

Learning and Lower Bounds for AC^0 with Threshold Gates

Parikshit Gopalan¹ and Rocco A. Servedio^{2*}

¹ Microsoft Research Silicon Valley, Mountain View, CA, 94203, U.S.A.
parik@microsoft.com

² Columbia University, New York, NY, 10027, U.S.A.
rocco@cs.columbia.edu

Abstract. In 2002 Jackson et al. [JKS02] asked whether AC^0 circuits augmented with a threshold gate at the output can be efficiently learned from uniform random examples. We answer this question affirmatively by showing that such circuits have fairly strong Fourier concentration; hence the low-degree algorithm of Linial, Mansour and Nisan [LMN93] learns such circuits in sub-exponential time. Under a conjecture of Gotsman and Linial [GL94] which upper bounds the total influence of low-degree polynomial threshold functions, the running time is quasi-polynomial. Our results extend to AC^0 circuits augmented with a small super-constant number of threshold gates at arbitrary locations in the circuit. We also establish some new structural properties of AC^0 circuits augmented with threshold gates, which allow us to prove a range of separation results and lower bounds.

Keywords: computational learning theory, AC^0 , Fourier concentration, threshold gates, polynomial threshold functions

1 Introduction

The seminal result of Linial, Mansour and Nisan [LMN93] showed how to learn the class AC^0 of constant depth circuits in quasi-polynomial time under the uniform distribution with random examples. Their work introduced the *Low-Degree Algorithm* which can learn functions where the Fourier spectrum is concentrated on low-degree coefficients; this algorithm and its extensions have since found many applications in learning, see e.g. [FJS91,BT96] for some early work and [JKS02,KOS04,MOS04,OS07,BOW08,KKMS08,KOS08] for more recent results.

In the two decades since their work, despite much effort, there has been limited progress in designing learning algorithms for more expressive circuit classes. Circuit classes like AC^0 with parity gates ($AC^0[2]$) and depth-2 circuits of arbitrary threshold gates remain beyond the reach of currently known algorithms. One obstacle is that there are no lower bounds known for some of these classes,

* Supported by NSF grants CCF-0347282, CCF-0523664 and CNS-0716245, and by DARPA award HR0011-08-1-0069.

such as depth-2 TC^0 , and the existence of lower bounds seems to be a prerequisite for any learning algorithm (see [FK09]). Devising learning algorithms and lower bound techniques that can handle more powerful classes of circuits is a central open problem at the intersection of computational learning theory and circuit lower bounds.

Jackson et al. made some progress on learning circuits more expressive than AC^0 in [JKS02]. They gave a quasipolynomial-time algorithm that can learn Majority-of- AC^0 circuits – polynomial-size, constant-depth circuits augmented with a single Majority gate at the output – under the uniform distribution. Using a result of [Bei94], this yields a quasipolynomial-time algorithm that can learn AC^0 circuits augmented with $\text{polylog}(n)$ many Majority gates at arbitrary locations in the circuit. The algorithm of Jackson et al. uses the low-degree algorithm as a weak learner and combines it with boosting. [JKS02] posed as an open question whether any efficient algorithm can learn Threshold-of- AC^0 circuits, in which the topmost gate is a threshold gate (i.e. a weighted majority in which the weights may be arbitrary). It is observed in [JKS02] via an explicit counterexample that the analysis of their boosting-based algorithm breaks down for Threshold-of- AC^0 . In this work, we take a significant step towards answering the question of [JKS02].

AC^0 circuits augmented with a few threshold gates have been well studied in the complexity theory literature, see e.g. [ABFR94, Bei94, GHR92, Gol97, Han07]. This is a natural class of circuits lying between the classes AC^0 (which we understand well) and TC^0 (for which we do not know lower bounds). One focus of this work has been on understanding the difference in power between unweighted threshold gates (i.e. majorities) versus threshold gates with arbitrary weights. Aspnes et al. [ABFR94] prove that any AC^0 circuit with a single threshold gate at the top cannot compute (or even approximate) parity, and more recently Hansen [Han07] has established $n^{\Omega(\log n)}$ size bounds on AC^0 circuits augmented with up to $\epsilon \log^2 n$ threshold gates. In contrast, when we restrict ourselves to Majority gates, an elegant result of Beigel [Bei94] alluded to above shows that any polynomial-size AC^0 circuit with $\text{polylog}(n)$ Majority gates is equivalent to a quasi-polynomial size AC^0 circuit with a single majority gate at the top, and lower bounds for such circuits follow from [ABFR94].

1.1 Our Results

We show that AC^0 circuits augmented with a few threshold gates with arbitrary weights can be learned in subexponential time under the uniform distribution. In doing this we establish some new structural properties of such circuits, which allow us to prove new lower bounds and separations for such circuits.

Learning AC^0 with threshold gates. Our first main result is a Fourier concentration bound for Threshold-of- AC^0 circuits: roughly speaking, this bound says that any size- M , constant-depth Threshold-of- AC^0 circuit C must satisfy

$$\sum_{|\alpha| > t} \widehat{C}(\alpha)^2 \leq \epsilon \quad \text{for} \quad t = \frac{(\log M)^{\Theta(d)} 2^{\Theta((\log M)^{2/3})}}{\epsilon^{(\log M)^{1/3}}}.$$

This can be viewed as a natural extension of the [ABFR94] result showing that Threshold-of- AC^0 cannot compute parity; we show that such circuits in fact exhibit strong Fourier concentration. (Thus, roughly speaking, our result is to [ABFR94] as the [LMN93] Fourier concentration bound for AC^0 is to the earlier AC^0 lower bounds of Håstad [Hås86].) We note that Fourier concentration bounds of the sort we establish were not known even for Majority-of- AC^0 prior to this work; the [JKS02] algorithm requires boosting and its analysis does not establish Fourier concentration.

