Electronic Colloquium on Computational Complexity, Report No. 56 (2012)

Attribute-Efficient Learning and Weight-Degree Tradeoffs for Polynomial Threshold Functions

Rocco A. Servedio[*] Li-Yang Tan[†] Justin Thaler [‡]

May 1, 2012

Abstract

We study the challenging problem of learning decision lists attribute-efficiently, giving both positive and negative results.

Our main positive result is a new tradeoff between the running time and mistake bound for learning length-k decision lists over n Boolean variables. When the allowed running time is relatively high, our new mistake bound improves significantly on the mistake bound of the best previous algorithm of Klivans and Servedio [7].

Our main negative result is a new lower bound on the *weight* of any degree-d polynomial threshold function (PTF) that computes a particular decision list over k variables (the "ODD-MAX-BIT" function). The main result of Beigel [1] is a weight lower bound of $2^{\Omega(k/d^2)}$, which was shown to be essentially optimal for $d \leq k^{1/3}$ by Klivans and Servedio. Here we prove a $2^{\Omega(\sqrt{k/d})}$ lower bound, which improves on Beigel's lower bound for $d > k^{1/3}$. This lower bound establishes strong limitations on the effectiveness of the Klivans and Servedio approach and suggests that it may be difficult to improve on our positive result. The main tool used in our lower bound is a new variant of Markov's classical inequality which may be of independent interest; it provides a bound on the derivative of a univariate polynomial in terms of both its degree *and* the size of its coefficients.

1 Introduction

Learning in the presence of irrelevant information is a central problem both in machine learning theory and in many practical applications. A clean theoretical formulation of this problem is provided by the framework of *attribute-efficient learning* that was introduced by Blum more than twenty years ago [2]. A class $\mathcal{C}$ of Boolean functions over $\{0,1\}^n$ is said to be *attribute-efficiently learnable* if there is an online algorithm that learns any k-variable function in $\mathcal{C}$ with a mistake bound that is only $\text{poly}(k, \log n)$.[1] While attribute-efficient algorithms are known for simple classes of functions

[*] Department of Computer Science, Columbia University.

[†] Department of Computer Science, Columbia University.

[‡] School of Engineering and Applied Sciences, Harvard University. Supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

[1] Throughout this paper we describe all learning results in terms of mistake bounds for online learning algorithms; standard conversions [11] can be used to translate such statements into sample complexity bounds for PAC-model learning algorithms.

ISSN 1433-8092

such as k-variable conjunctions, disjunctions, and linear threshold functions with coefficients that are integers of magnitude at most poly(k) [12], researchers have been less successful in developing attribute-efficient algorithms for richer classes of Boolean functions.

A particularly well-studied – and challenging – problem in this area is that of attribute-efficiently learning *decision lists*. This question was first posed by Blum [2] and considered subsequently by many authors [3, 4, 16, 14, 7, 13]. We revisit this problem, presenting both positive and negative results that significantly improve upon prior work. Both our upper and lower bounds rely on a careful study of the achievable tradeoffs between the degree and weight of *polynomial threshold functions* (PTFs) for decision lists. (See Section 2 for the standard definitions of decision lists, polynomial threshold functions, and the ODD-MAX-BIT function that we use throughout the paper.)

Prior Work on Attribute-Efficiently Learning Decision Lists. Given the apparent difficulty of obtaining a poly$(k, \log n)$-mistake-bound algorithm that runs in time poly(n) for learning length-k decision lists, it is natural to try to obtain some nontrivial tradeoff between running time and mistake bound for this problem. It has long been known that the Winnow algorithm [12] can learn any length-k decision list running in poly(n) time per trial but with a mistake bound of $2^{\Theta(k)} \log n$. On the other hand, it is also well known that by running Winnow over an "expanded feature space" of all monomials (i.e. monotone conjunctions) of length at most k, the algorithm can learn any length-k decision list with a mistake bound of poly$(k, \log n)$ but running in time n^k per trial. Is it possible to trade off between these two simple results?

Klivans and Servedio [7] observed that by running the Winnow algorithm over an expanded feature space of all monomials of length at most d, weight upper bounds for degree-d PTFs that compute decision lists directly yield tradeoffs between running time and mistake bound for learning decision lists.

More precisely, they showed the following:

Fact 1 *Let $\mathcal{C}$ be a class of Boolean functions with the property that each $f \in \mathcal{C}$ has a polynomial threshold function (over $\{0,1\}^n$ or $\{-1,1\}^n$) of degree at most d and weight at most W. Then the Expanded-Winnow algorithm runs in $n^{O(d)}$ time per trial and has mistake bound $O(W^2 \cdot d \cdot \log n)$ for $\mathcal{C}$.*

As their main result, Klivans and Servedio showed that for every $1 \leq d \leq \sqrt{k}$, any length-k decision list has a PTF of degree d and weight $2^{\tilde{O}(k/d^2+d)}$. Together with Fact 1, for $d \leq k^{1/3}$ this yields a nontrivial tradeoff between running time ($n^{O(d)}$) and mistake bound ($2^{\tilde{O}(k/d^2)} \log n$) for learning length-$k$ decision lists over $\{0,1\}^n$. The tradeoff breaks down for larger settings of d, though, because for $d > k^{1/3}$ the weight bound of Klivans and Servedio is worse than the bound obtained at $d = k^{1/3}$.

In earlier work Beigel [1] had already proved that any degree-d PTF for the ODD-MAX-BIT decision list over k variables must have weight $2^{\Omega(k/d^2)}$. The Klivans and Servedio upper bound shows that this lower bound is essentially optimal for $d \leq k^{1/3}$, but again it was not clear what is the right bound for larger values of d.

Several natural questions present themselves given this state of prior work: for $d > k^{1/3}$, is there an n^d-time algorithm that can learn length-k decision lists with a better mistake bound than $2^{\tilde{O}(k^{1/3})} \log n$? Is there a stronger lower bound than $2^{\Omega(k/d^2)}$ for the weight of degree-d PTFs that compute length-k decision lists? As the main results of this paper we give affirmative answers to both these questions.

Our Positive Result for Learning Decision Lists. For any $d \geq k^{1/3}$, we give an algorithm that learns an unknown length-k decision list in time $n^{O(d)}$ per trial with a mistake bound of $\tilde{O}(kd^2) \cdot 2^{\tilde{O}(\sqrt{k/d})} \log n$. For all $d \geq k^{1/3}$ this improves on the $2^{\tilde{O}(k/d^2+d)} \log n$ mistake bound of Klivans and Servedio [7], and it also improves on the $O(d \log(n)) \cdot 2^{O(k/d)}$ mistake bound, which holds for all $d < k$, achieved by an intermediate result of that paper [7, Page 593].

Our approach uses a simple extension of Fact 1 to "generalized PTFs" (GPTFs), which are linear combinations of *arbitrary* conjunctions of length at most d (note that in contrast a "standard" degree-d PTF over $\{0,1\}^n$ is a linear combination of *monotone* conjunctions of length at most d). In order to use this extension of Fact 1, we prove a new weight upper bound on degree-d GPTFs that compute decision lists; see Section 3.

Our Lower Bounds for Decision Lists. In Section 4, we show that for $d > k^{1/3}$, any degree-d PTF for the length-k decision list $\mathsf{ODD\text{-}MAX\text{-}BIT}_k$ must have weight $2^{\Omega(\sqrt{k/d})}$. (This weight lower bound holds over both domains $\{0,1\}^k$ and $\{-1,1\}^k$.) This strictly improves Beigel's lower bound for $k^{1/3} < d \leq k$, and suggests that our positive result may be essentially optimal for d in this range.

Our lower bound proof augments the approach of Beigel with a new technical ingredient which may be of independent interest, namely a variant of the classical Markov's inequality for real polynomials. While the original inequality gives a uniform upper bound on the derivative of a real univariate polynomial in terms of its degree, our new variant also takes into account the *size of the coefficients* of the polynomial to obtain a sharper bound when the coefficients are small; see Lemma 1. As described in Sections 3 and 4, a common intuition underlies both the key lower bound ingredient, Lemma 1, and the GPTF construction that gives our positive result.

We also improve Beigel's $2^{\Omega(k/d^2)}$ lower bound in another direction, by extending it to a wider space of allowable features. Beigel's result may be interpreted as saying that any sign-representation of $\mathsf{ODD\text{-}MAX\text{-}BIT}_k$ as an integer combination of *length-d monotone conjunctions* must have total weight $2^{\Omega(k/d^2)}$. We show that if $\mathcal{F}$ is any family of 2^d functions over k-bit inputs that is closed under restrictions, then there is some decision list over at most k variables such that any sign-representation as an integer combination of functions from $\mathcal{F}$ must have total weight $2^{\Omega(k/d^2)}$. See Section 5.

