(s(f)/\sqrt{n})$. If this conjecture is true, then the new algorithm, proposed in this paper, to learn monotone Boolean functions can achieve the optimal error $1/2 - \Omega(\log n/\sqrt{n})$.

2 Distance between Fair Monotone Function and Majority Function

For $x \in \{0,1\}^n$, $||x||$ represents the number of 1's in x, i.e., $||x|| = \sum_i x_i$ where x_i denotes the i-th bit of x. For a set S, $\sharp S$ denotes the cardinality of S. For a set of vectors $S \subseteq \{0,1\}^n$ and an integer $0 \le k \le n$, we use S_k to represent the set of vectors that contains k 1's, i.e., $S_k = \{x \in S \mid ||x|| = k\}$. For a natural number n, $[n]$ stands for the set $\{1, 2, \ldots, n\}$. For a Boolean function f, we define $I^f = \{x \mid f(x) = 1\}$ and $O^f = \{x \mid f(x) = 0\}$. For simplicity, we abbreviate $(I^f)_k$ and $(O^f)_k$ as I^f_k and O^f_k, respectively. A Boolean function f is called *fair* if f outputs 1 for exactly half of its inputs, i.e., $\sharp I^f = \sharp O^f$ holds. A Boolean function f of n variables is *monotone* if $f(x_1, \ldots, x_n) \le f(y_1, \ldots, y_n)$ whenever $x_i \le y_i$ for all i.

The *majority* function of n variables, denoted by MAJ_n, is a function that outputs 1 if and only if the number of 1's in its input exceeds $n/2$, i.e.,

$$\mathrm{MAJ}_n(x) = \begin{cases} 1 & \text{if } ||x|| \ge \lfloor n/2 \rfloor + 1, \\ 0 & \text{if } ||x|| \le \lfloor n/2 \rfloor. \end{cases}$$

In [3], Blum et al. conjectured that any single variable function has the minimum correlation with the majority function among all fair monotone functions. In this section, we prove that their conjecture is true.

Theorem 1 *For any fair monotone Boolean function f of n variables,*

$$\sharp\{x \mid MAJ_n(x) \ne f(x)\} \le \sharp\{x \mid MAJ_n(x) \ne x_1\} = 2^{n-1} - \binom{n-1}{\lfloor n/2 \rfloor}.$$

□

The single variable function disagrees with the majority function on

$$2^{n-1} - \binom{n-1}{\lfloor n/2 \rfloor} \sim \left(\frac{1}{2} - \frac{1}{\sqrt{2\pi n}}\right) 2^n \le \left(\frac{1}{2} - \frac{0.4}{\sqrt{n}}\right) 2^n$$

points. Blum et al. obtained a slightly weaker result to guarantee the performance of the learning algorithm for monotone Boolean functions under the uniform distribution described in Section 1.

Theorem 2 (Blum, Burch, Langford [3]) *For any fair monotone Boolean* *function f of n variables,*

$$\sharp\{x \mid MAJ_n(x) \neq f(x)\} \leq \left(\frac{1}{2} - \frac{0.1}{\sqrt{n}}\right) 2^n.$$

□

To prove Theorem 2, they analyzed the quantities p_k, which are defined as $p_k = \sharp I_k^f / \binom{n}{k}$, i.e., the fraction of inputs x contains exactly k 1's such that $f(x) = 1$. The key to proving the theorem is to apply the Kruskal-Katona Theorem (e.g., [6,5]) that says the p_k must increase with k at a reasonable rate. More specifically, they insisted that, for every monotone Boolean function and for every $0 \leq i < j \leq n$, $(p_i)^i \leq (p_j)^j$ holds ([3, Lemma 2]) as a corollary to the Kruskal-Katona theorem, and applied this to prove the theorem.

Remark: Unfortunately, the above proposition ([3, Lemma 2]) does not hold for certain ranges of i and j. For $i \geq 2$, define the Boolean function f as

$$f = x_1 \left(\bigvee_{k_2, k_3, \ldots, k_i \in \{2, \ldots, n\}} x_{k_2} x_{k_3} \cdots x_{k_i} \right).$$

It is easy to see that $p_j = \binom{n-1}{j-1} / \binom{n}{j} = j/n$ for every $i \leq j \leq n$. If we set $i = 0.3n$ and $j = 0.4n$, then $(p_i)^i = (3/10)^{0.3n} \sim (0.697)^n > (0.693)^n \sim (4/10)^{0.4n} = (p_j)^j$ holds and this contradicts the above proposition. Generally, the above proposition holds when the values of i and j are sufficiently close to $n/2$ (this can be proved by Corollary 4 stated below and a simple calculation) and may not hold for other cases. In the proof of Theorem 2 Blum et al. applied the above lemma for $i = n/2 - c\sqrt{n}$ and $j = n/2 + c\sqrt{n}$ where c is a small constant. So the fact that $(p_i)^i \leq (p_j)^j$ does not necessarily hold for some $i < j$ does not affect the correctness of Theorem 2 in [3].

