Mach Learn (2007) 69: 97–114
DOI 10.1007/s10994-007-5010-1

Unconditional lower bounds for learning intersections of halfspaces

Adam R. Klivans · Alexander A. Sherstov

Received: 9 September 2006 / Revised: 18 March 2007 /
Accepted: 13 April 2007 / Published online: 25 May 2007

Abstract We prove new lower bounds for learning intersections of halfspaces, one of the most important concept classes in computational learning theory. Our main result is that any statistical-query algorithm for learning the intersection of $\sqrt{n}$ halfspaces in n dimensions must make $2^{\Omega(\sqrt{n})}$ queries. This is the first non-trivial lower bound on the statistical query dimension for this concept class (the previous best lower bound was $n^{\Omega(\log n)}$). Our lower bound holds even for intersections of *low-weight* halfspaces. In the latter case, it is nearly tight.

We also show that the intersection of two majorities (low-weight halfspaces) cannot be computed by a polynomial threshold function (PTF) with fewer than $n^{\Omega(\log n/\log\log n)}$ monomials. This is the first super-polynomial lower bound on the PTF length of this concept class, and is nearly optimal. For intersections of $k = \omega(\log n)$ low-weight halfspaces, we improve our lower bound to $\min\{2^{\Omega(\sqrt{n})}, n^{\Omega(k/\log k)}\}$, which too is nearly optimal. As a consequence, intersections of even two halfspaces are not computable by polynomial-weight PTFs, the most expressive class of functions known to be efficiently learnable via Jackson's Harmonic Sieve algorithm. Finally, we report our progress on the *weak* learnability of intersections of halfspaces under the uniform distribution.

Keywords Intersections of halfspaces · Halfspace learning · PAC learning · SQ learning · Statistical queries · Query learning · Lower bounds for learning · Polynomial threshold functions · Harmonic sieve

1 Introduction

Learning intersections of halfspaces is a fundamental and well-studied problem in computational learning theory. In addition to generalizing well-known concept classes such as DNF

Editors: Hans Ulrich Simon, Gabor Lugosi, Avrim Blum.

A.R. Klivans (✉) · A.A. Sherstov
Department of Computer Sciences, The University of Texas at Austin, Austin, TX 78712, USA
e-mail: klivans@cs.utexas.edu

A.A. Sherstov
e-mail: sherstov@cs.utexas.edu

formulas, intersections of halfspaces are capable of representing arbitrary convex sets. While many efficient algorithms exist for PAC learning a *single* halfspace, the problem of learning the intersection of even two halfspaces remains a difficult challenge. A variety of efficient algorithms have been developed for learning natural restrictions of intersections of halfspaces in various learning models (Vempala 1997; Klivans et al. 2004; Klivans and Servedio 2004; Kwek and Pitt 1998).

Progress on proving *hardness* results for learning intersections of halfspaces has been more limited. Klivans and Sherstov (2006) have recently given the first representation-independent (cryptographic) hardness results for PAC learning intersections of halfspaces. Feldman et al. (2006) have obtained closely related results. The only other relevant hardness results are for representation-dependent (proper) learning: if the learner's output hypothesis must be from a restricted class of functions (e.g., intersections of halfspaces), then the learning problem in question is NP-hard with respect to randomized reductions (Alekhnovich et al. 2004).