With our Fourier concentration bound for Threshold-of- AC^0 in hand, applying the Low-Degree Algorithm of [LMN93] we get the first subexponential-time learning result for this class: any size- M , constant-depth Threshold-of- AC^0 can be learned to any constant accuracy ϵ in time $n^{2^{\Theta((\log M)^{2/3})}}$.

An important ingredient in our proof is a recent $2^{O(d)}n^{1-1/O(d)}$ upper bound on the total influence of degree- d polynomial threshold functions over n Boolean variables, proved recently by [HKM09] and [DRST09]. In 1994 Gotsman and Linial [GL94] conjectured a stronger bound, that every degree- d PTF has total influence $O(d\sqrt{n})$. We show that under the [GL94] conjecture our results become significantly stronger: every size- M depth- d Threshold-of- AC^0 circuit C has Fourier concentration

$$\sum_{|\alpha| > t} \widehat{C}(\alpha)^2 \leq \epsilon \quad \text{for} \quad t = \frac{2^{O(d)}(\log M)^d}{\epsilon^2}$$

and consequently such circuits can be learned to constant accuracy in time $n^{2^{O(d)}(\log M)^d}$.

We extend the above results by giving Fourier concentration and learning results for AC^0 circuits with r threshold gates in arbitrary locations in the circuit. We unconditionally learn such circuits with $r = O((\log M)^{1/3})$ many threshold gates, to any constant accuracy, in time $n^{2^{\Theta((\log M)^{2/3})}}$. Assuming the [GL94] conjecture, we learn such circuits with $r = O(\log \log M)$ to any constant accuracy in time $n^{2^{O(d)}(\log M)^{O(d)}}$. These results are achieved building on our results for Threshold-of- AC^0 .

Lower bounds and separation results. To complement the positive (learning) results described above, in Section 6 we establish new lower bounds and separation results for AC^0 circuits augmented with threshold gates. These results separate the classes Majority-of- AC^0 and Threshold-of- AC^0 and highlight some interesting contrasts between them.

1. Since Majority-of- AC^0 is already known to be learnable in quasi-polynomial time, our learning results are only of interest if Threshold-of- AC^0 is actually a broader class than Majority-of- AC^0 . We show that this is indeed the case, by exhibiting a single threshold gate for which any equivalent depth- d Majority-of- AC^0 circuit must have size $2^{\Omega(n^{1/(d-1)})}$. (See Section 6.1.)
2. Beigel [Bei94] showed that any size- s , depth- d circuit that contains m Majority gates is computed by a size- $2^{m(O(\log s))^{2d+1}}$, depth- $(d+2)$ circuit with

a single Majority gate at the root. We show that this size bound cannot be improved to polynomial, by showing that a simple AND of two Majority gates requires any constant-depth circuit with a single Majority gate at the top (or even an arbitrary Threshold gate at the top) to have $n^{\Omega_d(\log n)}$ size. (See Section 6.2.)

3. A natural question is whether Beigel’s result can be extended from Majority gates to arbitrary Threshold gates. Perhaps every AC^0 circuit which contains $\text{polylog}(n)$ many Threshold gates is equivalent to a $\text{quasipoly}(n)$ -size Threshold-of- AC^0 ? In fact the answer is no: we show that no analogue of Beigel’s result is possible for Threshold gates, by showing that any Threshold-of- AC^0 circuit that computes the AND of two (high-weight) Threshold gates must have exponential size. (See Section 6.3.)
4. We also give lower bounds for AC^0 circuits with relatively many Threshold gates. We prove that any AC^0 circuit with $\epsilon \log n$ Threshold gates cannot compute parity, for a small constant $\epsilon > 0$. Previously, Aspnes et al. [ABFR94] proved this claim for AC^0 with a single threshold gate at the top. Beigel [Bei94] showed that any AC^0 circuit must be augmented with $n^{\Omega(1)}$ many Majority gates in order to compute parity. Our bound allows for a smaller number of gates augmenting the basic AC^0 circuit, but the gates (Threshold instead of Majority) are more powerful. (See Section 6.4.)

We note that the previous lower bounds on Threshold-of- AC^0 due to [ABFR94] apply to functions which have high PTF degree. This approach cannot be used for results (1) and (2) above, where we are proving lower bounds against functions which have low PTF degree. As mentioned above, using different techniques Hansen [Han07] has established $n^{\Omega(\log n)}$ size bounds on AC^0 circuits augmented with up to $\epsilon \log^2 n$ threshold gates. Very recently V. Podolskii [Pod10] has shown that any AC^0 circuit augmented with $O(\log n)$ threshold gates that approximates the parity function to high accuracy must have exponential size.