To prove Theorem 1, we use another version of the Kruskal-Katona Theorem which was introduced by Lovász([10, Exercises 13.31], [6, Theorem 2]). This is a slightly weaker but much handier form of the Kruskal-Katona Theorem. Intuitively this theorem says that for every monotone Boolean function and for every $1 \leq k \leq n-1$, if $p_k \geq k/n$ then $p_{k+1} \geq (k+1)/n$ holds. This is a stronger statement than the proposition used by Blum et al. when k is close to $n/2$.

Now we state Lovász's version of the Kruskal-Katona Theorem. For $\mathcal{F} \subseteq 2^{[n]}$, define $\Delta(\mathcal{F}) = \{E \subseteq [n] \mid \exists F \in \mathcal{F}\ E \subseteq F, \sharp E = \sharp F - 1\}$.

Lemma 3 (Lovász [10]) *Let l and k be two integers such that $l \geq k \geq 1$. If $\mathcal{F} \subseteq \{F \in 2^{[n]} \mid \sharp F = k\}$ and $\sharp \mathcal{F} = \binom{l}{k}$, then $\sharp \Delta(\mathcal{F}) \geq \binom{l}{k-1}$.* □

An elegant proof of the lemma using the so-called *shifting technique* appears in [6].

Corollary 4 *For any monotone Boolean function f of n variables and for any $1 \leq k \leq n-1$, if $\sharp I_k^f \geq \binom{n-1}{k-1}$ then $\sharp I_{k+1}^f \geq \binom{n-1}{k}$.*

PROOF. Set $S_x = \{i \in [n] \mid x_i = 0\}$ for $x \in \{0,1\}^n$ and set $\mathcal{F}_{n-k} = \{S_x \mid x \in I_k^f\}$ for $0 \leq k \leq n$. Then there is a one-to-one correspondence between I_k^f and $\mathcal{F}_{n-k}$ and so $\sharp I_k^f = \sharp \mathcal{F}_{n-k}$. By the monotonicity of f, we have $\Delta(\mathcal{F}_{n-k}) \subseteq \mathcal{F}_{n-k-1}$, which implies $\sharp \mathcal{F}_{n-k-1} \geq \sharp \Delta(\mathcal{F}_{n-k})$. On the other hand, from Lemma 3 we can claim that if $\sharp \mathcal{F}_{n-k} (= \sharp I_k^f) \geq \binom{n-1}{k-1} = \binom{n-1}{n-k}$, then $\sharp \Delta(\mathcal{F}_{n-k}) \geq \binom{n-1}{n-k-1}$. Combining these inequalities we have

$$\sharp I_{k+1}^f = \sharp \mathcal{F}_{n-k-1} \geq \sharp \Delta(\mathcal{F}_{n-k}) \geq \binom{n-1}{n-k-1} = \binom{n-1}{k},$$

completing the proof. $\square$

The following lemma follows easily from Corollary 4.

Lemma 5 *For any fair monotone Boolean function f of n variables and for any $1 \leq t \leq n-1$,*

$$\sum_{k=0}^{t} \sharp I_k^f \leq \sum_{k=0}^{t-1} \binom{n-1}{k}, \tag{1}$$

and

$$\sum_{k=t}^{n} \sharp I_k^f \geq \sum_{k=t-1}^{n-1} \binom{n-1}{k}. \tag{2}$$

PROOF. First we prove (1). Suppose in contradiction that $\sum_{k=0}^{t} \sharp I_k^f > \sum_{k=0}^{t-1} \binom{n-1}{k}$ for some $1 \leq t \leq n-1$. Let $t' \in [t]$ be the minimum integer such that $\sum_{k=0}^{t'} \sharp I_k^f > \sum_{k=0}^{t'-1} \binom{n-1}{k}$. Then $\sharp I_{t'}^f > \binom{n-1}{t'-1}$ and $\sum_{k=0}^{t'} \sharp I_k^f = \sum_{k=0}^{t'-1} \binom{n-1}{k} + c$ for some integer $c \geq 1$. By Corollary 4, we have

$$\sharp I_{t'+1}^f \geq \binom{n-1}{t'}.$$

By applying Corollary 4 repeatedly, we obtain

$$\sharp I_t^f \geq \binom{n-1}{t-1}$$

for every $t'+1 \leq t \leq n$. Thus

$$\begin{aligned}
\sharp I^f = \sum_{k=0}^{n} \sharp I_k^f &= \sum_{k=0}^{t'} \sharp I_k^f + \sum_{k=t'+1}^{n} \sharp I_k^f \\
&\geq \sum_{k=0}^{t'-1} \binom{n-1}{k} + c + \sum_{k=t'}^{n-1} \binom{n-1}{k} \\
&\geq \sum_{k=0}^{n-1} \binom{n-1}{k} + c > 2^{n-1}.
\end{aligned}$$

This contradicts the assumption that f is fair and thus completes the proof of (1).