The PAC hardness results surveyed above are *conditional*, i.e., they depend on widely believed but unproven assumptions from cryptography or complexity theory. Our paper complements that work by proving lower bounds that are *unconditional* but valid only for a restriction of the PAC model. Specifically, we study the problem of learning intersections of halfspaces in Kearns' *statistical query* model of learning (Kearns 1993), an elegant restriction of Valiant's PAC model (Valiant 1984). A learner in the statistical query model is allowed queries of the form "What is $\Pr_{x\sim\mu}[Q(x, f(x)) = 1]$, approximately?" Here μ is the underlying distribution on $\{-1,1\}^n$, the function $Q : \{-1,1\}^n \times \{-1,1\} \to \{-1,1\}$ is a polynomial-time computable predicate, and $f : \{-1,1\}^n \to \{-1,1\}$ is the unknown concept. The motivation behind the statistical query model is that efficient algorithms in this model are robust to classification noise. Kearns showed that concept classes learnable via a polynomial number of statistical queries are efficiently PAC learnable. Perhaps surprisingly, virtually all known PAC learning algorithms can be adapted to work via statistical queries only; the one exception known to us is the algorithm of Blum et al. (2003) for learning parity functions.

The *SQ dimension* of a concept class $\mathcal{C}$ under distribution μ is defined as the size of the largest subset $\mathcal{A} \subseteq \mathcal{C}$ of concepts such that the elements of $\mathcal{A}$ are "almost" orthogonal under μ (see Sect. 2.2 for a precise definition). Blum et al. (1994) proved the SQ dimension of a concept class to be a measure of the number of statistical queries required to learn that class. It is well known that the concept class of parity functions has SQ dimension 2^n (the maximum possible) under the uniform distribution. This observation has been the basis of all known statistical query lower bounds.

1.1 Our results

Our main contribution is a lower bound for learning intersections of halfspaces in the statistical query model. We construct distributions under which intersections of halfspaces have a large SQ dimension. Let MAJ_k denote the concept class of intersections of k majorities, a subclass of intersections of halfspaces.

Theorem 1.1 *There are* (*explicitly given*) *distributions on* $\{-1,1\}^n$ *under which*

$$\mathrm{sqdim}(\mathrm{MAJ}_k) = \begin{cases} n^{\Omega(k/\log k)} & \text{if } \log n \leqslant k \leqslant \sqrt{n}, \\ \max\{n^{\Omega(k/\log\log n)}, n^{\Omega(\log k)}\} & \text{if } k \leqslant \log n. \end{cases}$$

Our result is essentially optimal. Namely, the SQ dimension of MAJ_k (and more generally, of intersections of k polynomial-weight halfspaces) is known to be at most $n^{O(k\cdot\log k\cdot\log n)}$ under all distributions. For completeness, we recall a proof of this upper bound in Sect. 4. An illustrative instantiation of our main theorem is as follows: for any constant $0<\epsilon\leqslant 1/2$, the intersection of n^ϵ halfspaces has SQ dimension $2^{\Omega(n^\epsilon)}$, the known upper bound being $2^{O(n^\epsilon\log^3 n)}$.

The previous best lower bound for this concept class was $n^{\Omega(\log n)}$. The $n^{\Omega(\log n)}$ bound holds even for n^ϵ-term DNF, a subclass of the intersection of n^ϵ halfspaces. The proof is as follows. A DNF formula with 2^t terms can compute any function on t variables. Thus, a polynomial-size DNF can compute parity on any subset of $\log n$ variables. Since any two distinct parity functions are orthogonal under the uniform distribution, the SQ dimension of polynomial-size DNF is at least $\binom{n}{\log n}=n^{\Omega(\log n)}$.

Our second contribution is a series of lower bounds for the representation of MAJ_k as a polynomial threshold function (PTF). Jackson gave the first polynomial-time algorithm, the celebrated *Harmonic Sieve* (Jackson 1995), for learning polynomial-size DNF formulas with membership queries under the uniform distribution. More generally, he showed that the concept class of polynomial-weight PTFs is learnable in polynomial time using the Harmonic Sieve. A natural question to ask is whether every intersection of k low-weight halfspaces, a straightforward generalization of k-term DNF, can be represented as a polynomial-weight PTF. We answer this question in the negative even for $k=2$. Let MAJ denote the majority function, which can be represented as the low-weight halfspace $\sum x_i\geqslant 0$. We prove that the intersection of two majority functions requires not only large weight but also large length:

Theorem 1.2 *The function* $\mathrm{MAJ}(x_1,\dots,x_n)\wedge\mathrm{MAJ}(y_1,\dots,y_n)$ *requires PTF length* $n^{\Omega(\log n/\log\log n)}$.