2 Preliminaries: Decision lists, PTF weight and degree, and Markov's inequality

A *decision list* $L : \{0,1\}^n \to \{-1,1\}$ of length k over the Boolean variables $x_1, \ldots, x_n$ computes a Boolean function in the following way: given an input assignment, the output bit is determined according to the rule

$$\text{“}\textbf{if } \ell_1 \textbf{ then output } b_1 \textbf{ else } \ldots \textbf{ else if } \ell_k \textbf{ then output } b_k \textbf{ else output } b_{k+1},\text{”}$$

where each ℓ_i is a literal and each b_i is either -1 or 1.

A decision list of particular interest to us is the function $\mathsf{ODD\text{-}MAX\text{-}BIT}_n$: $\{0,1\}^n \to \{-1,1\}$. This function outputs 1 on an n-bit input string $z \in \{0,1\}^n$ if and only if the rightmost bit of z that is set to 1 is in an odd bit position, i.e. z is of the form $x10^j$ where the length of x is even,

In the above definition, decision lists are defined over the domain $\{0,1\}^n$, but we may just as well view these functions as being defined over domain $\{-1,1\}^n$.

PTFs. Let $p(x_1, \ldots, x_n)$ be an n-variable polynomial of total degree d with integer coefficients whose absolute values sum to W. Let f be any function from $\{0,1\}^n$ to $\{-1,1\}$. If $f(x) = \text{sign}(p(x))$ for all $x \in \{0,1\}^n$, then we say that p is a *polynomial threshold function for f* (PTF) over $\{0,1\}^n$ of degree d and weight W. Similarly, for $f : \{-1,1\}^n \to \{-1,1\}$ we say that p is a *PTF for f over $\{-1,1\}^n$ of degree d and weight W* if $f(x) = \text{sign}(p(x))$ for all $x \in \{-1,1\}^n$. Finally, we say that a PTF p has *length s* if the polynomial p has at most s non-zero coefficients.

It is easy to see that a PTF of degree d and weight W over $\{0,1\}^n$ corresponds to a depth-2 circuit with a majority gate as the output gate and W AND gates each of fan-in at most d at the bottom, where we allow negations between each AND gate and the majority gate. A PTF over $\{-1,1\}^n$ corresponds to a similar circuit but with W parity gates of fan-in at most d, rather than AND gates of fan-in at most d, at the bottom layer.

Markov's Inequality. We will also need the following well-known inequality due to Markov.

Fact 2 (Markov) *Let $P(t)$ be a univariate real polynomial of degree k. Then for all $t \in [-1,1]$ we have*

$$|P'(t)| \leq k^2 \cdot \max_{t \in [-1,1]} |P(t)|.$$

3 An improved tradeoff between runtime and mistake bound for large d.

In this section we present our main positive result, giving an improved mistake bound for learning length-k decision lists in time n^d when $k^{1/3} \leq d \leq k$:

Corollary 1 *For any $k^{1/3} \leq d \leq k$, there is an algorithm that learns length-k decision lists over n Boolean variables in time $n^{O(d)}$ with a mistake bound of $\tilde{O}(kd^2) \cdot 2^{\tilde{O}(\sqrt{k/d})} \cdot \log n$.*

We first introduce our new approach of learning via GPTFs and state a GPTF analogue of Fact 1. Then, after briefly reviewing the Klivans and Servedio construction, we modify their construction to fit the new GPTF framework and thereby prove the main positive result. We close this subsection by noting that our GPTF approach can be combined with ingredients from Krause and Pudlák [9] to obtain a weight upper bound for honest-to-goodness PTFs over the domain $\{-1,1\}^n$ (which is somewhat weaker, though, than the weight upper bound that we establish for GPTFs.)

Learning using GPTFs. Let $\mathcal{C}_d$ denote the class of all conjunctions (not necessarily monotone) of at most d literals over Boolean variables $x_1, \ldots, x_n$, where we view conjunctions as outputting either 0 (false) or 1 (true). A *degree-d generalized PTF (GPTF)* for $f : \{0,1\}^n \to \{-1,1\}$ is an expression of the form

$$p(x_1, \ldots, x_n) = \sum_{c \in \mathcal{C}_d} w_c \cdot c(x_1, \ldots, x_n),$$

where each coefficient w_c is an integer, which is such that $f(x) = \text{sign}(p(x))$ for all $x \in \{0,1\}^n$. The *weight* of p is $\sum_{c \in \mathcal{C}_d} |w_c|$.

What is the point of using GPTFs instead of PTFs over $\{0,1\}^n$? It is clear that if f has a GPTF of degree d and weight W, then f also has a PTF of degree d and weight $W2^d$ over $\{0,1\}^n$:

simply replace each conjunction $c(x)$ in the GPTF with its interpolating 0/1-valued polynomial (which has integer coefficients whose magnitudes sum to at most 2^d). Thus for weight bounds $W \geq 2^d$, GPTFs cannot have much smaller weight than PTFs over $\{0,1\}^n$; but if $2^d \gg W$, then it is possible that GPTFs may offer a significant savings in weight over ordinary PTFs. We exploit precisely this phenomenon to obtain our main positive result. For $d \geq n^{1/3}$ we do not know how to prove a stronger weight bound than $2^{\tilde{O}(n^{1/3})}$ for degree-d PTFs over $\{0,1\}^n$ that compute length-n decision lists, but for GPTFs we will show the following (which, together with Fact 3, immediately gives Corollary 1):

Theorem 1 *Let f be any length-n decision list. For all $n^{1/3} \leq d \leq n$, there is a degree-d GPTF for f that has weight $\tilde{O}(\sqrt{nd}) \cdot 2^{\tilde{O}\left(\sqrt{n/d}\right)}$.*

The following fact is a straightforward analogue of Fact 1:

Fact 3 *Let $\mathcal{F}$ be a class of Boolean functions over $\{0,1\}^n$ with the property that each $f \in \mathcal{F}$ has a degree-d GPTF of weight at most W. Then there is an online learning algorithm for $\mathcal{F}$ which runs in $n^{O(d)}$ time per trial and has mistake bound $O(W^2 \cdot d \cdot \log n)$.*

The algorithm simply runs Winnow over the feature space of all (not just monotone) conjunctions of length at most d over $x_1, \ldots, x_n$. Fact 3 follows directly from [7, Theorem 1] since there are $\sum_{i=0}^{d} \binom{n}{i} 2^i \leq (2n)^d$ such conjunctions.

Proof outline for Theorem 1. We first recall the high-level structure of the Klivans and Servedio [7] construction of low-weight PTFs for decision lists. Given a decision list L over n variables and a degree bound d, Klivans and Servedio first break the decision list into $n/\tilde{O}(d^2)$ "inner lists" of length $\tilde{O}(d^2)$, and view L as a consisting of an "outer" decision list over the $n/\tilde{O}(d^2)$ inner lists. (This is referred to as the "Outer Construction.") Next, using Chebyshev polynomials, they construct a polynomial of degree d that gives a very accurate uniform approximation for each $\tilde{O}(d^2)$-length inner list; they call this the "Inner Construction." The final PTF is obtained by composing the inner and outer constructions; it has weight $2^{\tilde{O}(n/d^2+d)}$.

As mentioned in the Introduction, the lower bound of Beigel [1] shows that the Klivans and Servedio weight upper bound is essentially optimal for $d \leq n^{1/3}$. However, for $d \geq n^{1/3}$ the weight bound of Klivans and Servedio is essentially 2^d; roughly speaking this is because the degree-d Chebyshev polynomial used in the inner construction has weight essentially 2^d, which is quite high. In our construction we avoid this 2^d weight factor by exploiting the fact that, with GPTFs at our disposal, we can instead use a *lower*-degree Chebychev polynomial within the inner construction. Intuitively, this is because with GPTFs we can "use non-monotone ANDs as variables in our polynomial" without paying for it in the weight bound, as each such AND has GPTF weight 1. (This is reminiscent of the intuition that underlies the main technical lemma for our lower bounds, see Section 4.)

Outer construction. Fix any decision list f, and recall that f can be written as "if ℓ_1 then output b_1 else $\cdots$ else if ℓ_n then output b_n else output b_{n+1}," where each ℓ_i is a Boolean literal and each b_i is an output bit in $\{-1,1\}$.