Next we prove (2). Let f' be the dual function of f, i.e., $f'(x) = 1 - f(\bar{x})$. It is obvious that f' is monotone and fair. Since $\sharp I_k^f = \sharp O_{n-k}^{f'} = \binom{n}{n-k} - \sharp I_{n-k}^{f'}$ and $\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$ for $1 \leq k \leq n-1$,

$$\begin{aligned}
(2) &\Longleftrightarrow \sum_{k=t-1}^{n-1} \binom{n-1}{k} - \sum_{k=t}^{n} \sharp I_k^f \leq 0 \\
&\Longleftrightarrow \sum_{k=t-1}^{n-1} \binom{n-1}{k} - \sum_{k=t}^{n} \binom{n}{k} + \sum_{k=t}^{n} \sharp I_{n-k}^{f'} \leq 0 \\
&\Longleftrightarrow \sum_{k=t}^{n} \binom{n-1}{k-1} - \sum_{k=t}^{n} \binom{n}{k} + \sum_{k=t}^{n} \sharp I_{n-k}^{f'} \leq 0 \\
&\Longleftrightarrow - \sum_{k=t}^{n-1} \binom{n-1}{k} + \sum_{k=t}^{n} \sharp I_{n-k}^{f'} \leq 0 \\
&\Longleftrightarrow - \sum_{k=0}^{n-t-1} \binom{n-1}{k} + \sum_{k=0}^{n-t} \sharp I_k^{f'} \leq 0.
\end{aligned}$$

This is identical to inequality (1) when we replace t by $n-t$. □

Now we proceed to the proof of Theorem 1.

PROOF. (of Theorem 1) Let f be a fair monotone Boolean function of n variables. The number of points on which f and the majority function differ can be written as

$$
\begin{aligned}
\sharp\{x \in \{0,1\}^n \mid \mathrm{MAJ}_n(x) \neq f(x)\} \\
&= \sharp\{x \in \{0,1\}^n \mid \mathrm{MAJ}_n(x) \neq x_1\} + \sharp N - \sharp P,
\end{aligned}
$$

where

$$N = \{x \in \{0,1\}^n \mid (\mathrm{MAJ}_n(x) = x_1) \wedge (\mathrm{MAJ}_n(x) \neq f(x))\}$$

and

$$P = \{x \in \{0,1\}^n \mid (\mathrm{MAJ}_n(x) \neq x_1) \wedge (\mathrm{MAJ}_n(x) = f(x))\}.$$

So it is sufficient to show that $\sharp P \geq \sharp N$. In what follows, x denotes a vector of length n, whereas $\tilde{x}$ denotes a vector of length $n-1$. We have

$$
\begin{aligned}
\sharp N &= \sharp\{x \mid \mathrm{MAJ}_n(x) = 0, x_1 = 0, f(x) = 1\} \\
&\quad + \sharp\{x \mid \mathrm{MAJ}_n(x) = 1, x_1 = 1, f(x) = 0\} \\
&= \sharp\{\tilde{x} \mid ||\tilde{x}|| \leq \lfloor n/2 \rfloor, f(0\tilde{x}) = 1\} \\
&\quad + \sharp\{\tilde{x} \mid ||\tilde{x}|| \geq \lfloor n/2 \rfloor, f(1\tilde{x}) = 0\}, \qquad (3)
\end{aligned}
$$

and

$$
\begin{aligned}
\sharp P &= \sharp\{x \mid \mathrm{MAJ}_n(x) = 0, x_1 = 1, f(x) = 0\} \\
&\quad + \sharp\{x \mid \mathrm{MAJ}_n(x) = 1, x_1 = 0, f(x) = 1\} \\
&= \sharp\{\tilde{x} \mid ||\tilde{x}|| \leq \lfloor n/2 \rfloor - 1, f(1\tilde{x}) = 0)\} \\
&\quad + \sharp\{\tilde{x} \mid ||\tilde{x}|| \geq \lfloor n/2 \rfloor + 1, f(0\tilde{x}) = 1)\}. \qquad (4)
\end{aligned}
$$

We denote the first and second terms in (3) by N_0 and N_1, respectively, and the first and second terms in (4) by P_0 and P_1, respectively. From now on, we prove $\sharp P_0 - \sharp N_0 \geq 0$ and $\sharp P_1 - \sharp N_1 \geq 0$, which implies the theorem since we have $\sharp P - \sharp N = \sharp P_0 + \sharp P_1 - (\sharp N_0 + \sharp N_1) \geq 0$.