The lower bound of Theorem 1.2 nearly matches the $n^{O(\log n)}$ upper bound of Beigel et al. (1995), proving that their PTF construction is essentially optimal. As a corollary to Theorem 1.2, we observe that intersections of even two low-weight halfspaces cannot be computed by polynomial-weight PTFs, the most expressive class of concepts known to be learnable via Jackson's Harmonic Sieve. We note here that intersections of a constant number of halfspaces are learnable with membership and equivalence queries in polynomial time via Angluin's algorithm for learning finite automata. For the case of intersections of $k=\omega(1)$ halfspaces, however, no polynomial-time algorithms are known. For this case, we prove PTF length lower bounds with an exponential dependence on k:

Theorem 1.3 *Let* $k\leqslant\sqrt{n}$. *Then there are* (*explicitly given*) *functions in* MAJ_k *that require PTF length* $n^{\Omega(k/\log k)}$.

This lower bound is almost tight: Klivans et al. (2004, Theorem 29), have shown that every function in MAJ_k has a PTF of length $n^{O(k\cdot\log k\cdot\log n)}$. Note that Theorem 1.3 improves on Theorem 1.2 for $k=\omega(\log n)$.

Finally, we consider the feasibility of learning intersections of halfspaces *weakly* in polynomial time under the uniform distribution. (Recall that strong learning refers to constructing a hypothesis with error ϵ in time $\mathrm{poly}(n,1/\epsilon)$; weak learning refers to constructing a hypothesis with error $1/2-1/\mathrm{poly}(n)$ in time $\mathrm{poly}(n)$.) We report our progress on this problem in Sect. 5, proving negative results for generalizations of the problem and positive results for several restricted cases.

1.2 Our techniques

Most of our results follow from a variety of new applications of *bent* functions, i.e., functions whose Fourier coefficients are as small as possible. Although the Fourier analysis of Boolean functions is usually relevant only to uniform-distribution learning, we apply an observation due to Bruck (1990) that the flatness of a function's spectrum is directly related to the length of its PTF representation, a quantity involved with arbitrary-distribution learning. We construct non-uniform distributions under which various intersections of low-weight halfspaces are capable of computing bent functions. This in turn yields a variety of lower bounds on their PTF length, depending on the construction we employ. We then extend the construction of a single bent function to a family of bent functions and prove that this yields a large set of orthogonal functions, the critical component of our SQ dimension lower bound. All functions and distributions we construct are explicitly defined.

For the near-optimal lower bound on the PTF length of the intersection of two majority functions, we combine results on the PTF degree of intersections of halfspaces due to O'Donnell and Servedio (2003) with a translation lemma in circuit complexity due to Krause and Pudlák (1997).

1.3 Organization

We first prove PTF length lower bounds for intersections of majorities in Sect. 3. We build on these results to prove our main SQ dimension lower bound in Sect. 4. Our discussion of weak learning appears in Sect. 5.

2 Preliminaries

A *Boolean function* is a mapping $\{-1,1\}^n \to \{-1,1\}$, where 1 corresponds to "true." In this representation, the parity χ_S of a set $S \subseteq [n]$ of bits is given by the product of the corresponding variables: $\chi_S \overset{\text{def}}{=} \bigoplus_{i\in S} x_i = \prod_{i\in S} x_i$. A *majority function* is a Boolean function of the form

$$\operatorname{sign}(x_{j_1} + x_{j_2} + \cdots),$$

where the x_{j_i} are distinct variables from among $x_1, \ldots, x_n$. A generalization of majority is a *halfspace*

$$\operatorname{sign}(a_1 x_{j_1} + a_2 x_{j_2} + \cdots),$$

where the a_i are integer weights. Finally, a *polynomial threshold function* (PTF) has the form

$$\operatorname{sign}(a_1\chi_1 + a_2\chi_2 + \cdots),$$

where the a_i are integer coefficients and the χ_i are distinct parity functions over $x_1, \ldots, x_n$, possibly including the constant function 1. Note that halfspaces and majorities are PTFs. One can assume w.l.o.g. that the polynomial $a_1\chi_1 + a_2\chi_2 + \cdots$ sign-representing a PTF is nonzero on all inputs.

