1 Introduction

In [2], O’Donnell and Servedio show that any monotone function given by a depth-\(d\) decision tree can be learned to constant accuracy from random samples in polynomial time. The impact of this result is somewhat lessened by an apparent lack of interesting monotone functions given by low-depth decision trees. In particular, it has been conjectured that all such functions essentially depend on few variables.

**Conjecture 1.** For every \(\epsilon > 0\) and every monotone function \(f : \{0,1\}^n \to \{0,1\}\) given by a depth-\(d\) decision tree, there is a \(k\)-junta, \(g\) for \(k = \text{poly}_\epsilon(d)\) so that \(f\) and \(g\) agree on all but an \(\epsilon\)-fraction of inputs.

In this note, we disprove the above conjecture, and in particular provide an example of a monotone low-degree function that is not well approximated by any small junta. In particular we prove:

**Theorem 2.** There exists a constant \(\epsilon > 0\) so that for every positive integer \(d\), there exists a \(k = \exp(\Omega(\sqrt{d}))\) and a monotone function \(f : \{0,1\}^n \to \{0,1\}\) given by a depth-\(d\) decision tree, so that for every \(k\)-junta \(g\), \(f\) and \(g\) disagree on at least an \(\epsilon\)-fraction of inputs.

In fact it is not hard to show that the bound on \(k\) in Theorem 2 is tight up to the constant in the exponent. In particular, it is shown in [2] that any monotone function given by a depth-\(d\) decision tree has total influence \(I(f) = O(\sqrt{d})\). We combine this with the main result of [1], which says that any boolean function \(f\) can be \(\epsilon\)-approximated by a \(k\)-junta for \(k = \exp(O(I(f)/\epsilon))\). Combining these results we find that:
Corollary 3. If $f$ is a monotone function given by a depth-$d$ decision tree, and if $\epsilon > 0$, then there is a $k$-junta $g$ that agrees with $f$ on all but an $\epsilon$ fraction of coordinates for $k = \exp(O(\sqrt{d}/\epsilon))$.

By this result of [1], we know that any $f$ satisfying the conditions of Theorem 2 must not only have near the maximum possible total influence for a low-depth monotone function, but also must not be approximable by any function with much lower total influence. Because of this restriction, our construction will look somewhat similar to a construction of Talagrand in [3]. In particular, Talagrand constructs a monotone function $f$ on $\{0, 1\}^n$ so that on a constant fraction of inputs, $f$ has influence $\Omega(\sqrt{n})$. We note that since the total influence of $f$ must be $O(\sqrt{n})$, that this condition is equivalent to saying that for any subset $A \subseteq \{0, 1\}^n$ with $|A| = \Omega(2^n)$ that $\sum_{x \in A} \# \{i : f(x) \neq f(x^{\oplus i})\} = \Omega(|A|\sqrt{n})$, which is a strengthening of the condition that $f$ is not close to any function of small total influence.

2 The Construction

In order to define the function $f$ with the properties specified by Theorem 2, we first introduce some background notation. We let $d, t$ and $m$ be integers with $t = \Theta(\sqrt{d})$ and $m = \Theta(2^t)$. We furthermore assume that $2^{-t}m$ is sufficiently small given the value of $t/\sqrt{d}$. We let $\mathcal{S} = (S_1, \ldots, S_m)$ be a random sequence of sets, where the $S_i$ are chosen independently and uniformly from the set of subsets of $\{1, 2, \ldots, d-1\}$ of size exactly $t$. Given this $\mathcal{S}$, we define the function $T_{\mathcal{S}}$ on $\{0, 1\}^{d-1}$ as follows:

$$T_{\mathcal{S}}(x_1, \ldots, x_{d-1}) = \{1 \leq i \leq m : x_j = 1 \text{ for all } j \in S_i\}.$$

We will hereafter abbreviate $T$ by suppressing the explicit dependence on $\mathcal{S}$, and abbreviate $(x_1, \ldots, x_{d-1})$ by $x$.

We finally define $f$ as

$$f_{\mathcal{S}}(x_1, \ldots, x_{d-1}, y_1, \ldots, y_m) = \begin{cases} 1 & \text{if } |T(x)| \geq 2 \\ 0 & \text{if } |T(x)| = 0 \\ y_i & \text{if } T(x) = \{i\} \end{cases}$$

Again, we will often suppress the dependence of $f$ on $\mathcal{S}$. It is clear that $f$ is monotone. Furthermore, $f$ is given by a depth-$d$ decision tree, since after fixing the values of the $x_i$, the value of $f$ depends on at most one more coordinate. In the next Section, we show that $f$ cannot be approximated by any $k$-junta for small $k$. 

2
3 Approximation Bounds

Theorem 2 will follow from the following Proposition:

**Proposition 4.** There exists and \( \epsilon > 0 \) so that for \( f_S \) defined as above, with constant probability over the choice of \( S \), \( f \) is not \( \epsilon \)-approximated by any \( k \)-junta for \( k = o(2^t) \).