Claim 1 *Let f be a length-n decision list. For every $r \leq n$ there is a degree-r GPTF for f that has weight $r \cdot 3^{O(n/r)}$.*

Proof: We assume below that r divides n evenly and prove an $O(r \cdot 3^{n/r})$ weight bound; this implies the claim as stated above. We break f into length-r "modified decision lists" $f_1, \ldots, f_{n/r}$ which output values in $\{-1, 0, 1\}$ as described below. The first function f_1 may be expressed as a ± 1 sum of r conjunctions as follows:

$$f_1(x) = b_1\ell_1 + b_2(\bar{\ell}_1 \wedge \ell_2) + \cdots + b_r(\bar{\ell}_1 \wedge \cdots \wedge \bar{\ell}_{r-1} \wedge \ell_r)$$

(recall that each $b_i \in \{-1, 1\}$ and each conjunction gives a value in $\{0, 1\}$). It is clear that f_1 outputs the correct value b_i of the decision list f if any literal $\ell_i \in \{\ell_1, \ldots, \ell_r\}$ is satisfied by input x, and otherwise f_1 outputs the value 0 (since no conjunction is satisfied). The other functions $f_2, \ldots, f_{n/r}$ are defined similarly. Notice that each modified decision list f_i has GPTF weight r and GPTF degree r. Combining these expressions for the modified decision lists as

$$P(x) = 3^{n/r}f_1 + 3^{n/r-1}f_2 + \cdots + 3^1 f_{n/r} + b_{n+1}, \tag{1}$$

we get that $P(x)$ is a GPTF for f that has weight $O(r \cdot 3^{n/r})$ and degree r. ■

Inner Construction. In the outer construction we broke f into $\ell := n/r$ modified decision lists $f_1, \ldots, f_\ell$ each of length r, as in Eq. (1). We set parameters $\ell = \sqrt{n/d}$ and $r = \sqrt{nd}$.

Claim 2 *Let $n^{1/3} \leq d \leq n$ and let f_i be a modified decision list of length $r = \sqrt{nd}$. There is an integer linear combination p_i of conjunctions of width $2d\log(n/d)$, such that*

- *The weight of p_i is $2^{\tilde{O}\left(\sqrt{n/d}\right)}$;*
- *There is an integer $C = 2^{\tilde{O}\left(\sqrt{n/d}\right)}$ such that for every input $x \in \{0,1\}^r$ that satisfies at least one literal in f_i, we have $|p_i(x) - Cf_i(x)| \leq C/r$; and*
- *$f_i(x) = 0$ implies $p_i(x) = 0$.*

We note that in [7, Corollary 9] the Chebyshev polynomial is used to construct a polynomial of degree $2\sqrt{r}\log r$ and weight $2^{O(\sqrt{r}\log^2 r)}$ that satisfies the last two properties of the lemma. Here, exploiting the fact that our p_i may be a linear combination of conjunctions, we show how to achieve the same properties using higher weight but lower degree; this will be useful for us later when we combine with the outer construction.

Proof of Claim 2: Recall that f_i may be expressed as $f_i(x) = b_1T_1(x) + \cdots + b_rT_r(x)$, where each T_j is a conjunction of length $j \leq r$. We view T_j as an AND-of-ANDs, where

- The top AND has fanin $t := \lceil jr/d^2 \rceil \leq \lceil r^2/d^2 \rceil = \lceil n/d \rceil$, and
- Each bottom AND has fanin at most $b := \lceil j/t \rceil \leq \lceil d^2/r \rceil = \lceil \sqrt{d^3/n} \rceil$. (Note that $d^3/n \geq 1$ since $d \geq n^{1/3}$ by assumption.)

Let $\mathsf{AND}_t(y_1, \ldots, y_t)$ denote the top AND gate, so the inputs $y_1, \ldots, y_t$ are the bottom-level ANDs. To construct p we first apply the construction of [7, Corollary 9] to AND_t (which is itself a decision list on t variables). This yields a polynomial q (in 0/1 variables $y_1, \ldots, y_t$) of degree at most $2\sqrt{n/d}\log(n/d)$ and weight $2^{O(\sqrt{n/d}\log^2(n/d))}$, and an integer $C = 2^{O(\sqrt{n/d}\log^2(n/d))}$ such that $|q(y_1, \ldots, y_t) - C \cdot \mathsf{AND}_t(y_1, \ldots, y_t)| \leq \frac{C}{t}$ for all $y \in \{0,1\}^t \setminus \{0^t\}$. Moreover, if $y = 0^t$ then $q(y) = 0$.

The inputs to the top AND gate are themselves conjunctions of width $b \leq \sqrt{d^3/n}$, so we obtain our final representation p_i by simply composing q with the exact representation of each bottom AND gate as a conjunction of width b. (For example, if the polynomial q contains a monomial $y_1 y_3 y_4$ we simply AND together the first, third and fourth bottom-level AND gates.) Since q has degree at most $2\sqrt{n/d}\log(n/d)$, p_i is a linear combination of conjunctions of width at most $b \cdot 2\sqrt{n/d}\log(n/d) = 2d\log(n/d)$ as required. The weight of the linear combination equals the weight of q, which is $2^{O(\sqrt{n/d}\log^2(n/d))}$. ■

Proof of Theorem 1: The remainder of the analysis necessary to prove Theorem 1 follows exactly as in [7, Theorem 10]; we present the details for completeness. We assume that f is the decision list $(x_1, b_1), \ldots, (x_n, b_n), b_{n+1}$ (the case when f contains negated literals is entirely similar). We begin with the outer construction and note that

$$f(x) = \operatorname{sign}\left(C \cdot \left(\sum_{i=1}^{\ell} 3^{\ell-i+1} f_i(x) + b_{n+1}\right)\right),$$

where C is the value from Claim 2 and each f_i is a modified decision list of length $r := \sqrt{nd}$ computing the restriction of f to its i-th block (so the total number of modified decision lists is $\ell := n/r = \sqrt{n/d}$). Next we replace each Cf_i by p_i, the approximating polynomial given by the inner construction in Claim 2, and consider

$$H(x) = \sum_{i=1}^{\ell}(3^{\ell-i+1} p_i(x)) + Cb_{n+1}.$$

We will argue that $\operatorname{sign}(H(x))$ is a GPTF for f with the claimed degree and weight. Fix any $x \in \{0,1\}^n$. If $x = 0^n$ then by Claim 2 each $p_i(x) = 0$ and so $H(x) = Cb_{n+1}$ agrees in sign with b_{n+1}. Now suppose that $t = (i-1)r + c$ is the first index such that $x_t = 1$ (i.e. the input x "exits the overall decision list" at the modified decision list f_t, and $f(x) = f_t(x) = b_t$). By Claim 2, we have

- $3^{\ell-j+1} p_j(x) = 0$ for all $j < i$,
- $3^{\ell-i+1} p_i(x)$ differs from $3^{\ell-i+1} C b_t$ by at most $C 3^{\ell-i+1} \cdot \frac{1}{r}$, and
- the magnitude of $3^{\ell-j+1} p_j(x)$ is at most $C 3^{\ell-j+1}(1 + \frac{1}{r})$ for all $j > i$.

Combining these bounds, the value of $H(x)$ differs from $3^{\ell-i+1} C b_t$ by at most

$$C\left(\frac{3^{\ell-i+1}}{r} + \left(1 + \frac{1}{r}\right)(3^{\ell-i} + 3^{\ell-i-1} + \ldots + 3) + 1\right),$$

which can be checked to be less than $C 3^{\ell-i+1}$ in magnitude for $r > 1$. Consequently, $\operatorname{sign}(H(x)) = b_t = f(x)$ as claimed. Finally, applying the degree and weight bounds for each p_i given by Claim 2 completes the proof.