2 Preliminaries

2.1 MAC^0 and TAC^0 and $\text{TAC}^0[r]$

Recall that a *threshold function*, or halfspace, over n variables is a Boolean function $h : \{-1, 1\}^n \rightarrow \{-1, 1\}$, $h(x) = \text{sign}(\sum_{i=1}^n w_i x_i - \theta)$, where w_1, \dots, w_n, θ may be arbitrary real values. We will sometimes write Thr to denote a single threshold gate and Maj to denote a single Majority gate, where the Majority function is the threshold function for which each w_i equals 1 and the threshold θ equals 0.

A *Threshold-of- AC^0 circuit*, or TAC^0 , is a circuit consisting of a threshold function (with arbitrary weights and fanin) as the output gate and AC^0 circuits feeding into it. A depth- d TAC^0 is one in which each of the AC^0 circuits feeding into the output threshold gate has depth at most $d - 1$. The size of a TAC^0 is the total number of gates (so in particular, in a size- M TAC^0 each of the AC^0 circuits is of size at most M).

A *Majority-of- AC^0 circuit*, or MAC^0 , is a TAC^0 in which the top threshold function is a majority gate.

Finally, we will also consider AC^0 circuits that have r arbitrary Thr gates buried at arbitrary locations in the circuit; we refer to such a circuit as a “Threshold-of- r - AC^0 s”, or $TAC^0[r]$.

We give standard definitions of polynomial threshold functions, influence of variables on Boolean functions, noise sensitivity, and the basics of Fourier analysis in the full version of the paper [GS10].

2.2 Random Restrictions and AC^0

We write “ $\rho \sim \mathcal{R}_p$ ” to indicate that ρ is a *random restriction with parameter p* . Such a restriction ρ is chosen by independently fixing each variable to $+1$ or -1 each with probability $\frac{1-p}{2}$, and leaving the variable unfixed with probability p . We write f_ρ to denote the function that results from applying ρ to f .

We will use several facts about the behavior of AC^0 circuits under random restrictions. The first of these facts is Håstad’s Switching Lemma:

Lemma 1 ([Hås86]). *Let C be a depth-2 circuit (i.e. a DNF or a CNF) of bottom fan-in s . Then $\Pr_\rho[C_\rho \text{ cannot be written as a depth-}t \text{ decision tree}] \leq (5ps)^t$, where ρ is a random restriction with parameter p .*

(The above statement is implicit in [Hås86] and is made explicit in e.g. [Hås01].) Repeated applications of the Switching Lemma can be used to prove the following in a rather straightforward way:

Lemma 2 ([LMN93], Lemma 2). *If C is a size- M depth- d AC^0 circuit, then for any $t \geq 0$ we have $\Pr_\rho[C_\rho \text{ cannot be written as a depth-}t \text{ decision tree}] \leq M2^{-t}$, where ρ is a random restriction with parameter $p = \frac{1}{10^{d+t-1}}$.*

([LMN93] actually state a slightly weaker form in which the LHS is replaced by “ $\Pr_\rho[\deg(C_\rho) > t]$.” It is easy to check that using Lemma 1, the [LMN93] proof directly yields Lemma 2 as stated above.)

2.3 Sketch of the Random Restriction Argument.

The high-level idea of our proof is quite simple, and is similar to that of [LMN93]. We show that when a TAC^0 is hit with a random restriction, with high probability it collapses into a “much simpler function,” specifically a low-degree PTF. Recent results on the Fourier concentration of low-degree PTFs due to [DRST09,HKM09] let us infer that the original TAC^0 must also have had good Fourier concentration. In the rest of this section we elaborate on this argument.

We begin by recalling the basic outline of [LMN93]’s Fourier concentration bound for AC^0 circuits. It will be useful for us to view the [LMN93] argument as proceeding in two stages:

1. The first stage analyzes what happens to a size- M , depth- d AC^0 circuit C when it is hit with a random restriction with parameter $p \approx \frac{1}{(\log M)^{d-1}}$ (recall that p is the probability that a variable “survives” the restriction, i.e. is left unfixed). [LMN93] show that with high probability such a restriction causes C_ρ to collapse down to a $(\log M)$ -depth decision tree.
2. The second stage is the observation that a $(\log M)$ -depth decision tree T , being a degree $\log M$ polynomial, has extremely strong Fourier concentration: $\sum_{|\alpha| > \log M} \hat{T}(\alpha)^2 = 0$. Linial et al. then use the Fourier concentration of C_ρ to argue that the original AC^0 function computed by C must have had most of its Fourier weight at levels $\leq (\log M)^d$.

Our argument for TAC^0 has a similar high-level structure, but with some significant differences in both stages. Let C now denote a size- M , depth- d TAC^0 circuit.

- 1'. In the first stage, we consider hitting C with a “stronger” random restriction with a smaller value of p (so fewer variables survive the restriction). We show that with high probability such a restriction causes C_ρ to collapse down to a “low-degree” PTF of degree $k \ll \log M$. The stronger restriction is necessary since the results of [DRST09,HKM09] are non-trivial only when the degree of the PTF is $o(\sqrt{\log n})$.
- 2'. The results of [DRST09,HKM09] imply that C_ρ must have some nontrivial Fourier concentration. The Fourier concentration for C_ρ is much weaker than what one gets for decision trees, but one can adapt the original [LMN93] argument to show that the original circuit C itself must have had some Fourier concentration.