Two important characteristics of PTFs from a learning standpoint are its weight and length. The *weight* of a PTF $\operatorname{sign}(\sum_i a_i\chi_i)$ is $\sum_i |a_i|$. The *length* of a PTF is the number of monomials, i.e., distinct parity functions. Thus, a PTF's weight is never less than its length. A PTF is *light* (respectively, *short*) if its weight (respectively, length) is bounded by a polynomial in n.

In the above description, the polynomial (weighted sum of parities) computing a PTF f agrees in sign with f on every input. We refer to this type of sign-representation as *strong*: a polynomial p *strongly* represents a Boolean function f iff for all x we have $p(x) \neq 0$ and $f(x) = \operatorname{sign}(p(x))$. We will also need the following relaxed version of threshold computation (Saks 1993): a polynomial p *weakly* represents a Boolean function f iff $p(x) \neq 0$ for some x, and $f(x) = \operatorname{sign}(p(x))$ on any such x. We say that a function has a strong (respectively, weak) representation on a set of parities $\mathcal{A} \subseteq \mathcal{P}([n])$ iff there is a polynomial $\sum_{S \in \mathcal{A}} a_S \chi_S$ that strongly (respectively, weakly) represents f. The following is a useful tool in analyzing PTFs.

Theorem 2.1 (Theorem of the Alternative, Aspnes et al. 1994; O'Donnell and Servedio 2003) *Let $\mathcal{A} \subseteq \mathcal{P}([n])$ denote any set of parities on $x_1, \ldots, x_n$, and let $\mathcal{P}([n])$ denote the full set of the 2^n parities. Then for any function $f : \{-1,1\}^n \to \{-1,1\}$, exactly one of the following statements holds*:

(a) *f has a strong representation on $\mathcal{A}$*;
(b) *f has a weak representation on $\mathcal{A}^\perp = \mathcal{P}([n]) \setminus \mathcal{A}$.*

2.1 Fourier transform

Consider the vector space of functions $\{-1,1\}^n \to \mathbb{R}$, equipped with the inner product $\langle f, g\rangle = \mathbf{E}_{x \sim U}[f(x) \cdot g(x)]$. The parity functions $\{\chi_S\}_{S \subseteq [n]}$ form an orthonormal basis for this inner product space. As a result, every Boolean function f can be uniquely written as its *Fourier polynomial*

$$f = \sum_{S \subseteq [n]} \hat{f}(S)\chi_S,$$

where $\hat{f}(S) \stackrel{\text{def}}{=} \langle f, \chi_S\rangle$. Observe that $\hat{f}(\emptyset) = 2\Pr_x[f(x) = 1] - 1$. The f-specific constants $\hat{f}(S)$ are called *Fourier coefficients*. The orthonormality of the parities yields *Parseval's identity* for Boolean functions:

$$\sum_{S \subseteq [n]} \hat{f}(S)^2 = \langle f, f\rangle = 1.$$

As in signal processing, one can obtain an approximation to a function by identifying and estimating its large Fourier coefficients (the "dominant frequencies"). Although there are 2^n coefficients to consider, the large ones can be retrieved efficiently by the elegant algorithm of Kushilevitz and Mansour (1993), to which we refer as "KM":

Theorem 2.2 (Kushilevitz and Mansour 1993) *Let f be any Boolean function and let $\delta, \theta > 0$ be parameters. With probability $\geqslant 1 - \delta$,* KM *outputs every $S \subseteq [n]$ for which $|\hat{f}(S)| \geqslant \theta$, and no $S \subseteq [n]$ for which $|\hat{f}(S)| \leqslant \theta/2$.* KM *runs in time* $\operatorname{poly}(n, \frac{1}{\theta}, \log\frac{1}{\delta})$.