■

3.1 A New Upper Bound for PTFs over $\{-1,1\}^n$

We do not know whether n-variable decision lists have degree-d PTFs (as opposed to GPTFs) of weight $\tilde{O}(\sqrt{nd}) \cdot 2^{\tilde{O}\left(\sqrt{n/d}\right)}$, though in the next section we will prove that they do *not* have PTFs of weight $2^{o\left(\sqrt{n/d}\right)}$. For PTFs over $\{-1,1\}^n$, though, we can combine the approach of Theorem 1 with ideas from Krause and Pudlák [9] to give a new upper bound on weight which improves on the Klivans and Servedio weight upper bound for $d \geq n^{1/2}$. In Appendix A we prove:

Theorem 2 *For any $n^{1/4} \leq d \leq n$, any decision list f on n variables has a degree-d PTF over the domain $\{-1,1\}^n$ of weight $2^{(n/d)^{2/3} \cdot \mathrm{polylog}(n)}$.*

4 Degree-d PTFs for $\mathsf{ODD\text{-}MAX\text{-}BIT}_n$ Require Weight $2^{\Omega(\sqrt{n/d})}$

In this section we prove our main lower bound:

Theorem 3 *For any $d = o\left(\frac{n}{\log^2 n}\right)$, any degree-$d$ PTF for $\mathsf{ODD\text{-}MAX\text{-}BIT}_n$ requires weight $2^{\Omega(\sqrt{n/d})}$. This holds for PTFs over both domains $\{0,1\}^n$ and $\{-1,1\}^n$.*

Proof Outline. To explain our proof it is helpful to first recall the $2^{\Omega(n/d^2)}$ lower bound proof of Beigel [1]. Beigel first breaks the list of n variables into $k = n/\ell$ blocks of length $\ell := \Theta(d^2)$ each. His argument proceeds for k stages; in the i-th stage he argues that there exists an n-bit input x_i with $|p(x_i)| \geq 2|p(x_{i-1})|$ which is such that x_i and x_{i-1} differ only in the i-th block. After the k-th stage this yields an input x_k that has $|p(x_k)| \geq 2^k$; such an input clearly implies that the weight of p is at least 2^k.

The existence of the desired x_i is established in each stage using Markov's inequality for univariate real polynomials (see Fact 2 of Section 2). Beigel shows that if no such x_i existed, then by using x_{i-1} and symmetrizing p with respect to an appropriate distribution, one would obtain a univariate polynomial P of degree d such that (roughly speaking) P "stays bounded" on the interval $[0,1]$, and P has large derivative at some point in this interval (i.e. $|P'(x)| = \Omega(\ell)$ for some $x \in [0, 1/\ell]$). But Markov's inequality implies that any such polynomial must have $\Omega(\sqrt{\ell}) > d$; consequently, the desired x_i must indeed exist.

While Beigel's lower bound is essentially optimal for $d \leq n^{1/3}$, it becomes vacuous for $d \geq n^{1/2}$ (and we show below that it is not tight even for $d = \omega(n^{1/3})$.) Intuitively, the reason that the bound is loose in the high-degree regime is that Markov's inequality only takes the *degree* of the polynomial P into account and does not use information about the *weight* of P. With this intuition in mind, we prove a generalization of Markov's inequality that takes into account P's weight. More specifically, we show that if P has degree d and weight W, "stays bounded" in the interval $[0,1]$, and satisfies $\max_{x\in[0,1]} |P(x)| \geq 1/2$, then $|P'(x)| = O(d \log W)$ for all $x \in [0,1]$. Notice that for $W = 2^d$ this bound matches Markov's inequality, but – crucially – it is significantly stronger if W is subexponential in d.

With this Markov-type inequality in hand, our proof follows the same outline as Beigel. We break the list into blocks of length ℓ; with our stronger inequality we are able to take ℓ to be $o(d^2)$, which is why we improve on Beigel's bound. We iteratively construct inputs x_i with $|p(x_i)| \geq 2|p(x_{i-1})|$, arguing at each stage i that if no such x_i existed, then one could obtain a univariate polynomial P of weight $2^{O(\ell/d)}$ which satisfies the conditions of our Markov-type inequality. By invoking our inequality, we conclude that P must have degree greater than d, and so this implies that the desired x_i must exist at each stage.

The Markov-Type Inequality. Below we state our Markov-type inequality as Lemma 1. This lemma generalizes a result of Borwein and Erdélyi [5, Corollary 3.2] and may be of independent interest.

Lemma 1 *Let $P : \mathbb{R} \to \mathbb{R}$ be a degree-d polynomial satisfying the following:*

1. *The coefficients of P (which need not be integers) each have absolute value at most W; and*
2. $1/2 \le \max_{x \in [0,1]} |P(x)| \le R$.

Then $\max_{x\in[0,1]} |P'(x)| = O(d \cdot R \cdot \max\{\log W, \log d\})$.

At a high level, the proof first shows that $|P(y)|$ is not just "small" for all $y \in [0,1]$, but in fact is small everywhere within a sufficiently large ellipse B_ρ surrounding $[0,1]$ in the complex plane. Cauchy's Integral Formula then completes the result, as this formula expresses $P'(y)$ as an average of P's values on B_ρ, scaled by (the squared reciprocal of) the diameter of B_ρ.

Proof: Borwein and Erdélyi proved the following lemma, which is a consequence of Hadamard's Three Circle Theorem, a classical result in complex analysis.

Lemma 2 *[5, Corollary 3.2] Let $M \in \mathbb{R}$ and $d \in \mathbb{N}$ satisfy $1 \le M \le 2d$. Suppose f is analytical inside and on the complex ellipse $A_{d,M}$ with foci at 0 and 1 and major axis $\left[-\frac{M}{d}, 1+\frac{M}{d}\right]$. Let $B_{d,M}$ be the complex ellipse with foci at 0 and 1 and with major axis $\left[-\frac{1}{dM}, 1+\frac{1}{dM}\right]$. Then there is an absolute constant $c_1 > 0$ such that*

$$\max_{z\in B_{d,M}} \ln|f(z)| \le \max_{z\in[0,1]} \ln|f(z)| + \frac{c_1}{M}\left(\max_{z\in A_{d,M}} \ln|f(z)| - \max_{z\in[0,1]} \ln|f(z)|\right).$$

Note that for all $z \in A_{d,M}$, we have

$$\begin{aligned} \ln|P(z)| &\le \ln\left((d+1)W\left(1+\frac{M}{d}\right)^d\right) \\ &\le \ln(d+1) + \ln(W) + d\cdot\frac{M}{d} = \ln(d+1) + \ln(W) + M, \end{aligned}$$

where the first inequality holds because P has at most $d+1$ coefficients, each of which has absolute value at most W, and $|z| \le 1 + \frac{M}{d}$ for all $z \in A_{d,M}$ and each monomial in P has degree at most d. If we take $M \ge \ln(W) + \ln(d+1)$, the upper bound on $\ln|P(z)|$ is at most $2M$.

Now applying Lemma 2 to the function P and $M = \ln(W) + \ln(d+1)$, we get that there is an absolute constant c_1 such that

$$\max_{z\in B_{d,M}} \ln|P(z)| \le \ln R + \frac{c_1}{M}(3M + \ln 2) \le \ln R + c_2$$

for some universal constant c_2, where we have used the assumption that $1/2 \le \max_{z\in[0,1]} |P(z)| \le R$. As a result, we have that

$$\max_{z\in B_{d,M}} |P(z)| \le c_3 R$$

for some universal constant $c_3 > 0$.

Now for any $y \in [0,1]$, there is a positive number $\rho = \Theta(\frac{1}{dM})$ such that $B_\rho(y) = \{z \in \mathbb{C} : |z - y| = \rho\}$, the closed disk of radius ρ around y, is contained in $B_{d,M}$. Consequently Cauchy's integral formula implies that for all $y \in [0,1]$, we have

$$\begin{aligned} |P'(y)| &= \left| \frac{1}{2\pi i} \int_{B_{d,M}} \frac{P(z)}{(z-y)^2} dz \right| = \left| \frac{1}{2\pi i} \int_{B_\rho} \frac{P(z)}{(z-y)^2} dz \right| \\ &\leq \frac{1}{2\pi} c_3 R \left| \int_{B_\rho} \frac{1}{(z-y)^2 dz} \right| \leq c_3 R \frac{1}{\rho} = \Theta(R \cdot d \cdot M). \end{aligned}$$

Here, the first equality holds by Cauchy's Theorem, because $\frac{P(z)}{(z-y)^2}$ is analytical except at y, and the penultimate inequality combines parametric evaluation of the complex integral $\int_{B_\rho} \frac{1}{(z-y)^2 dz}$ with the fact that $|z - y| = \rho$ for all $z \in B_\rho$. Since we took $M = \log(W) + \log(d + 1)$, the lemma follows. ■

Discussion. For $R = 1$, Lemma 1 is tight up to constant factors for a wide range of values of W. To see this, let $W \in [d, 3^d]$ be a power of 3, and let $d' = \lceil \frac{d}{\log_3 W} \rceil$. Consider the function $P(x) = T_{\log_3 W}(x^{d'})$, where T_k denotes the degree-k Chebyshev polynomial of the first kind. Basic properties of T_k give that all of P's coefficients have absolute value at most $3^{\log_3 W} = W$; that P has degree $d' \cdot \log_3 W = O(d)$; and that $\max_{x \in [0,1]} |P(x)| = 1$. Since $T_k'(1) = k^2$ for any k, the chain rule implies that $P'(1) = (\log_3 W)^2 \cdot d' = \Omega(d \log W)$.