For $i \in \{0,1\}$ and for $0 \leq k \leq n-1$, define the sets of vectors of length $n-1$ by $I_k^i = \{\tilde{x} \mid ||\tilde{x}|| = k, f(i\tilde{x}) = 1\}$ and $O_k^i = \{\tilde{x} \mid ||\tilde{x}|| = k, f(i\tilde{x}) = 0\}$. Then

$$
\begin{aligned}
\sharp P_0 - \sharp N_0 &= \sum_{k=0}^{\lfloor n/2 \rfloor - 1} \sharp O_k^1 - \sum_{k=0}^{\lfloor n/2 \rfloor} \sharp I_k^0 \\
&= \sum_{k=0}^{\lfloor n/2 \rfloor - 1} \binom{n-1}{k} - \sum_{k=0}^{\lfloor n/2 \rfloor - 1} \sharp I_k^1 - \sum_{k=0}^{\lfloor n/2 \rfloor} \sharp I_k^0
\end{aligned}
$$

$$= \sum_{k=0}^{\lfloor n/2 \rfloor -1} \binom{n-1}{k} - \sum_{k=0}^{\lfloor n/2 \rfloor} \sharp I_k^f \geq 0.$$

The last inequality follows from (1) in Lemma 5. Similarly,

$$\begin{aligned}
\sharp P_1 - \sharp N_1 &= \sum_{k=\lfloor n/2 \rfloor+1}^{n-1} \sharp I_k^0 - \sum_{k=\lfloor n/2 \rfloor}^{n-1} \sharp O_k^1 \\
&= \sum_{k=\lfloor n/2 \rfloor+1}^{n-1} \sharp I_k^0 - \sum_{k=\lfloor n/2 \rfloor}^{n-1} \binom{n-1}{k} + \sum_{k=\lfloor n/2 \rfloor}^{n-1} \sharp I_k^1 \\
&= \sum_{k=\lfloor n/2 \rfloor+1}^{n} \sharp I_k^f - \sum_{k=\lfloor n/2 \rfloor}^{n-1} \binom{n-1}{k} \geq 0.
\end{aligned}$$

The last inequality follows from (2) in Lemma 5. $\square$

To construct an algorithm for learning monotone functions based on our theorem, the following lemma of Blum et al. [3] is useful.

Lemma 6 (Blum, Burch and Langford[3]) *Suppose that there exists an algorithm A for learning fair monotone functions, that uses no examples and outputs a hypothesis with error at most* $1/2 - \epsilon$*. Then there exists an algorithm A' for learning monotone functions that, for any* $\alpha, \delta > 0$*, produces a hypothesis with error at most* $1/2 - \epsilon/(2+\alpha)$ *with probability* $1 - \delta$*, accessing the oracle EXAMPLE* $\frac{2(2+\alpha)^2}{\alpha\epsilon^2}$ *times and calling the algorithm A.* $\square$

We apply Lemma 6 with $\alpha = 1/2$ and $\epsilon = 0.4/\sqrt{n}$ to obtain the following corollary which slightly improves on the result of Blum et al.

Corollary 7 *In linear time, we can learn monotone Boolean function guaranteeing error at most* $1/2 - 0.15/\sqrt{n}$ *under the uniform distribution.* $\square$

An algorithm satisfying the conditions of Corollary 7 is quite simple: Draw $O(n)$ examples from EXAMPLE. If at least a $1/2 + 0.2/\sqrt{n}$ fraction of examples are labeled by 1 then the algorithm outputs the constant-one function. Likewise, if at least a $1/2 + 0.2\sqrt{n}$ fraction of examples are labeled by 0, then the algorithm outputs the constant-zero function. Otherwise, the algorithm blindly outputs as a hypothesis the majority function over all the variables.

3 Relationship between Influences and Average Sensitivity

In the previous section, we proved that the single variable function has the minimum correlation with the majority function among all fair monotone Boolean functions. This leads us to consider a new learning algorithm for monotone Boolean functions under the uniform distribution: Draw enough examples from EXAMPLE and then output as a hypothesis the most correlated function in $\{0, 1, x_1, x_2, \ldots, x_n, \mathrm{MAJ}_n\}$ for the examples.