It is thus useful to recognize classes of functions that have large Fourier coefficients. We denote by $\mathbf{L}_\infty(f)$ the largest absolute value of a Fourier coefficient of f. Formally, $\mathbf{L}_\infty(f) \stackrel{\text{def}}{=} \max_S\{|\hat{f}(S)|\}$. This quantity places a lower bound on the length of a PTF computing f:

Theorem 2.3 (Bruck 1990, Theorem 5.11) *Any PTF computing f has length at least $1/\mathbf{L}_\infty(f)$.*

Theorem 2.3 implies that functions with short PTFs are weakly learnable under the uniform distribution:

Proposition 2.4 *Let $\mathcal{C}$ be a class of Boolean functions. If each $f \in \mathcal{C}$ has a PTF of length ℓ, then $\mathcal{C}$ is learnable to accuracy $\frac{1}{2} + \frac{1}{2\ell}$ under the uniform distribution in time* $\operatorname{poly}(n, \ell)$.

Proof Let $f \in \mathcal{C}$ be the unknown target function. In time $\operatorname{poly}(n, \ell)$, KM identifies all parities that predict f with advantage $1/\ell$ or better. It thus suffices to show that for some parity χ, $|\mathbf{E}_x[\chi \cdot f]| \geqslant 1/\ell$. The latter is equivalent to showing that $\mathbf{L}_\infty(f) \geqslant 1/\ell$. But if we had $\mathbf{L}_\infty(f) < 1/\ell$, then any PTF implementing f would require more than ℓ monomials (by Theorem 2.3). Thus, some parity χ predicts f with advantage $1/\ell$ or better. □

Proposition 2.4 shows that PTF *length* is an indicator of weak learnability under the uniform distribution. Additionally, PTF *weight* is an indicator of strong learnability under the uniform distribution: Jackson (1995) proves that the Harmonic Sieve strongly learns an unknown Boolean function if it can be written as a polynomial-weight PTF.

For all $f : \{-1, 1\}^n \to \{-1, 1\}$, we have $\mathbf{L}_\infty(f) \geqslant 2^{-n/2}$ by Parseval's identity. For n even, f is called *bent* if all Fourier coefficients of f are $2^{-n/2}$ in absolute value. It is known (Bruck 1990) that bent functions include *inner product mod 2*

$$\mathrm{IP}_n(x) = (x_1 \wedge x_2) \oplus (x_3 \wedge x_4) \oplus \cdots \oplus (x_{n-1} \wedge x_n)$$

and *complete quadratic*

$$\mathrm{CQ}_n(x) = \begin{cases} 1 & \text{if } (\|x\| \bmod 4) \in \{0, 1\}, \\ -1 & \text{otherwise.} \end{cases}$$

Above and throughout the paper, $\|x\|$ stands for the number of -1 bits in x. In particular, $\|x \oplus y\|$ yields the number of bit positions where x and y differ.

2.2 Statistical query dimension

The *statistical query* model, first defined by Kearns (1993), is an elegant model of learning that can withstand classification noise. The SQ model has proven to be a useful formalism. In fact, a vast majority of today's efficient learning algorithms fit in this framework. The SQ dimension of a concept class, defined shortly, is a tight measure of the hardness of learning in this model. As a result, SQ dimension estimates are of considerable interest in learning theory.