This tight example formalizes the following intuition: Recalling that the Chebyshev polynomials are a tight example for Markov's inequality, Markov's inequality can be understood to say that in order to maximize the derivative of a polynomial P of degree d which stays bounded on the interval $[0,1]$, the best approach is to "spend all the allowed degree on a Chebyshev polynomial." Lemma 1 says that in order to maximize the derivative of a polynomial of degree d *and weight* W which stays bounded on the interval $[0,1]$, the best approach is to "spend all of the allowed *weight* on a Chebyshev polynomial T, and then spend any remaining allowed degree by composing T with a suitable monomial $x^{d'}$." This is the intuition that underlies our construction in the positive result Claim 2, where we use conjuntions (the bottom-level AND gates) as inputs for the top-level polynomial q (which is essentially a Chebychev polynomial).

Proof of Theorem 3. Here we prove the result for PTFs over the domain $\{0,1\}^n$; the proof over the domain $\{-1,1\}^n$ is similar and is given in Appendix B.

Let p be a degree d, weight W PTF for $\mathsf{ODD\text{-}MAX\text{-}BIT}_n$ over $\{0,1\}^n$. Suppose further that $W < 2^{\sqrt{n/d}-1}$. We partition the n variables $x_1, \ldots, x_n$ into n/ℓ blocks of length $\ell := 2\sqrt{nd}$. We refer to a string $y_i \in \{0,1\}^{\ell \cdot i}$ as a *partial input* to p. Let $V(y_i)$ denote $p(y_i \circ 0^{n-\ell \cdot i})$, where $\circ$ denotes concatenation.

We say that partial input $y_i \in \{0,1\}^{\ell \cdot i}$ is *good* if $|V(y_i)| \geq 2^{i-1}$. We will attempt to iteratively construct good partial inputs $y_1, \ldots, y_{n/\ell}$; if we succeed at all iterations, then $y = y_{n/\ell}$ is an input such that $|p(y)| \geq 2^{\sqrt{n/d}-1}$, which clearly contradicts the assumption that $W < 2^{\sqrt{n/d}-1}$. Thus, there must be some $i^* < n/\ell$ such that there is a good partial input $y_{i^*} \in \{0,1\}^{i^* \cdot \ell}$, but no good partial input $y_{i^*+1} \in \{0,1\}^{(i^*+1)\cdot \ell}$. Moreover we have $i^* \geq 1$, because it is easy to construct y_1: since p has integer coefficients we may assume without loss of generality that $|p(x)| \geq 1$ for all

$x \in \{-1,1\}^n$ (if $p(x) = 0$ for some x we can simply add $1/2$ to p and multiply the resulting polynomial by two). So we may take y_1 to be any string in $\{0,1\}^\ell$.

Below we show how to use the y_{i^*} whose existence was established above to construct a certain univariate polynomial P. By applying Lemma 1 to P, we will conclude that $W = 2^{\Omega(\sqrt{n/d})}$ as desired.

We may assume without loss of generality that $V(y_{i^*}) < 0$. Given a string $w \in \{0,1\}^{\ell/2}$, let $z_w \in \{0,1\}^n$ be the string defined as follows:

- the first i^* blocks of z_w agree with y_{i^*};
- the even bits in the (i^*+1)-th block are all 0;
- the odd bits in the (i^*+1)-th block are successively $w_1, w_2, \ldots, w_{\ell/2}$; and
- the remaining blocks $i^*+2, \ldots, n/\ell$ have all bits set to 0.

Define q to be the function which, on input $w \in \{0,1\}^{\ell/2}$, outputs $p(z_w)$, i.e. $q(w) := p(z_w)$. Note that $\deg(q) \le \deg(p)$.

Define the univariate polynomial

$$Q(t) := \mathbb{E}_{x \leftarrow \mu_t}[q(x)].$$

Here μ_t is the product distribution over the discrete cube $\{0,1\}^{\ell/2}$ where each coordinate j satisfies $\mathbb{E}_{x \leftarrow \mu_t}[x_j] = t$ (that is, each x_j is independently set to 1 with probability t). The following properties of Q are easily verified:

1. $Q(0) = q(0^{\ell/2})$ and $Q(1) = q(1^{\ell/2})$.
2. $|Q(t)| \le 2^{i^*}$ for all $t \in [0,1]$. This holds by the assumption that there is no good partial input y_{i^*+1}, i.e. $|V(y_{i^*+1})| \le 2^{i^*}$ for all $y_{i^*+1} \in \{0,1\}^{(i^*+1)\cdot\ell}$.
3. $\deg(Q) \le \deg(q) \le \deg(p)$. Indeed, given the multivariate polynomial expansion of $q(w_1, \ldots, w_{\ell/2})$, we can obtain Q easily just by "erasing all the subscripts in each variable". For example, if $q(w) = 5w_1w_2w_3 - 8w_1w_2 - 4w_1w_3w_4 - 3w_3$ then $Q(t) = 5t^3 - 8t^2 - 4t^3 - 3t = t^3 - 8t^2 - 3t$. This follows from linearity of expectation along with the fact that μ_t is defined to be the product distribution satisfying $E_{x \leftarrow \mu_t}[x_i] = t$ for all coordinates i. So indeed we have $\deg(Q) \le \deg(q) \le \deg(p)$.
4. The sum of the absolute value of the coefficients of Q is at most W. (Since this holds for p it holds for q, and since it holds for q it holds for Q by the previous remark.)

Note that we have $q(0^{\ell/2}) = V(y_{i^*}) < -2^{i^*-1}$. Crucially, we also have $q(w) \ge 1/2$ for all $w \in \{0,1\}^{\ell/2} \setminus \{0^{\ell/2}\}$ (this is because for any such w, every input $z \in Z_w$ has $p(z) \ge 1$, because p is a PTF for $\mathsf{ODD\text{-}MAX\text{-}BIT}_n$).

The final polynomial we seek is the univariate polynomial

$$P(t) := Q(t) - Q(0).$$

We trivially have $P(0) = 0$. On the other hand, for any $t \in [2/\ell, 1]$ we have $P(t) = \mathbb{E}_{x \leftarrow \mu_t}[q(x) - q(0^{\ell/2})]$. Recalling the bounds on $q(0^{\ell/2})$ and on $q(w)$ from the previous paragraph, and observing that for $t \in [2/\ell, 1]$ we have $\Pr_{x \leftarrow \mu_t}[x = 0^{\ell/2}] \leq (1 - \frac{2}{\ell})^{\ell/2} < 1/e$, we get that $P(t) \geq 2^{i^*-1} \cdot (1 - \frac{1}{e}) \geq 2^{i^*-2}$ for $t \in [2/\ell, 1]$.

Thus, $P(t)$ is a degree-d polynomial that has the following properties:

1. The coefficients of P each have absolute value at most $W + |V(y_{i^*})| \leq 2W$;
2. $1/2 \leq \max_{x \in [0,1]} |P(x)| \leq 2^{i^*+1}$;
3. There exists some $t \in [0, 2/\ell]$ such that $P'(t) \geq \ell \cdot 2^{i^*-3}$. This follows from the Mean Value Theorem, as $P(0) = 0$ and $P(2/\ell) \geq 2^{i^*-2}$.