The conjecture of Gotsman & Linial significantly strengthens the bounds on total influence and noise sensitivity of low-degree PTFs that are currently known; it implies non-trivial bounds as long as the degree is $o(\sqrt{n})$. This in turn strengthens the Fourier concentration that we get for C_ρ in Stage 2', and hence also for C . We present each of the stages of the above argument in as self-contained a way as possible in Section 3. Section 4 puts the pieces together to prove the main results.

3 Random Restrictions of TAC^0 .

3.1 Stage 1: Collapsing TAC^0 to a low-degree PTF

In this section we prove the following:

Lemma 3. *Let C be a size- M , depth- d TAC^0 . Let ρ be a random restriction with parameter p (specified below) and let $k \geq 1$. Then for any $0 < p' < 1$, with failure probability at most δ the function C_ρ is a degree- k PTF, where*

$$\delta = M^{-2} + M^5 \left(\frac{4e \log(M)p'}{k} \right)^k \quad \text{and} \quad p = \frac{1}{10^{d-1}(4 \log M)^{d-2}} \cdot p'.$$

Proof. The proof is conceptually quite simple. Let $C = \text{Thr}(C_1, \dots, C_\ell)$ where Thr is the topmost threshold gate, $\ell \leq M$ is its fan-in, and each C_i is an AC^0 circuit of depth at most $d-1$ and size M_i , where $M_1, \dots, M_\ell \leq M$. We view the restriction ρ as being obtained in two steps. The first step collapses each C_i to a decision tree of depth $O(\log M)$. The second step significantly reduces the depth of each decision tree, down to k . After these two steps, with high probability each C_i has collapsed down to $(C_i)_\rho$ which is a degree- k polynomial. Thus C_ρ is a PTF of degree k .

In the first step we take a random restriction ρ_1 with parameter $p_1 = \frac{1}{10^{d-1}(4 \log M)^{d-2}}$. For a given i , Lemma 2 gives that with failure probability at most $M_i \cdot M^{-4}$, the function $(C_i)_{\rho_1}$ is equivalent to a decision tree T_i of depth $4 \log M$. Summing failure probabilities over all $i = 1, \dots, \ell$, this occurs for every C_i with overall failure probability at most $(M_1 + \dots + M_\ell)M^{-4} \leq M^{-2}$.

In the second step, we take a random restriction with parameter p' (thus the overall probability that a variable survives the combined restriction is $p = p_1 p'$ as desired). The following simple lemma analyzes the effect of a random restriction on a depth- t decision tree:

Lemma 4. *Let T be a depth- t decision tree and ρ be a random restriction with parameter p' . Then for $k \geq 1$, we have $\Pr[T_\rho \text{ cannot be written as a depth-}k \text{ decision tree}] \leq 2^t ((etp')/k)^k$.*

Proof. Suppose that under ρ at most k variables survive in each root-to-leaf path in T . Then it is clear that T_ρ can be written as a decision tree of depth at most k . So fix any given path of length at most t in T ; wlog the variables appearing on this path are x_1, \dots, x_t . The probability that at least k of these variables survive ρ is at most

$$\binom{t}{k} (p')^k \leq \left(\frac{et}{k}\right)^k (p')^k = \left(\frac{etp'}{k}\right)^k.$$

A union bound over all (at most 2^t) paths in T finishes the proof.

We apply this lemma to each of the $\ell \leq M$ decision trees T_i from step 1, taking $t = 4 \log M$. A union bound gives that the probability that any T_i fails to have its depth reduced to k is at most $M \cdot 2^t \cdot (etp'/k)^k$. Any decision tree of depth k is exactly computed by a Fourier polynomial of degree at most k ; the top-level Thr gate takes the sign of a weighted sum of these polynomials, and we obtain Lemma 3.

3.2 Stage 2: From Fourier concentration of C_ρ to Fourier concentration of C

We will use the following recent bound on the noise sensitivity of degree- k PTFs due to Diakonikolas et al. [DRST09] and Harsha et al. [HKM09]:

Theorem 1. *For any degree- k PTF f over $\{-1, 1\}^n$ and any $0 \leq \epsilon \leq 1$, we have $\text{ns}_\epsilon(f) \leq 2^{O(k)} \cdot \epsilon^{\frac{1}{O(k)}}$.*

The following simple result (Corollary 17 of [KOS04]) converts noise sensitivity upper bounds to Fourier concentration bounds:

Lemma 5. *Let $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$ be any Boolean function and let $\kappa : [0, 1/2] \rightarrow \mathbf{R}^+$ be an increasing function such that $\text{ns}_\epsilon(f) \leq \kappa(\epsilon)$. Then*

$$\sum_{|\alpha| \geq m} \widehat{f}(\alpha)^2 \leq \epsilon \quad \text{for } m = \frac{1}{\kappa^{-1}(\epsilon/2.32)}.$$

Plugging in Theorem 1 gives the following Fourier concentration bound:

Corollary 1. *For any degree- k PTF f over $\{-1, 1\}^n$ and any $0 \leq \epsilon \leq 1$, we have*

$$\sum_{|\alpha| \geq m(\epsilon)} \widehat{f}(\alpha)^2 \leq \epsilon \quad \text{where } m(\epsilon) = \frac{2^{\Theta(k^2)}}{\epsilon^{\Theta(k)}}.$$

We now show that if f_ρ has good Fourier concentration (w.h.p. over the choice of random restriction ρ), then f itself has good Fourier concentration. This is done by the following lemma, adapting arguments from [LMN93]; the proof is in the full version [GS10].