It is interesting to see whether the above algorithm can achieve error better than $1/2 - \omega(1/\sqrt{n})$. In this section, we analyze the performance of this algorithm by using ideas from Harmonic analysis.

As in the previous section, we can assume that a target function is fair. Now our problem turns out to be estimating $\max_f \min_{h \in \mathcal{H}} err(f, h)$ for $\mathcal{H} = \{x_1, x_2, \ldots, x_n, \mathrm{MAJ}_n\}$, where the maximum is taken over all fair monotone Boolean functions of n variables.

Because of the hardness result described below and Corollary 7, it can be seen that the above value is lower bounded by $1/2 - O(\log n/\sqrt{n})$.

Theorem 8 (Blum, Burch, Langford [3]) *There are no algorithms, given only a polynomial number of accesses to the oracle MEMBER, which can guarantee error $1/2 - \omega(\log n/\sqrt{n})$ for learning monotone Boolean functions under the uniform distribution.* □

Note that given access to MEMBER, we do not need to have access to EXAMPLE because a randomized learning algorithm can simulate EXAMPLE by simply calling MEMBER on uniform random inputs.

Throughout this section we view Boolean functions of n variables as real valued functions which map $\{0, 1\}^n$ to $\{-1, 1\}$: A Boolean function $f(x)$ with range $\{0, 1\}$ is considered to take value $(-1)^{1-f(x)}$. For $x \in \{0, 1\}^n$ and for $1 \le i \le n$, $x \oplus e_i$ denotes x with its i-th bit flipped. For a Boolean function f of n variables, the *influence of* x_i *on* f, denoted by $L^f(x_i)$, is defined as the probability that $f(x)$ differs from $f(x \oplus e_i)$ when x is drawn from the uniform distribution on $\{0, 1\}^n$. That is

$$L^f(x_i) = \Pr[f(x) \neq f(x \oplus e_i)].$$

Note that, for every monotone Boolean function f, $\mathrm{err}(f, x_i) = 1/2(1 - L^f(x_i))$. The *average sensitivity* of f, denoted by $s(f)$, is defined as

$$s(f) = \sum_{i=1}^{n} L^f(x_i).$$

For example, if $f = x_1$ then $L^f(x_1) = 1$, $L^f(x_i) = 0$ for every $i \neq 1$ and $s(f) = 1$. Theorem 1 implies $\mathrm{err}(f, \mathrm{MAJ}_n) = \frac{1}{2} - \Theta(1/\sqrt{n}) = \frac{1}{2} - \Theta(s(f)/\sqrt{n})$. If $f = \mathrm{MAJ}_n$ then $L^f(x_i) = \Theta(1/\sqrt{n})$ for every $1 \leq i \leq n$, $s(f) = \Theta(\sqrt{n})$ and $\mathrm{err}(f, \mathrm{MAJ}_n) = 0 = \frac{1}{2} - \Theta(s(f)/\sqrt{n})$. Here we consider another example. Let $f = t_1 \vee t_2 \vee \cdots \vee t_m$ be a monotone DNF formula of n variables such that each term t_i is a product of exactly l variables satisfying $(1-1/2^l)^m \sim 1/2$ and every variable appears exactly once in f. It is easy to see that f is almost fair and $L^f(x_i) = \Theta(\log n/n)$ for every $1 \leq i \leq n$ and so $s(f) = \Theta(\log n)$. A simple but tedious calculation shows $\mathrm{err}(f, \mathrm{MAJ}_n) = \frac{1}{2} - \Theta(\log n/\sqrt{n}) = \frac{1}{2} - \Theta(s(f)/\sqrt{n})$. These observations motivate us to divide our problem into two subproblems.

Problem 9 *Is it true that for every fair monotone Boolean function f on n variables*

$$err(f, MAJ_n) = \frac{1}{2} - \Omega\left(\frac{s(f)}{\sqrt{n}}\right)?$$

Problem 10 *Let f be a fair monotone Boolean function on n variables. Is it true that if, for every $1 \leq i \leq n$,*

$$err(f, x_i) = \frac{1}{2} - O\left(\frac{\log n}{\sqrt{n}}\right),$$

which is equivalent to the statement that $L^f(x_i) = O(\log n/\sqrt{n})$, then $s(f) = \Omega(\log n)$?

If answers to both of the problems are true, then the above algorithm achieves an optimal error of $1/2 - \Omega(\log n/\sqrt{n})$ for learning monotone Boolean functions under the uniform distribution. Problem 10 seems somewhat curious since it says "If the influence of every variable is small, then the sum of them is large." In the rest of this section, we prove that the answer to Problem 10 is true, even if we drop the condition that f is fair and weaken the condition $L^f(x_i) = O(\log n/\sqrt{n})$ to $L^f(x_i) = O(1/n^c)$ for some constant $c > 0$. Unfortunately, we could not solve Problem 9 at the time this paper was written.