A *concept class* $\mathcal{C}$ is a set of functions $\{-1, 1\}^n \to \{-1, 1\}$. The *statistical query dimension* of $\mathcal{C}$ under distribution μ, denoted $\operatorname{sqdim}_\mu(\mathcal{C})$, is the largest N for which there are N functions $f_1, \ldots, f_N \in \mathcal{C}$ with

$$|\mathbf{E}_{x\sim\mu}[f_i(x) \cdot f_j(x)]| \leqslant \frac{1}{N}$$

for all $i \neq j$. We denote $\operatorname{sqdim}(\mathcal{C}) \overset{\text{def}}{=} \max_\mu\{\operatorname{sqdim}_\mu(\mathcal{C})\}$. The SQ dimension of a concept class fully characterizes its weak learnability in the statistical query model: a low SQ dimension implies an efficient weak-learning algorithm, and a high SQ dimension rules out such an algorithm (see Blum et al. 1994 and Yang 2005, Corollary 1).

2.3 Notation

We adopt the notation $\mathbf{L}_\infty^+(f) \overset{\text{def}}{=} \max_{S\neq\emptyset}\{|\hat{f}(S)|\}$. We denote by MAJ_k the family of functions computable by the intersection of k majorities, each on some subset of the n variables. Throughout the paper, we view k as an arbitrary function of n, including a constant. $\mathrm{MAJ}(x_{i_1}, x_{i_2}, \ldots)$ stands for the majority value of $x_{i_1}, x_{i_2}, \ldots$. We denote the set $\{1, 2, \ldots, a\}$ by $[a]$. $\mathbf{I}[A]$ denotes 1 if the statement A is true, and 0 otherwise. The vector with -1 in the ith position and 1's elsewhere is e_i. In particular, $x \oplus e_i$ represents x with its ith bit flipped.

Recall that a Boolean function is called *monotone* if flipping a bit from -1 to 1 in any input does not decrease the value of the function. For example, the majority function $\sum x_i \geqslant 0$ is monotone. A function $f(x_1, \ldots, x_n)$ is *unate* if $f(\sigma_1 \oplus x_1, \ldots, \sigma_n \oplus x_n)$ is monotone for some fixed $\sigma \in \{-1, 1\}^n$. Here σ is called the *orientation* of f. For example, the function $x_1 - 2x_2 + x_3 - 4x_5 \geqslant 3$ is unate with orientation $\sigma = (1, -1, 1, -1)$.

3 PTF length lower bounds for MAJ_k

We begin by developing lower bounds on the PTF representation of intersections of low-weight halfspaces. In particular, this section establishes two of the main results of this paper: Theorems 1.2 and 1.3. We will also need these structural results to prove our main lower bound on the SQ dimension of intersections of halfspaces.

3.1 PTF length of MAJ_k: an $n^{\Omega(\log k)}$ bound

Unlike the lower bound for MAJ_2, the results in this section and the next require $k = \omega(1)$ for a super-polynomial lower bound. However, they rely solely on the fundamental Theorem 2.3 and are thus considerably simpler. Furthermore, the constructions below (Lemmas 3.3 and 3.5) will allow us to prove a lower bound on the SQ dimension of MAJ_k in Sect. 4. A key to these results is the following observation.

Lemma 3.1 *Let $f(x_1, \ldots, x_n)$ have a PTF of length ℓ. Then so does $f(\chi_1, \ldots, \chi_n)$, where each χ_i is a parity over $x_1, \ldots, x_n$ or the negation of a parity.*

Proof Given a polynomial of length ℓ that strongly sign-represents f, make the replacement $x_i \to \chi_i$. This does not increase the number of monomials, while yielding a PTF for $f(\chi_1, \ldots, \chi_n)$. □

By Lemma 3.1, it suffices to show that $f(\chi_1, \ldots, \chi_n)$ does not have a short PTF in order to prove that neither does $f(x_1, \ldots, x_n)$. We accomplish the former via a reduction to a known hard function.