Lemma 6. *Let $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$ and let t, p be parameters such that $pt > 8$. Then*

$$\sum_{|\alpha| > t} \widehat{f}(\alpha)^2 \leq 2\mathbf{E}_\rho \left[\sum_{|\beta| > pt/2} \widehat{f}_\rho(\beta)^2 \right],$$

where ρ is a random restriction with parameter p .

As an easy corollary of Lemma 6 we have the following:

Corollary 2. *Let $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$ and let t, p be parameters such that $tp > 8$. Suppose that with probability at least $1 - \delta$ (over the choice of a random restriction ρ with parameter p) the function f_ρ has Fourier concentration $\sum_{|\beta| > pt/2} \widehat{f}_\rho(\beta)^2 \leq \epsilon$. Then we have $\sum_{|\alpha| > t} \widehat{f}(\alpha)^2 \leq 2\epsilon + 2\delta$.*

(This follows from the lemma because \widehat{f}_ρ is a Boolean function and consequently always has total Fourier weight at most 1.)

4 Proof of the Fourier concentration results for TAC^0

Throughout this section C is a size- M , depth- d TAC^0 . The regime we are most interested in is when the circuit size M is $\text{poly}(n)$ and the error parameter ϵ is something like a small constant; in particular, we are most interested in situations where $\epsilon > M^{-1}$. (We note that even the Majority function has $\widehat{f}([n])^2 = \Theta(1/n)$, so Fourier concentration bounds for TAC^0 must certainly be vacuous for $\epsilon < 1/n$.)

4.1 The unconditional result

Putting together all the pieces, we have established a Fourier concentration bound for TAC^0 :

Theorem 2. *Let C be a size- M , depth- d TAC^0 . Let $\epsilon \geq 2M^{-2}$. Then C has Fourier concentration*

$$\sum_{|\alpha| > t} \widehat{C}(\alpha)^2 \leq 4\epsilon \quad \text{for} \quad t = \frac{(\log M)^{\Theta(d)} \cdot 2^{\Theta((\log M)^{2/3})}}{\epsilon^{\Theta((\log M)^{1/3})}}. \quad (1)$$

Proof. In Stage 1 we shall take (with foresight) $k = (\log M)^{1/3}$ and $p' = \frac{k}{4eM^{7/k} \log M}$, so consequently $p = \frac{k}{(40 \log M)^{d-1} \cdot e \cdot M^{7/k}}$. This choice of parameters gives failure probability at most $\delta = 2M^{-2}$ in Lemma 3, so with this failure probability we have that C_ρ is a degree- k PTF which satisfies

$$\sum_{\alpha \geq m} \widehat{f}(\alpha)^2 \leq \epsilon \quad \text{where} \quad m = \frac{2^{\Theta((\log M)^{2/3})}}{\epsilon^{\Theta((\log M)^{1/3})}}.$$

In Step 3, we take $t = 2m/p$ so $tp/2 = m$ which is at least 8. Corollary 2 thus gives us

$$\sum_{|\alpha| > t} \widehat{C}(\alpha)^2 \leq 2\epsilon + 2\delta \quad \text{where} \quad t = \frac{(\log M)^{\Theta(d)} \cdot 2^{\Theta((\log M)^{2/3})}}{\epsilon^{\Theta((\log M)^{1/3})}}.$$

Applying the well-known [LMN93] machinery for uniform distribution learning of Boolean functions with good Fourier concentration, we get the following:

Corollary 3. *Size- M depth- d TAC^0 circuits can be learned to accuracy ϵ in time n^t where*

$$t = \frac{(\log M)^{\Theta(d)} \cdot 2^{\Theta((\log M)^{2/3})}}{\epsilon^{\Theta((\log M)^{1/3})}}.$$

Thus as long as $\epsilon \geq 1/2^{O((\log M)^{1/3})}$ and $d \leq O((\log M)^{2/3}/(\log \log M))$ this gives an algorithm to learn size- M depth- d TAC^0 in time $n^{2^{\Theta((\log M)^{2/3})}}$, i.e. sub-exponential time ($2^{n^{o(1)}}$) for any $M = \text{poly}(n)$.

4.2 The Gotsman-Linial Conjecture and its consequences

In 1994 Gotsman and Linial [GL94] asked the question of what is the maximum total influence of any degree- k PTF over n variables. They conjectured that the symmetric function which changes sign on the k middle layers of the Boolean hypercube has the highest total influence of any degree- k PTF (it is easy to see that this function is indeed a degree- k PTF). Since each layer of edges in the Boolean hypercube contains at most $\sqrt{n}2^{n-1}$ edges, a direct consequence of their conjecture (which is nearly equivalent to it for $k = o(\sqrt{n})$) is the following:

Conjecture 1 ([GL94]). Every degree- k PTF f over n variables has $\text{Inf}(f) \leq k\sqrt{n}$.