To solve this problem we use ideas from Harmonic analysis. The Fourier transform of Boolean functions over the uniform distribution is defined as follows (see e.g. [4]). Given any subset $A \subseteq [n]$, the Fourier basis function $\chi_A : \{0,1\}^n \to \{-1,1\}$ is defined as $\chi_A(x) = (-1)^{|A \cap X|}$, where X is the subset of $[n]$ defined as $i \in X$ if and only if $x_i = 1$. It is well known that the 2^n basis functions χ_A form an orthonormal basis for the vector space of real valued functions on $\{0,1\}^n$. For a subset $A \subseteq [n]$, the Fourier coefficient of the basis χ_A, denoted by $\hat{f}_A$, is given by

$$\hat{f}_A = E[f\chi_A] = \frac{1}{2^n} \sum_x f(x)(-1)^{|A \cap X|}.$$

Note that any function f can be uniquely expressed as $f(x) = \sum_A \hat{f}_A \chi_A(x)$. Because of orthonormality, we have Parseval's equation : $\sum_A \hat{f}_A^2 = 1$.

If f is a monotone Boolean function then it is obvious that $L^f(x_i) = -\hat{f}_{\{i\}} \geq 0$ for every $1 \leq i \leq n$. For $1 \leq i \leq n$, we define function $f^i : \{0,1\}^n \to \{-1,0,1\}$ as

$$f^i(x) = \frac{1}{2}\left(f(x) - f(x \oplus e_i)\right).$$

It is easy to check that if $i \notin A$ then $\hat{f^i}_A = 0$ and if $i \in A$ then $\hat{f^i}_A = \hat{f}_A$. So we have $L^f(x_i) = \sum_x f^i(x)^2/2^n = \sum_A (\hat{f^i}_A)^2 = \sum_{A:i \in A} \hat{f}_A^2$ (see, e.g., [4, Lemma 1]). Summing this over all $1 \leq i \leq n$ we obtain

$$s(f) = \sum_{i=1}^{n} L^f(x_i) = \sum_A |A| \hat{f}_A^2. \tag{5}$$

Now we state our theorem.

Theorem 11 *Let f be a monotone Boolean function of n variables such that $\gamma \leq \Pr[f(x) = 1] \leq 1 - \gamma$ for some constant $\gamma > 0$. If there is a constant $c > 0$ such that $L^f(x_i) < 1/n^c$ for any $1 \leq i \leq n$, then $s(f) = \sum_{i=1}^n L^f(x_i) = \Omega(\log n)$.* $\square$

Note that the above theorem is the best possible in the sense that there exists a monotone Boolean function f such that $L^f(x_i) = \Theta(\log n/n)$ for every i and $s(f) = \Theta(\log n)$ [2].

The outline of the proof of Theorem 11 is as follows : Because of (5) and Parseval's equation ($\sum_A \hat{f}_A^2 = 1$), it is sufficient to show that $\sum_{A:|A| \geq \Omega(\log n)} \hat{f}_A^2$ is not negligible under the assumption that $L^f(x_i)$ is small for every i. To prove this, we use the idea of the proof of the theorem by Kahn et al. [8, Theorem 3.1] which says that the sum of squares of the individual influences is $\Omega((\log n)^2/n)$ for every monotone Boolean function f. To prove this theorem, they applied the inequalities of Beckner [1] to establish a slightly weaker version of our statement without the assumption that the influences are small. As in the proof of their theorem, the next lemma by Kahn et al. [8] is the key of the proof of our theorem.

Lemma 12 (Kahn, Kalai, Linial ([8], Lemma 3.4)) *Let g be a function from $\{0,1\}^n$ into $\{-1,0,1\}$. Let t be the probability that $g \neq 0$ and let $g = \sum_A \hat{g}_A \chi_A$ be the Fourier expansion of g. Then,*

$$t^{\frac{2}{1+\delta}} \geq \sum_A \delta^{|A|} \hat{g}_A^2$$

for every $0 \le \delta \le 1$. $\square$

Now we proceed to the proof of Theorem 11.