By Lemma 1, the first two properties of P imply that $\max_{x \in [0,1]} |P'(x)| = O(d \cdot 2^{i^*} \cdot \max\{\log W, \log d\})$. Combining this with the third property of P, we get that $d \cdot \max\{\log W, \log d\} = \Omega(\ell) = \Omega(\sqrt{nd})$. Recalling that $d = o\left(\frac{n}{\log^2 n}\right)$, we conclude that $W = 2^{\Omega\left(\sqrt{n/d}\right)}$ as desired. This concludes the proof of Theorem 3. ■

5 Extending Beigel's Lower Bound to More General Features

Our next result extends Beigel's lower bound to significantly more general feature spaces. We say that a class $\mathcal{F}$ of functions over $\{0,1\}^n$ is *closed under restrictions* if for all $i \in [n]$ and all $b \in \{0,1\}$, we have that if $f(x_1, ..., x_n) \in \mathcal{F}$ then $f(x_1, \ldots, x_{i-1}, b, x_{i+1}, \ldots, x_n) \in \mathcal{F}$. In Appendix C we prove the following:

Theorem 4 *Fix $d > 0$, and let $\mathcal{F}$ be any set of at most 2^d boolean-valued functions from $\{0,1\}^n$ to $\{0,1\}$. Assume $\mathcal{F}$ is closed under restrictions. There exists a decision list L on n variables such that, for any integer-weight representation $L(x) = \mathrm{sign}(\sum_{f_i \in \mathcal{F}} w_i f_i + \theta)$ of L as a linear threshold function over the feature space $\mathcal{F}$, the absolute value of the w_i's and θ sum to $2^{\Omega\left(n/d^2\right)}$.*

Several remarks are in order. First, Theorem 4 should be compared to a closely related result of Klivans and Sherstov [8, Theorem 4.6], building on work of Buhrman *et al.* [6]. Klivans and Sherstov show that if $\mathcal{F}$ is *any* feature space of size $2^{O\left(n^{1/3}\right)}$ such that every decision list can be uniformly approximated within error $1 - 1/2^{\Omega\left(n^{1/3}\right)}$ by a linear combination of features in S, then $|S| = 2^{\Omega\left(n^{1/3}\right)}$. This result implies Theorem 4 in the case $d = n^{1/3}$, and holds for arbitrary classes of functions $\mathcal{F}$, not just for classes closed under restrictions.

Second, we clarify that given any set $\mathcal{F}$ of boolean-valued functions, the decision list L whose existence is guaranteed by Theorem 4 is actually a restriction of $\textsf{ODD-MAX-BIT}_n$ – this is easily seen by examining the proof of the theorem. However, the specific decision list L necessarily depends on $\mathcal{F}$. Indeed for any fixed decision list L, there exists a trivial set $\mathcal{F}$ which computes L exactly (specifically, the unique function in $\mathcal{F}$ is L itself).

Third, we note that Theorem 4 implies a new lower bound on the weight of degree-d GPTFs that compute $\textsf{ODD-MAX-BIT}_n$. A $2^{\Omega(n/d^2)}$ weight lower bound follows immediately from Beigel's weight lower bound for degree-d PTFs over $\{0,1\}^n$ when $d \leq n^{1/3}$ (see the discussion preceding Theorem 1), but not for larger d. For $d = n^\alpha$ where $1/3 < \alpha < 1/2$, Theorem 4 can be seen to imply that any degree-d GPTF for $\textsf{ODD-MAX-BIT}_n$ must have weight $2^{\Omega\left(n/(d^2 \log^2 n)\right)}$. Indeed, the set $\mathcal{F}$ of all conjunctions of width d has size $n^{O(d)} = 2^{O(d \log n)}$. Thus, applying Theorem 4 to

$\mathcal{F}$ guarantees the existence of some decision list L such that any degree d GPTF for L requires a coefficient of size $2^{\Omega\left(n/(d^2 \log^2 n)\right)}$. Since L is in fact a restriction of $\mathsf{ODD\text{-}MAX\text{-}BIT}_n$, this lower bound applies to $\mathsf{ODD\text{-}MAX\text{-}BIT}_n$ itself.

6 Conclusion

We have given new constructions of low-weight GPTFs, and new weight lower bounds for ordinary PTFs, for decision lists. Our new constructions yield improved tradeoffs between runtime and mistake bounds for learning decision lists, while our lower bounds imply that new algorithms or analyses are needed to improve over current results.

A number of interesting questions remain open. For example, it would be interesting to generalize our $2^{\Omega(\sqrt{n/d})}$ weight lower bound for decision lists to more general feature spaces. In particular, generalizing the lower bound to arbitrary length-d conjunctions would show that the weight upper bound of Theorem 1 is the best possible for GPTFs. It would also be interesting to prove stronger lower bounds for PTF weight-degree tradeoffs for richer classes of functions such as DNFs or decision trees.

References

[1] Richard Beigel. Perceptrons, PP, and the Polynomial Hierarchy. *Computational Complexity*, 4:339–349, 1994.

[2] Avrim Blum. Learning boolean functions in an infinite atribute space (extended abstract). In *STOC*, pages 64–72. ACM, 1990.

[3] Avrim Blum, Lisa Hellerstein, and Nick Littlestone. Learning in the presence of finitely or infinitely many irrelevant attributes. *J. Comput. Syst. Sci.*, 50(1):32–40, 1995.

[4] Avrim Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artif. Intell.*, 97(1-2):245–271, 1997.

[5] Peter Borwein and Tamás Erdélyi. Markov-Bernstein type inequalities under Littlewood-type coefficient constraints. *Indagationes Mathematicae*, 11(2):159 – 172, 2000.

[6] Harry Buhrman. On computation and communication with small bias. In *In Proc. of the 22nd Conf. on Computational Complexity (CCC*, pages 24–32, 2007.

[7] Adam R. Klivans and Rocco A. Servedio. Toward attribute efficient learning of decision lists and parities. *Journal of Machine Learning Research*, 7:587–602, 2006.

[8] Adam R. Klivans and Alexander A. Sherstov. Lower bounds for agnostic learning via approximate rank. *Computational Complexity*, 19(4):581–604, 2010.

[9] Matthias Krause and Pavel Pudlák. On computing boolean functions by sparse real polynomials. In *FOCS*, pages 682–691. IEEE Computer Society, 1995.

[10] Matthias Krause and Pavel Pudlák. Computing boolean functions by polynomials and threshold circuits. *Computational Complexity*, 7(4):346–370, 1998.

[11] N. Littlestone. From online to batch learning. In *Proceedings of the Second Annual Workshop on Computational Learning Theory*, pages 269–284, 1989.

[12] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1987.

[13] P. Long and R. Servedio. Attribute-efficient learning of decision lists and linear threshold functions under unconcentrated distributions. In *Proc. 20th Annual Conference on Neural Information Processing Systems (NIPS)*, 2006.

[14] Ziv Nevo and Ran El-Yaniv. On online learning of decision lists. *Journal of Machine Learning Research*, 3:271–301, 2002.

[15] Noam Nisan and Mario Szegedy. On the degree of boolean functions as real polynomials. *Computational Complexity*, 4:301–313, 1994.

[16] Leslie G. Valiant. Projection learning. *Machine Learning*, 37(2):115–130, 1999.

A Proof of Theorem 2: Weight Upper Bounds for PTFs over $\{-1,1\}^n$.

Since Theorem 1 employs GPTFs rather than PTFs, it does not show that the weight lower bound of Theorem 3 is tight. It is thus a natural question to determine whether the lower bound of Theorem 3 is tight for PTFs. Towards this goal, Theorem 2 gives an upper bound showing that Theorem 3 is close to tight for PTFs over $\{-1,1\}^n$. This result improves on the PTF weight upper bound in Klivans and Servedio [7] for any $d \geq n^{1/2}$. We restate Theorem 2 for the reader's convenience:

Theorem 2 *For any $n^{1/4} \leq d \leq n$, any decision list f on n variables has a degree-d PTF over the domain $\{-1,1\}^n$ of weight $2^{(n/d)^{2/3} \cdot \mathrm{polylog}(n)}$.*

Intuitively, the difficulty in extending Theorem 1 to PTFs over $\{0,1\}^n$ is that in the inner construction, the bottom AND gates in the proof of Claim 2 may not have low-weight exact representations over $\{0,1\}^n$ even if we allow very high degree. Instead, we will use the fact that over $\{-1,1\}^n$ any AND gate has a low-weight *uniform approximation.* More specifically, the following fact is an easy extension of [10, Proposition 2.1] (we note explicitly that the function $\mathsf{AND}^{(j)}$ in the following fact does not have any dependence on j; we use this notation because it will be useful in the context of the proof of Theorem 2):

Fact 4 *Let $\mathsf{AND}^{(j)} : \{-1,1\}^n \to \{0,1\}$ denote some conjunction of exactly n literals, so there is a distinguished input $z \in \{-1,1\}^n$ such that $\mathsf{AND}^{(j)}(z) = 1$ and $\mathsf{AND}^{(j)}(x) = 0$ for all $x \in \{-1,1\}^n \setminus \{z\}$. Then for any $\varepsilon > 0$, there is a polynomial $g(x_1,\ldots,x_n)$ of degree at most n such that*

1. *All coefficients of g are integers, and the sum of their absolute values is $O(n/\epsilon^2)$;*

2. *There is an integer $C_1 = O(n/\epsilon^2)$ such that $|g_j(y) - C_1\mathsf{AND}^{(i)}(y)| \leq C_1\epsilon$ for all $y \in \{-1,1\}^n$; and*

3. $g_j(-z) = 0$. *That is, g evaluates to precisely* 0 *on the input in* $\{-1,1\}^n$ *that satisfies none of the literals in* $\mathsf{AND}^{(j)}$.