We show that using our approach, Conjecture 1 yields significantly improved Fourier concentration (and significantly more efficient learnability) for TAC^0 . The noise sensitivity bounds of [DRST09] and [HKM09] follow from a bound of $2^{O(k)}n^{1-1/O(k)}$ on the average sensitivity of degree- k PTFs. This bound becomes trivial for $k = \Omega(\sqrt{\log n})$, and hence we needed to use a very strong random restriction in order to reduce our initial TAC^0 to a PTF of degree $o(\sqrt{\log n})$. Conjecture 1 implies that a weaker random restriction will suffice. We use the following noise sensitivity and Fourier concentration consequences of the Gotsman-Linial conjecture:

Corollary 4. *If Conjecture 1 holds, then for any degree k PTF f over $\{-1, 1\}^n$ and any $0 \leq \epsilon \leq 1$,*

$$\text{ns}_\epsilon(f) \leq 2k\sqrt{\epsilon} \quad \text{and} \quad \sum_{|\alpha| \geq m} \widehat{f}(\alpha)^2 \leq \epsilon \quad \text{where} \quad m = \frac{24k^2}{\epsilon^2}.$$

The first inequality follows from the reduction from total influence to noise sensitivity for PTFs given in [DRST09] (see Section 7), and the second inequality then follows from Lemma 5. We thus obtain:

Theorem 3. *Let C be a size- M , depth- d TAC^0 . Let $\epsilon \geq 2M^{-2}$. If Conjecture 1 is true, then we have*

$$\sum_{|\alpha| > t} \widehat{C}(\alpha)^2 \leq 4\epsilon \quad \text{for} \quad t = \frac{2^{O(d)}(\log M)^d}{\epsilon^2}.$$

Proof. In Stage 1 we shall take $k = \log M$ and $p' = 10^{-4}$, so $p = \frac{1}{10^{d+3}(4 \log M)^{d-2}}$. Lemma 3 gives that with probability at least $1 - 2M^{-2}$, the function C_ρ is a degree- k PTF, in which case we have, for $m = 24k^2/\epsilon^2$, $\sum_{|\alpha| > m} \widehat{C}_\rho(\alpha)^2 \leq \epsilon$. For Stage 3, in Corollary 2 we take $t = \frac{2m}{p} = \frac{(c_1 \log M)^d}{\epsilon^2}$ for some absolute constant c_1 . Corollary 2 thus gives us $\sum_{|\alpha| > t} \widehat{C}(\alpha)^2 \leq 2\epsilon + 2\delta$.

Similar to before, the [LMN93] low-degree algorithm gives us:

Corollary 5. *If Conjecture 1 is true, then size- M , depth- d TAC^0 can be learned to accuracy ϵ in time*

$$n^{\frac{2^{O(d)}(\log M)^d}{\epsilon^2}}.$$

This gives quasi-polynomial time learning for $M = \text{poly}(n)$ -size TAC^0 for any constant (or even $1/\text{polylog}(n)$) accuracy ϵ .

5 Learning $\text{TAC}^0[r]$

Our learning results can be extended from TAC^0 circuits to $\text{TAC}^0[r]$ circuits for small (but superconstant) values of r . The high-level approach is as follows: We first prove a general result showing that if a class \mathcal{C} has Fourier concentration, then any R -junta-of-functions-from- \mathcal{C} must also have fairly good Fourier concentration provided that R is not too large. We then argue that any $\text{TAC}^0[r]$ is equivalent to a R -junta-of- TAC^0 for $R = (r + 1)2^r$. This lemma and the arguments used in its proof are similar to arguments found in [BRS95]. Combining the above two ingredients with the Fourier concentration bounds for TAC^0 which we obtained in Section 4, we get Fourier concentration bounds for $\text{TAC}^0[r]$.

Because of space limits here we only state the results and defer full proofs to [GS10]. We show unconditionally that $\text{TAC}^0[O((\log M)^{1/3})]$ circuits can be learned in essentially the same time bound that we achieved for unconditionally learning TAC^0 circuits:

Theorem 4. *The class of $\text{TAC}^0[O((\log M)^{1/3})]$ circuits of size M and depth d can be learned to accuracy ϵ (for $\epsilon > 2M^{-2}$) in time n^t , where*

$$t = (\log M)^{\Theta(d)} \cdot 2^{\Theta((\log M)^{2/3})} \epsilon^{-\Theta((\log M)^{1/3})}.$$

Assuming the Gotsman-Linial conjecture, we obtain

Theorem 5. *If Conjecture 1 is true, then the class of $\text{TAC}^0[r]$ circuits of size M and depth d can be learned to accuracy ϵ (for $\epsilon > 2M^{-2}$) in time n^t , where*

$$t = 2^{O(d+r)} \cdot (\log M)^{3d} \epsilon^{-3}.$$

For constant d this gives quasi-polynomial time learning for r as large as $O(\log \log M)$.

6 Lower bounds

6.1 MAC^0 cannot compute TAC^0

In this section we prove that there are TAC^0 circuits that have no small equivalent MAC^0 circuit.