PROOF. (of Theorem 11) If

$$\sharp\left\{i \in [n] \;\middle|\; L^f(x_i) \ge \frac{1}{n^{1/2^\alpha + c'}}\right\} \ge n^{1/2^\alpha + c'} \log n$$

for some $\alpha > 0$ and for some $c' > 0$, then $s(f) = \sum_{i=1}^{n} L^f(x_i) \ge \log n$ holds. Thus, without loss of generality, we can assume that

$$\sharp\left\{i \in [n] \;\middle|\; L^f(x_i) \ge \frac{1}{n^{1/2^\alpha + c'}}\right\} < n^{1/2^\alpha + c'} \log n, \tag{6}$$

for any $\alpha > 0$ and for any $c' > 0$. Let $\beta = \lceil \log \frac{1}{c} \rceil$ and define the sets $S_1, S_2, \ldots, S_{\beta+1}$ as follows :

$$\begin{aligned} S_1 &= \left\{i \in [n] \;\middle|\; 0 \le L^f(x_i) \le \frac{1}{n^{1/2 + c'}}\right\}, \\ S_j &= \left\{i \in [n] \;\middle|\; \frac{1}{n^{1/2^{j-1} + c'}} < L^f(x_i) \le \frac{1}{n^{1/2^j + c'}}\right\}, \qquad \text{for } 2 \le j \le \beta, \\ S_{\beta+1} &= \left\{i \in [n] \;\middle|\; \frac{1}{n^{1/2^\beta + c'}} < L^f(x_i) < \frac{1}{n^c}\right\}. \end{aligned}$$

Since $0 \le L^f(x_i) < 1/n^c$ for any $1 \le i \le n$, the S_j's form a partition of the set $[n]$. Hence, from (6) we have

$$\begin{aligned} &\sum_{i=1}^{n} L^f(x_i)^2 \\ &< \left(\frac{1}{n^{1/2+c'}}\right)^2 \sharp S_1 + \sum_{j=2}^{\beta} \left(\frac{1}{n^{1/2^j + c'}}\right)^2 \sharp S_j + \left(\frac{1}{n^c}\right)^2 \sharp S_\beta \\ &< \left(\frac{1}{n^{1/2+c'}}\right)^2 n + \sum_{j=2}^{\beta} \left(\frac{1}{n^{1/2^j + c'}}\right)^2 n^{1/2^{j-1} + c'} \log n + \left(\frac{1}{n^c}\right)^2 n^{1/2^\beta + c'} \log n \\ &= n^{-2c'} + (\beta - 1) n^{-c'} \log n + n^{1/2^\beta - 2c + c'} \log n \end{aligned}$$

Let c' be a constant satisfying $1/2^\beta - 2c + c' \le -c'$. (The condition can be satisfied by putting $c' = c/2$.) Then we have

$$\sum_{i=1}^{n} L^f(x_i)^2 = O(n^{-c'} \log n) = O(n^{-c'/2}). \tag{7}$$

Now we apply Lemma 12 with $g = f^i$. Note that the probability that $f^i \neq 0$ is equal to $L^f(x_i)$. So we have

$$L^f(x_i)^{\frac{2}{1+\delta}} \geq \sum_A \delta^{|A|}(\hat{f^i}_A)^2 = \sum_{A: i \in A} \delta^{|A|} \hat{f}_A^2.$$

Summing this over $1 \leq i \leq n$, we have

$$\sum_{i=1}^{n} L^f(x_i)^{\frac{2}{1+\delta}} \geq \sum_A \delta^{|A|}|A|\hat{f}_A^2 \geq \delta^b \sum_{A:|A| \leq b} |A|\hat{f}_A^2 \geq \delta^b (\sum_{A:|A| \leq b} \hat{f}_A^2 - \hat{f}_\emptyset^2), \tag{8}$$

where b is a parameter which we will choose shortly. Since $\frac{2}{1+\delta} \leq 2$ and due to the monotonicity of r-th power averages, which says that for any n non-negative numbers $y_1, \ldots, y_n$ and for any $s \geq r \geq 1$,

$$\left(\frac{\sum_i y_i^r}{n}\right)^{\frac{1}{r}} \leq \left(\frac{\sum_i y_i^s}{n}\right)^{\frac{1}{s}}$$

holds (see, e.g., [7, p.26, Eq. (2.9.1)] or [8]), we have

$$\begin{aligned} \sum_{i=1}^{n} L^f(x_i)^{\frac{2}{1+\delta}} &\leq \left(n^{\frac{1+\delta}{2} - \frac{1}{2}} \left(\sum_{i=1}^{n} L^f(x_i)^2\right)^{\frac{1}{2}}\right)^{\frac{2}{1+\delta}} \\ &= \left(n^{\frac{\delta}{2}} \cdot O(n^{-\frac{c'}{4}})\right)^{\frac{2}{1+\delta}} = O\left(n^{\frac{2\delta - c'}{2(1+\delta)}}\right). \end{aligned} \tag{9}$$