With this fact in hand we now prove Theorem 2.

Proof: We break f into $\ell := (n/d)^{2/3}$ modified decision lists as in Equation 1, each of length $r = n/\ell = n^{1/3}d^{2/3}$. The proof uses the following variant of Claim 2.

Claim 3 *Fix $d \geq n^{1/3}$ and let f_i be a modified decision list of length $r = n^{1/3}d^{2/3}$. There is an degree-$\tilde{O}(d)$ polynomial $p_i(x_1, \ldots, x_r)$ with integer coefficients such that*

- *The weight of p_i is $2^{\tilde{O}\left((n/d)^{2/3}\right)}$;*
- *There is an integer $C = 2^{\tilde{O}((n/d)^{2/3})}$, such that for every input $x \in \{-1,1\}^r$ that satisfies at least one literal in f_i, we have $|p(x) - Cf_i(x)| \leq C/r$; and*
- *$f_i(x) = 0$ implies $p_i(x) = 0$.*

Proof: Recall that f_i may be expressed as $f_i(x) = b_1T_1(x) + \cdots + b_rT_r$, where T_j is a conjunction of length $j \leq r$. We view T_j as an AND-of-ANDs where

- The top AND has fanin $t := \lceil jr/d^3 \rceil \leq \lceil r^2/d^2 \rceil = \lceil (n/d)^{2/3} \rceil$,
- Each bottom AND has fanin at most $b := \lceil j/t \rceil \leq \lceil d^2/r \rceil = \lceil \frac{d^{4/3}}{n^{1/3}} \rceil$. Note that $\frac{d^{4/3}}{n^{1/3}} \geq 1$ since $d \geq n^{1/4}$ by assumption.

Compared to the proof of Claim 2, we have reduced the top fan-in to minimize the propagation of error due to the approximations we will use for the bottom AND gates.

Let $\mathsf{AND}_t(y_1, \ldots, y_t)$ denote the top AND gate, so the inputs $y_1, \ldots, y_t$ are the bottom-level ANDs (which take values in $\{0,1\}$). To construct p we apply [7, Corollary 9] to AND_t to obtain a polynomial q of degree $d' = \tilde{O}(\sqrt{t})$ and weight $2^{\tilde{O}(\sqrt{t})}$, and an integer $C_2 = 2^{\tilde{O}(\sqrt{t})}$ such that $|q(y_1, \ldots, y_t) - C_2 \cdot \mathsf{AND}_t(y_1, \ldots, y_t)| \leq \frac{C_2}{t}$ for all $y \in \{0,1\}^t$. Moreover, $q(0^t) = 0$.

The inputs to the top AND gate are themselves conjunctions of width b over $\{-1,1\}$ inputs. Denote the j-th bottom-level AND gate by $\mathsf{AND}^{(j)}$. Set $\epsilon = 2^{-\tilde{\Theta}(\sqrt{t})}$, and let g_j denote the polynomials for the functions $\mathsf{AND}^{(j)}$ given by Fact 4. We obtain our final representation p by composing q with the functions g_j/C_1, and then scaling the resulting coefficients to be integers. More precisely, viewing the polynomials g_j as functions over r variables (though only t variables are relevant for each g_j), the polynomial p is $p(x_1, \ldots, x_r) = (C_1)^{d'} q(\frac{g_1(x)}{C_1}, \ldots, \frac{g_t(x)}{C_1})$. This polynomial p has degree $\tilde{O}(\sqrt{t}) \cdot b = \tilde{O}((n/d)^{1/3}) \cdot \frac{d^{4/3}}{n^{1/3}} = \tilde{O}(d)$ as desired. It has integer coefficients, and the sum of the absolute values of its coefficients is at most $(n/\varepsilon^2)^{\tilde{O}(\sqrt{t})} = 2^{\tilde{O}(t) \cdot \log n} = 2^{(n/d)^{2/3} \cdot \text{polylog}(n)}$.

All that remains to show is that there is an integer $C = 2^{(n/d)^{2/3} \cdot \text{polylog}(n)}$ such that for every input $x \in \{-1,1\}^r$, it holds that $|p(x) - Cf_i(x)| \leq C/r$, and that $f_i(x) = 0$ implies $p(x) = 0$. The latter property follows from the fact that each $g_j(0^t) = 0$ for all j, and $q(0^t) = 0$. To see the first property, notice that q has at most $2^{\tilde{O}(\sqrt{t})} \leq \frac{1}{r\epsilon}$ monomials, since it has degree $d' = \tilde{O}(\sqrt{t})$. Since each $\frac{g_j}{C_1}$ is a pointwise ϵ-approximation to the j'th bottom gate $\mathsf{AND}^{(j)}$, for each monomial the difference between its value evaluated at $\left(\mathsf{AND}^{(1)}(x), \ldots, \mathsf{AND}^{(t)}(x)\right)$ and at $\left(\frac{g_1(x)}{C_1}, \ldots, \frac{g_t(x)}{C_1}\right)$ is most ϵ. Thus, letting $C = C_2 \cdot C_1^{d'}$, the total error $|p(x) - C \cdot f_i(x)|$ at any point is at most C/r. This concludes the proof of the claim. ■

Theorem 2 now follows by replacing each modified decision list f_i with the corresponding polynomial p_i from Claim 3. The result is a PTF P over $\{-1,1\}^n$ for f which has weight $2^{(n/d)^{2/3} \cdot \mathrm{polylog}(n)}$ and has degree $\tilde{O}(d)$. The analysis showing that $\mathrm{sign}(P(x)) = f(x)$ for all $x \in \{-1,1\}^n$ follows exactly as in Theorem 1 and is omitted. ■

B Proof of Theorem 3 over domain $\{-1,1\}^n$.

We describe the changes to the proof for domain $\{0,1\}^n$ that are required to obtain the lower bound for a PTF p over domain $\{-1,1\}^n$. As in the $\{0,1\}^n$ proof we may assume that $V(y_{i^*}) < 0$. Given a string $w \in \{-1,1\}^{\ell/2}$ we now let $z_w \in \{-1,1\}^n$ be the input defined as follows:

- the first i^* blocks of z_w agree with y_{i^*};
- the even bits in the (i^*+1)-th block are all 1;
- the odd bits in the (i^*+1)-th block are successively $w_1, w_2, \ldots, w_{\ell/2}$; and
- the remaining blocks $i^*+2, \ldots, n/\ell$ have all bits set to 1.

As before, define q to be the function which, on input $w \in \{-1,1\}^{\ell/2}$, outputs $p(z_w)$. Also as before, define

$$Q(t) := \mathbb{E}_{x \leftarrow \mu_t}[q(x)],$$

where now μ_t is the product distribution over the Hamming cube $\{-1,1\}^{\ell/2}$ where each coordinate j satisfies $\mathbb{E}_{x \leftarrow \mu_t}[x_j] = t$ (so each x_j is independently 1 with probability $\frac{1+t}{2}$ rather than with probability t as before). The earlier arguments show that Q satisfies the following properties:

1. $Q(-1) = q(-1, \ldots, -1)$ and $Q(1) = q(1, \ldots, 1)$.
2. $|Q(t)| \leq 2^{i^*}$ for all $t \in [-1,1]$.
3. $\deg(Q) \leq \deg(q) \leq \deg(p)$.
4. The sum of the absolute value of the coefficients of Q is at most W.

Finally, we now define P as

$$P(t) := Q(t) - Q(-1).$$

The earlier arguments imply that $P(t)$ is a degree-d polynomial with the following properties:

1. The coefficients of P have absolute value at most $W + |V(y_{i^*})| \leq 2W$.
2. $1/2 \leq \max_{x \in [-1,1]} |P(x)| \leq 2^{i^*+1}$.
3. $P(-1) = 0$ and $P(-1 + (4/\ell)) \geq 2^{i^*-2}$. In particular, there exists some $t \in [-1, -1 + (4/\ell)]$ such that $P'(t) \geq \ell \cdot 2^{i^*-4}$.