Theorem 6. *There is a threshold function over $N = O(n^2)$ variables such that any equivalent MAC^0 circuit of depth $d \geq 2$, $d = \Theta(1)$ must have size $2^{\Omega(n^{1/(d-1)})}$.*

The desired function is the function $U_{n,4n}(x)$ defined by Goldmann *et al.* in Section 4 of [GHR92] (all variables below take values ± 1):

$$U_{n,4n}(x) = \text{sign}(2r_{n,4n}(x) + 1), \quad r_{n,4n}(x) = \sum_{i=0}^{n-1} \sum_{j=0}^{4n-1} 2^i x_{ij}. \quad (2)$$

It is clear that $U_{n,4n}$ is a TAC^0 circuit (of depth 1), consisting of a single threshold gate over $N = 4n^2$ input variables. It remains to show that any depth- d MAC^0

circuit for $U_{n,4n}(x)$ must be large. We do this in two steps as follows. Suppose that C is a depth- d , size- M MAC^0 circuit that computes $U_{n,4n}(x)$. If $M = 2^{\Omega(n^{1/(d-1)})}$ then there is nothing to show, so we assume $M = 2^{O(n^{1/(d-1)})}$. We shall consider the effect of applying a random restriction with parameter $r = \frac{1}{10^{d-1}s^{d-2}}$ to C , where we select $s = 3 \log M$. We establish the following two lemmas in [GS10]:

Lemma 7. *With probability at least $1 - M^{-1}$ over the random choice of ρ , the function $(U_{n,4n})_\rho$ is a polynomial threshold function of total weight at most M^7 .*

Lemma 8. *With probability at least $1 - 2n^{-2}$ over the random choice of ρ , the function $(U_{n,4n})_\rho$ has a sub-function (obtained by possibly fixing some additional variables in $(U_{n,4n})_\rho$) that is equivalent, up to renaming variables, to $U_{m,4m}$ where $m = \Omega(n/(\log M)^{d-2})$.*

Fix a restriction ρ that satisfies both Lemmas (such a ρ must exist since each of the two events has probability greater than $1/2$). The function $U_{m,4m}$ is a restriction of the function $(U_{n,4n})_\rho$ from Lemma 7, and thus $(U_{m,4m})_\rho$ must have a polynomial threshold function of weight at most M^7 . However, the discussion following Corollary 8 of [GHR92] shows that the total weight of any PTF for $U_{m,4m}$ must be at least $\Omega(2^{m/2}/\sqrt{m})$. Since $m = \Omega(n/(\log M)^{d-2})$, straightforward manipulation yields the desired lower bound $M = 2^{\Omega(n^{1/(d-1)})}$.

6.2 Lower bounds on MAC^0

Beigel [Bei94] showed that any size- s , depth- d circuit that contains m Maj gates is computed by a size- $2^{m(O(\log s))^{2d+1}}$, depth- $(d+2)$ circuit with a single Maj gate at the root. It is natural to ask whether this simulation can be improved to a polynomial-size (rather than quasi-polynomial) Maj of AC^0 . In this section we observe that no such strengthened version of Beigel's theorem can exist, by proving that there is no polynomial-size MAC^0 (or even TAC^0) for an AND of two Maj gates:

Theorem 7. *For any constant d , any TAC^0 circuit of depth d that computes $f(x, y) = \text{Maj}(x_1, \dots, x_n) \wedge \text{Maj}(y_1, \dots, y_n)$ must have size $n^{\Omega_d(\log n)}$.*

Proof. The proof is by contradiction. Let $M = n^{o(\log n)}$ and let C be a depth- d TAC^0 of size M that computes $f(x, y)$. We analyze the effect of hitting C with a very strong random restriction ρ , one which has parameter $p = n^{-0.1}$. It is easy to see that with extremely high probability – much more than $1/2 - f_\rho$ turns into some function of the form

$$f_\rho(x, y) = \text{sign}\left(\sum_{i \in S_1} x_i + C_1\right) \wedge \text{sign}\left(\sum_{j \in S_2} y_j + C_2\right),$$

where $|S_1|, |S_2| \geq n^{0.8}$ and $|C_1|, |C_2| \leq n^{0.51}$. For any such ρ , by fixing at most $2n^{.51}$ additional variables, we get $\text{Maj}(x') \wedge \text{Maj}(y')$ where x', y' are $\Omega(n^{0.8})$ -bit

strings. By the recent result of Sherstov [She09], any PTF for this function must have degree at least $c_1 \log n$ for some absolute constant $c_1 > 0$.

On the other hand, let us consider what happens to the TAC^0 C under such a strong random restriction using Lemma 3. Since $p = n^{-0.1}$, we have $p' = n^{-0.1} \cdot 10^{d-1} (4 \log M)^{d-2} < n^{-0.09}$ for n sufficiently large. Taking $k = (c_1/2) \log n$, Lemma 3 gives us that C_ρ has a PTF of degree at most $(c_1/2) \log n$ with failure probability at most

$$M^{-2} + M^5 (4e \log(M) p' / k)^k = M^{-2} + M^5 n^{-\Omega(\log n)} < 1/2$$

since $M = n^{o(\log n)}$. Thus, there must be some restriction ρ such that f_ρ has PTF degree at least $c_1 \log n$, but C_ρ has PTF degree at most $(c_1/2) \log n$. This contradiction proves the theorem.

Aspnes et al. [ABFR94] prove lower bounds on the size of TAC^0 circuits that compute various functions such as parity. The method of [ABFR94] is useful for functions that have high weak PTF degree (such as parity). In contrast, our argument above gives us a TAC^0 lower bound for the function $\text{Maj}(x) \wedge \text{Maj}(y)$, which is known [BRS95] to have PTF degree only $O(\log n)$.

6.3 Lower bounds on TAC^0

We prove that no analogue of Beigel's theorem [Bei94] is possible for Thr gates: even an AND of two Thr gates may require a TAC^0 of more than quasi-polynomial size. The proof (see [GS10]) is similar to that of Theorem 7, it uses a recent result of Sherstov [She09] showing that the function $f(x, y) = U_{n,4n}(x) \wedge U_{n,4n}(y)$ (see Section 6.1) has PTF degree $\Omega(n)$.