Choose δ to be $c'/4$ and b to be $\frac{c'}{10}\frac{\log n}{\log(4/c')}$. From (8) and (9) we have

$$\begin{aligned} \sum_{A:|A| \leq b} \hat{f}_A^2 &\leq \left(\frac{1}{\delta}\right)^b \sum_{i=1}^{n} L^f(x_i)^{\frac{2}{1+\delta}} + \hat{f}_\emptyset^2 \leq \left(\frac{4}{c'}\right)^b O\left(n^{-c'/(4+c')}\right) + \hat{f}_\emptyset^2 \\ &\leq O\left(n^{c'/10 - c'/5}\right) + \hat{f}_\emptyset^2 \leq O\left(n^{-c'/10}\right) + \hat{f}_\emptyset^2 < c'', \end{aligned}$$

where c'' is some constant satisfying $c'' < 1$. The last inequality follows from the fact that $\hat{f}_\emptyset^2 = (2\Pr[f(x) = 1] - 1)^2 \leq (1 - 2\gamma)^2 < 1$. By Parseval's equation ($1 = \sum_A \hat{f}_A^2$), we have $\sum_{A:|A|>b} \hat{f}_A^2 = \Omega(1)$. Hence,

$$s(f) = \sum_A |A|\hat{f}_A^2 \geq \sum_{A:|A|>b} |A|\hat{f}_A^2 > b \sum_{A:|A|>b} \hat{f}_A^2 = \Omega(\log n) \cdot \Omega(1) = \Omega(\log n).$$

This completes the proof of the theorem. $\square$

4 Concluding Remarks

In the present paper, we proved that the single variable functions are the least correlated with the majority function among all fair monotone functions. We believe that the single variable functions are the only ones having this property, i.e., if a fair monotone function f satisfies err(f,MAJ_n) =err(x_1, MAJ_n), then f must be a single variable function. It would be interesting to prove this, or to find a counter-example to this.

All learning algorithms discussed in the paper have the same structure: Draw enough examples from EXAMPLE and output the most correlated function in a suitably chosen set of hypotheses $\mathcal{H}$ on examples. This paper proves that if we set $\mathcal{H}$ to $\{0, 1, \mathrm{MAJ}_n\}$, then the error of the algorithm is guaranteed to be $1/2 - \Omega(1/\sqrt{n})$. We believe that we can improve the performance of the algorithm by setting $\mathcal{H}$ to $\{0, 1, x_1, \ldots, x_n, \mathrm{MAJ}_n\}$. The problem of finding the smallest set $\mathcal{H}$ such that the error is shown to be at most $1/2 - \Omega(\log n/\sqrt{n})$ is also interesting. If such a set $\mathcal{H}$ consisting of a polynomial number of functions exists, then we will obtain an asymptotically optimal learning algorithm for monotone Boolean functions under the uniform distribution.

Acknowledgments

The authors would like to thank anonymous referees for their helpful comments to improve the presentation of the paper.

References

[1] W. Beckner, "Inequalities in Fourier Analysis," Annals of Mathematics, Vol. 102, pp. 159–182, 1975.

[2] M. Ben-Or, N. Linial, "Collective Coin Flipping, Robust Voting Games and Minima of Banzhaf Values," *Proc. 26th FOCS*, pp. 408–416, 1985.

[3] A. Blum, C. Burch and J. Langford, "On Learning Monotone Boolean Functions," *Proc. 39th FOCS*, pp. 408–415, 1998.

[4] N.H. Bshouty and C. Tamon, "On the Fourier Spectrum of Monotone Functions," *JACM*, Vol. 43, No. 4, pp. 747–770, 1996.

[5] D.E. Daykin, "A Simple Proof of the Kruskal-Katona Theorem," *J. Combinatorial Theory (A)*, Vol. 17, pp. 252–253, 1974.

[6] P. Frankl, "A New Short Proof for the Kruskal-Katona Theorem," *Discrete Mathematics*, Vol. 48, pp. 327–329, 1984.

[7] G.H. Hardy, J.E. Littlewood and G. Polya, *Inequalities (2nd edition)*, Cambridge University Press, 1952.

[8] J. Kahn, G. Kalai and N. Linial, "The influence of Variables on Boolean Functions (Extended Abstract)," *Proc. 29th FOCS*, pp. 68–80, 1988.

[9] M. Kearns, M. Li and L. Valiant, "Learning Boolean Formulas," *JACM*, Vol. 41, No. 6, pp. 1298–1328, 1994.

[10] L. Lovász, *Combinatorial Problems and Exercises*, North-Holland, Amsterdam, 1979.