Note that the main difference compared to the argument for the $\{0,1\}^n$ domain is that in item (3) above, the "jump" now occurs near -1, not in the domain $[0,1]$, so we cannot directly invoke Lemma 1. On the other hand, the polynomial P stays bounded on the larger domain $[-1,1]$ rather than $[0,1]$, and so we can easily transform P into a polynomial which satisfies all of the conditions of Lemma 1. Indeed, defining $P_1(x) = P(-x)$, it is easy to check that we may apply Lemma 1 to P_1 to conclude that $W = 2^{\Omega\left(\sqrt{n/d}\right)}$ as desired. ■

C Proof of Theorem 4

Proof Outline. Fix a polynomial p which sign-represents $\mathsf{ODD\text{-}MAX\text{-}BIT}$. In proving his lower bound, Beigel breaks the n variable domain into blocks of length ℓ, and iteratively constructs inputs x_i such that $|p(x_i)|$ grows exponentially. We explained in Section 4 that at each stage the existence of an input x_i with $|p(x_i)| \geq 2|p(x_{i-1})|$ follows from Markov's inequality. However, another way to understand the existence proof is as follows. If no such x_i existed, then we could in fact obtain a degree d polynomial which uniformly approximates the OR function on ℓ variables. By a result of Nisan and Szegedy [15], such a polynomial requires degree $\Omega(\sqrt{\ell})$.

In [8, Theorem 4.2], the authors prove a generalization of Nisan and Szegedy's result, showing that if one wants to uniformly approximate *all* disjunctions on ℓ variables using *any* feature space, then the size of the feature space must be $2^{\Omega(\sqrt{\ell})}$. It is therefore plausible that Beigel's argument can be extended to more general feature spaces, using the feature-independent result of Klivans and Sherstov in place of Nisan and Szegedy, and indeed this is how our proof proceeds. Details follow.

Proof of Theorem 4. Fix $\ell := \Theta(d^2)$. We will iteratively consider decision lists defined on $i \cdot \ell$ variables, for each $i \in \{1, \dots, n/\ell\}$. Refer to the set of all such decision lists as $\mathcal{DL}_i$, and refer to the function $\mathsf{ODD\text{-}MAX\text{-}BIT}$ defined over $i \cdot \ell$ variables as $\mathsf{ODD\text{-}MAX\text{-}BIT}_i$. Furthermore, let $\mathcal{F}_i$ denote the set of features over $i \cdot \ell$ variables obtained from $\mathcal{F}$ by "ignoring" the last $n - i$ variables in each feature. That is,

$$\mathcal{F}_i = \{\psi : \{0,1\}^{i \cdot \ell} \to \{0,1\} | \psi_i(x) = \phi_i(x \circ \mathbf{0}) \text{ for some } \phi \in \mathcal{F}\}.$$

Notice that since $\mathcal{F}$ is closed under restrictions, $\mathcal{F}_i$ is as well.

We iteratively construct decision lists $L_i \in \mathcal{DL}_i$ such that for *every* representation $L_i(x) = \text{sign}(h(x))$ of L_i as the sign of a integer linear combination of the features in $\mathcal{F}_i$, there exists a $y_i(L_i) \in \{0,1\}^{\ell \cdot i}$ such that $|h(y_i(L_i))| \geq 2^{i-1}$. Moreover, each L_i will be obtained from $\mathsf{ODD\text{-}MAX\text{-}BIT}_i$ by restricting some of the variables in block i. If we succeed at all iterations, it clearly follows that the decision list in the final iteration satisfies the properties guaranteed by the theorem. Notice we can take L_1 to simply be $\mathsf{ODD\text{-}MAX\text{-}BIT}_1$, as we can assume all sign-representations of L_1 have margin at least 1.

Suppose this process fails at stage $i > 1$. Then

1. There is a decision list $L_i \in \mathcal{DL}_i$ such that for *every* sign-representation h for L_i as an integer linear combination of features in $\mathcal{F}_i$, there is an input $y_i \in \{0,1\}^{i \cdot \ell}$ such that $|h(y_i)| \geq 2^{i-1}$. Moreover, L_i is obtained from $\mathsf{ODD\text{-}MAX\text{-}BIT}_i$ by restricting some of the variables.

2. For all decision lists $L \in \mathcal{DL}_{i+1}$, there exists a sign-representation $L(x) = \text{sign}(h(x))$ such that $|h(y_{i+1})| \leq 2^i$ for all $y_{i+1} \in \{0,1\}^{(i+1) \cdot \ell}$.

We show that there is a feature space $\mathcal{F}'$ with $|\mathcal{F}'| \leq |\mathcal{F}|^5$ which uniformly approximates the set of *all* disjunctions on $\ell/2$ variables within additive error 1/3. By [8, Theorem 4.2], this implies that $|\mathcal{F}'| = 2^{\Omega(d)}$, and hence $\mathcal{F} = 2^{\Omega(d)}$ as well.

Let $y_i \in \{0,1\}^{i \cdot \ell}$ be any vector, and assume that $L_i(y_i) < 0$. Consider the set $\mathcal{S}$ of decision lists in $\mathcal{DL}_{i+1}$ which are obtained from $\mathsf{ODD\text{-}MAX\text{-}BIT}_{i+1}$ by

- Restricting variables in blocks $\{1, \ldots, i\}$ in the same manner as L_i;
- Restricting all even variables in block $i+1$ to 0;
- Restricting some subset of the odd variables in block $i+1$ to 0.

There is a natural one-to-one correspondence between decision lists in $\mathcal{S}$ and disjunctions on $\ell/2$ variables: if $L_S \in \mathcal{S}$ denotes the decision list such that the "active" odd variables in block i are precisely those in S, then L_S naturally corresponds to the disjunction $\mathsf{OR}_S(x_1, \ldots, x_{\ell/2}) := \vee_{j \in S} x_j$. The key observation is that any sign-representation h_{L_S} for L_S as an integer linear combination of features in $\mathcal{F}_{i+1}$ can be transformed into a uniform approximation for OR_S, using features which are themselves (products of) restrictions of features in $\mathcal{F}_{i+1}$. Since $\mathcal{F}_i$ is closed under restrictions, this means that we have in fact obtained a uniform approximation for OR_S using features which are themselves products of those in $\mathcal{F}_{i+1}$.

Indeed, consider the disjunction $\mathsf{OR}_S(x_1, \ldots, x_{\ell/2}) := \vee_{j \in S} x_j$. Let $h_{L_S}(x) = \sum_{\phi_j \in \mathcal{F}_{i+1}} c_j \phi_j(x)$ be the sign-representation of L_S guaranteed by Property 2 above. Then we conclude the following.

- There is some y_i such that $W := |h_{L_S}(y_i \circ \mathbf{0})| \geq 2^{i-1}$. This holds by Property 1 above because $L_i(y_i) = \text{sign}(h_{L_S}(y_i \circ \mathbf{0}))$, and the function $g(y) = h_{L_S}(y \circ \mathbf{0})$ is an integer linear combination of features in $\mathcal{F}_i$.
- $h_{L_S}(y_{i+1}) \leq 2^i$ for all $y_{i+1} \in \{0,1\}^{(i+1)\cdot \ell}$ by definition of h_{L_S}.
- Assume without loss of generality that $L_i(y_i) < 0$. Then $\text{OR}_S(x_1, \ldots, x_{\ell/2}) = L_S(y_{i^*} \circ \tilde{x})$, where $\tilde{x} \in \{0,1\}^\ell$ denotes the vector obtained by placing x_i in the ith odd coordinate, and placing 0 in all even coordinates.

Consider the function $h_1 : \{0,1\}^{(i+1)\cdot\ell} \to \{0,1\}$ given by $h_1(x) = h_{L_S}(x) - h_{L_S}(y_i \circ \mathbf{0})$. Trivially, $h_1(y_i \circ \mathbf{0}) = 0$, and the first and second bullets above imply that, $W \leq h_1(x) \leq 3W$ for all x. Combining this with the third bullet point above, we conclude that the function

$$h_2(x_1, \ldots, x_{\ell/2}) = 2 \cdot \left(\frac{h_{L_S}(y_i \circ \tilde{x})}{3W} - 1 \right)^5 + 1$$

uniformly approximates OR_S to within error 1/3. Indeed, $h_2(\mathbf{0}) = -1$, and for all $x \neq \mathbf{0}$, $h_2(x) \in [2/3, 1]$. Notice h_2 is a linear combination of features which can be written as the product of at most 5 features in $\mathcal{F}_{i+1}$ (possibly after adding the constant-function to $\mathcal{F}_{i+1}$). As *every* disjunction over $\ell/2$ variables can be uniformly approximated in this way, we conclude from [8, Theorem 4.2], that $\mathcal{F}_{i+1} \geq 2^{\Omega(d)}$. Since $|\mathcal{F}_{i+1}| \leq |\mathcal{F}|$, we conclude that $|\mathcal{F}| = 2^{\Omega(d)}$, completing the proof.