Before we begin the proof, we will need one Lemma

**Lemma 5.** With \( T \) as above,

\[ \Pr_{S, x}(|T_S(x)| = 1) = \Omega(1). \]

*Proof.* We will show the further claim that

\[ E [ |T_S(x)| (2 - |T_S(x)|) ] = \Omega(1). \quad (1) \]

Since the term in the expectation is positive only if \( |T| = 1 \), this will complete our proof. We note that

\[
E [ |T_S(x)| ] = \sum_{i=1}^{m} \Pr(i \in T_S(x)) \\
= \sum_{i=1}^{m} \Pr(x_j = 1 \text{ for all } j \in S_i) \\
= m2^{-t}.
\]

On the other hand, we have that

\[
E [ |T_S(x)| (|T_S(x)| - 1) ] = \sum_{i \neq j} \Pr(i, j \in T_S(x)) \\
= \sum_{i \neq j} \Pr(i \in T_S(x)) \Pr(j \in T_S(x) | i \in T_S(x)) \\
= \sum_{i \neq j} 2^{-t} \Pr(x_\ell = 1 \text{ for all } \ell \in S_j | x_\ell = 1 \text{ for all } \ell \in S_i).
\]

To compute this conditional probability we let \( S_j = \{a_1, \ldots, a_t\} \) where the \( a_i \) are picked randomly from \( \{1, 2, \ldots, d - 1\} \) without replacement. After fixing the values of \( S_i, a_1, \ldots, a_{r-1} \) and conditioning on the event that \( x_\ell = 1 \) for \( \ell \in S_i \) and \( x_{a_1} = \cdots = x_{a_{r-1}} = 1 \), we compute the probability that

3
$x_{ar} = 1$. This probability is clearly $1/2$ if $a_r \notin S_i$ and 1 if $a_r \in S_i$. Thus the probability that $x_{ar} = 1$ is

$$(1 + \Pr(a_r \in S_i))/2 = \left(1 + \frac{|S_i \setminus \{a_1, \ldots, a_{r-1}\}|}{d-r}\right)/2 = (1/2 + O(t/d)).$$

Hence the probability that $j \in T_S(x)$ given that $i \in T_S(x)$ is

$$(1/2 + O(t/d))^t = 2^{-t} \exp(O(t^2/d)) = O(2^{-t}).$$

Therefore, we have that

$$\E[|T_S(x)|/|T_S(x)| - 1)] \leq \sum_{i \neq j} 2^{-2t} \exp(O(t^2/d)) \leq (2^{-t}m)^2 \exp(O(t^2/d)).$$

Therefore, we have that

$$\E[|T_S(x)|/(|T_S(x)| - 1)] = \sum_{i \neq j} 2^{-2t} \exp(O(t^2/d)) \leq (2^{-t}m)^2 \exp(O(t^2/d)).$$

As long as $2^{-t}m$ is bounded below by a constant and above by $\exp(-O(t^2/d))/2$, this is $\Omega(1)$.

We are now ready to prove Proposition 4. By Lemma 5, we note that with constant probability over $S$, that $\Pr_x(|T_S(x)| = 1) = \Omega(1)$. For such $S$, we claim that $f$ has the desired property. In particular we claim the following:

**Lemma 6.** If $f$ is as above and $g$ is a $k$-junta, then

$$\Pr(f(x,y) \neq g(x,y)) \geq \frac{\Pr_x(|T(x)| = 1) - k2^{-t}}{2}.$$

**Proof.** This follows from the simple observation that after fixing the value of $x$ that if $T = \{i\}$ and $g$ does not depend on $y_i$ that $\Pr_y(f(x,y) \neq g(x,y)) = 1/2$. This is because after further conditioning on the values of all $y_j$ for $j \neq i$, $g$ becomes a constant function (by assumption) and $f$ takes the values
0 and 1 each with probability 1/2. Therefore we have that

\[
\Pr(f(x, y) \neq g(x, y)) \geq \Pr(T(x) = \{i\} \text{ and } g \text{ does not depend on } y_i) \\
= \frac{\Pr(|T(x)| = 1) - \Pr(T(x) = \{i\} \text{ and } g \text{ depends on } y_i)}{2} \\
= \frac{\Pr(|T(x)| = 1) - \sum_{i: g \text{ depends on } y_i} \Pr(T(x) = \{i\})}{2} \\
\geq \frac{\Pr(|T(x)| = 1) - \sum_{i: g \text{ depends on } y_i} \Pr(i \in T(x))}{2} \\
\geq \frac{\Pr(|T(x)| = 1) - \sum_{i: g \text{ depends on } y_i} 2^{-t}}{2} \\
\geq \frac{\Pr_x(|T(x)| = 1) - k2^{-t}}{2}.
\]

\[\square\]

Proposition 4 and Theorem 2 now follow immediately.

Acknowledgements

I would like to thank Ryan O’Donnell for making me aware of this problem, and for his help with finding appropriate references for this paper. This research was done with the support of an NSF postdoctoral fellowship.

References

[1] E. Friedgut *Boolean Functions with Low Average Sensitivity Depend on Few coordinates*, Combinatorica Vol. 18(1), pp. 27-36, 1998.

[2] R. O’Donnell, R. Servedio *Learning monotone decision trees in polynomial time*, SIAM Journal on Computing Vol. 37(3), pp. 827-844, 2007.

[3] M. Talagrand *How much are increasing sets positively correlated?*, Combinatorica Vol. 16(2), pp. 243-258, 1996.