Definition 3.2 (Reflection) *Let $f : \{-1, 1\}^n \to \{-1, 1\}$ and $y \in \{-1, 1\}^n$. The y-*reflection *of f is the function $f_y(x) = f(x \oplus y)$. A function $g : \{-1, 1\}^n \to \{-1, 1\}$ is called a* reflection *of f if $g(x) = f(x \oplus y)$ for some fixed y and all x.*

We are now in a position to prove the desired reduction to a hard function.

Lemma 3.3 *Let $k \leqslant 2^{n^{o(1)}}$. Then there are explicitly given functions $\chi_1, \chi_2, \ldots, \chi_n$ (each a parity or the negation of a parity) such every reflection of* IP *on $\Omega(\log n \cdot \log k)$ variables is computable by $f(\chi_1, \chi_2, \ldots, \chi_n)$ for some $f \in \mathrm{MAJ}_k$.*

Proof Let $g_1, g_2, \dots, g_{\log k}$ be copies of the IP function, each on a distinct set of variables V_i with $|V_i| = v$ for some $v = v(n,k)$ to be chosen later. Thus, $g = \bigoplus g_i$ is IP on $v \log k$ variables. At the same time, g is computable by the AND of $2^{\log k - 1} < k$ functions, each of the form $h_1 \vee h_2 \vee \dots \vee h_{\log k}$, where $h_i \in \{g_i, \neg g_i\}$. Each $h_1 \vee h_2 \vee \dots \vee h_{\log k}$ can be computed by the PTF

$$\begin{aligned} &h_1 + h_2 + \dots + h_{\log k} \geqslant 1 - \log k, \quad \text{or} \\ &2^{v/2} h_1 + 2^{v/2} h_2 + \dots + 2^{v/2} h_{\log k} \geqslant 2^{v/2}(1 - \log k). \end{aligned} \tag{3.1}$$

Every h_i is a bent function on the v variables V_i, and thus $2^{v/2} h_i$ is simply the sum of the 2^v parities on V_i, each with a plus or a minus sign.

Create a new set of variables $U = \{\chi_1, \chi_2, \dots\}$ as follows. U will contain a distinct variable for each parity on V_i (for each $i = 1, 2, \dots, \log k$) and one for its negation. In addition, U will contain $2^{v/2}(\log k - 1) < 2^{v/2} \log k$ variables, each of which corresponds to the constant 1. As a result, each of the k PTFs of the form (3.1) is a majority function in terms of U. Therefore, IP(x) on $v \log k$ variables is computable by $f(\chi_1, \chi_2, \dots)$ for some $f \in \text{MAJ}_k$. Furthermore, for every fixed $y \in \{-1, 1\}^{v \log k}$, IP$(x \oplus y)$ is computable by $f_y(\chi_1, \chi_2, \dots)$ for some $f_y \in \text{MAJ}_k$. This is because for each parity, $U = \{\chi_1, \chi_2, \dots\}$ additionally contains its negation.

It remains to show that $|U| \leqslant n$. Setting $v = \log n - \log\log k - 2$ yields $|U| = 2 \cdot 2^v \log k + 2^{v/2} \log k \leqslant n$. Thus, for $k \leqslant 2^{n^{o(1)}}$ the above construction computes IP on the claimed number of variables:

$$v \log k = (\log n - \log\log k - 2) \log k = \Omega(\log n \cdot \log k). \qquad \square$$

Lemma 3.3 immediately yields the desired lower bound on PTF length.