Theorem 8. *Fix any absolute constant d . Any TAC^0 circuit of depth d that computes $f(x, y) = U_{n,4n}(x) \wedge U_{n,4n}(y)$ must have size $2^{\Omega(n^{1/(d-1)})}$.*

6.4 Lower bounds on $\text{TAC}^0[t(n)]$

Inspection of the proof of Theorem 4 is easily seen to imply that the parity function cannot be computed by a $\text{TAC}^0[(\log n)^{2/3}]$ circuit. In the full version [GS10] we give an improved bound that allows up to $O(\log n)$ threshold gates.

Theorem 9. *Fix any absolute constant d . Any $\text{poly}(n)$ -size, depth- d $\text{TAC}^0[t(n)]$ circuit that computes the parity function must have $t(n) = \Omega(\log n)$.*

References

- [ABFR94] J. Aspnes, R. Beigel, M. Furst, and S. Rudich. The expressive power of voting polynomials. *Combinatorica*, 14(2):1–14, 1994.
- [Bei94] R. Beigel. When do extra majority gates help? $\text{polylog}(n)$ majority gates are equivalent to one. *Computational Complexity*, 4:314–324, 1994.

- [BOW08] E. Blais, R. O'Donnell, and K. Wimmer. Polynomial regression under arbitrary product distributions. In *COLT*, pages 193–204, 2008.
- [BRS95] R. Beigel, N. Reingold, and D. Spielman. PP is closed under intersection. *Journal of Computer & System Sciences*, 50(2):191–202, 1995.
- [BT96] N. Bshouty and C. Tamon. On the Fourier spectrum of monotone functions. *Journal of the ACM*, 43(4):747–770, 1996.
- [DLM⁺07] I. Diakonikolas, H. Lee, K. Matulef, K. Onak, R. Rubinfeld, R. Servedio, and A. Wan. Testing for concise representations. In *FOCS*, pages 549–558, 2007.
- [DRST09] I. Diakonikolas, P. Raghavendra, R. Servedio, and L.-Y. Tan. Average sensitivity and noise sensitivity of polynomial threshold functions. Available at <http://arxiv.org/abs/0909.5011>, 2009.
- [FJS91] M. Furst, J. Jackson, and S. Smith. Improved learning of AC^0 functions. In *COLT*, pages 317–325, 1991.
- [FK09] Lance Fortnow and Adam Klivans. Efficient learning algorithms yield circuit lower bounds. *Journal of Computer & System Sciences*, 75(1):27–36, 2009.
- [GHR92] M. Goldmann, J. Håstad, and A. Razborov. Majority gates vs. general weighted threshold gates. *Computational Complexity*, 2:277–300, 1992.
- [GL94] C. Gotsman and N. Linial. Spectral properties of threshold functions. *Combinatorica*, 14(1):35–50, 1994.
- [Gol97] M. Goldmann. On the power of a threshold gate at the top. *Information Processing Letters*, 63(6):287–293, 1997.
- [GS10] P. Gopalan and R. A. Servedio. Learning and Lower Bounds for AC^0 with Threshold Gates. Available at <http://eccc.hpi-web.de/report/2010/074/>, 2010.
- [Han07] K. Hansen. Computing symmetric Boolean functions by circuits with few exact threshold gates. In *COCOON*, pages 448–458, 2007.
- [Hås86] J. Håstad. *Computational Limitations for Small Depth Circuits*. MIT Press, Cambridge, MA, 1986.
- [Hås01] J. Håstad. A slight sharpening of LMN. *Journal of Computer and System Sciences*, 63(3):498–508, 2001.
- [HKM09] P. Harsha, A. Klivans, and R. Meka. Bounding the sensitivity of polynomial threshold functions. Available at <http://arxiv.org/abs/0909.5175>, 2009.
- [JKS02] J. Jackson, A. Klivans, and R. Servedio. Learnability beyond AC^0 . In *STOC*, pages 776–784, 2002.
- [KKMS08] A. Kalai, A. Klivans, Y. Mansour, and R. Servedio. Agnostically learning halfspaces. *SIAM Journal on Computing*, 37(6):1777–1805, 2008.
- [KOS04] A. Klivans, R. O'Donnell, and R. Servedio. Learning intersections and thresholds of halfspaces. *Journal of Computer & System Sciences*, 68(4):808–840, 2004.
- [KOS08] A. Klivans, R. O'Donnell, and R. Servedio. Learning geometric concepts via Gaussian surface area. In *FOCS*, pages 541–550, 2008.
- [LMN93] N. Linial, Y. Mansour, and N. Nisan. Constant depth circuits, Fourier transform and learnability. *Journal of the ACM*, 40(3):607–620, 1993.
- [MOS04] E. Mossel, R. O'Donnell, and R. Servedio. Learning functions of k relevant variables. *Journal of Computer & System Sciences*, 69(3):421–434, 2004.
- [OS07] R. O'Donnell and R. Servedio. Learning monotone decision trees in polynomial time. *SIAM J. Comput.*, 37(3):827–844, 2007.
- [Pod10] V. Podolskii. Personal communication, 2010.
- [She09] A. Sherstov. The intersection of two halfspaces has high threshold degree. In *FOCS*, pages 343–362, 2009.