Theorem 3.4 *Let $k \leqslant 2^{n^{o(1)}}$. Then the intersection of k majorities requires a PTF with $n^{\Omega(\log k)}$ monomials.*

Proof Let $k \leqslant 2^{n^{o(1)}}$. By Lemma 3.3, there is a function $f \in \text{MAJ}_k$ and a choice of signed parities $\chi_1, \dots, \chi_n$ such that $f(\chi_1, \dots, \chi_n)$ computes IP on $v = \Omega(\log n \cdot \log k)$ variables. Since $\mathbf{L}_\infty(f(\chi_1, \dots, \chi_n)) = 2^{-v/2}$, any PTF computing $f(\chi_1, \dots, \chi_n)$ requires $2^{v/2} = n^{\Omega(\log k)}$ monomials by Theorem 2.3. By Lemma 3.1, the same holds for $f(x_1, \dots, x_n)$. □

3.2 PTF length of MAJ$_k$: an $n^{\Omega(k/\max\{\log\log n, \log k\})}$ bound

This section applies Lemma 3.1 with a different reduction. The resulting lower bound is better than that of Theorem 3.4 for some range of k.

Lemma 3.5 *Let $k \leqslant \sqrt{n}$. Then there are explicitly given functions $\chi_1, \chi_2, \dots, \chi_n$ (each a parity or the negation of a parity) such that every reflection of* CQ *on* $\min\{\Omega(\frac{k \log n}{\log\log n}), \Omega(\frac{k \log n}{\log k})\}$ *variables is computable by $f(\chi_1, \chi_2, \dots, \chi_n)$ for some $f \in \text{MAJ}_k$.*

Proof Consider CQ on v variables, for some $v = v(n,k)$ to be chosen later. Since CQ depends only on the sum of the input bits, it can be represented by the AND of v predicates as follows:

$$\text{CQ}(x) = 1 \iff \bigwedge_{s \in S} \left(\textstyle\sum_i x_i \neq s \right),$$

where $S \subseteq \{-v, \ldots, 0, \ldots, v\}$ and $|S| \leqslant v$. A single PTF can check any number t of these predicates:

$$\Big(\sum_i x_i - s_1\Big)^2 \Big(\sum_i x_i - s_2\Big)^2 \cdots \Big(\sum_i x_i - s_t\Big)^2 > 0, \tag{3.2}$$

where $s_1, \ldots, s_t \in S$.

Consider the PTF $(\sum_i x_i + v)^{2t} > 0$. Multiplying out the l.h.s. yields the sum of exactly $(2v)^{2t}$ parities (not all distinct). Construct a new set of variables $U = \{\chi_1, \chi_2, \ldots\}$ to contain a variable for each of these $(2v)^{2t}$ parities and their negations. Over U, the PTF $(\sum_i x_i + v)^{2t} > 0$ is a majority. In fact, any PTF of the form (3.2) is a majority over U. Hence, $\mathrm{CQ}(x)$ on v variables is computable by $f(\chi_1, \chi_2, \ldots)$ for some $f \in \mathrm{MAJ}_k$. Furthermore, for every fixed $y \in \{-1, 1\}^v$, $\mathrm{CQ}(x \oplus y)$ is computable by $f_y(\chi_1, \chi_2, \ldots)$ for some $f_y \in \mathrm{MAJ}_k$. This is because for each parity, $U = \{\chi_1, \chi_2, \ldots\}$ additionally contains its negation.

It remains to pick v such that $v \leqslant kt$ (the k PTFs must collectively check all v predicates) and $|U| \leqslant n$ (the new variable set can have size at most n):

$$\begin{aligned} v &= \max\{v' : v' \leqslant kt \text{ and } 2(2v')^{2t} \leqslant n \text{ for some integer } t \geqslant 1\} \\ &= \min\left\{\Omega(\sqrt{n}), \Omega\left(\frac{k \log n}{\log\log n}\right), \Omega\left(\frac{k \log n}{\log k}\right)\right\}, \end{aligned}$$

which is equivalent to $v = \min\{\Omega(k \log n/\log\log n),\ \Omega(k \log n/\log k)\}$ for $k \leqslant \sqrt{n}$. □

Theorem 3.6 *Let $k \leqslant \sqrt{n}$. Then the intersection of k majorities requires a PTF with* $\min\{n^{\Omega(k/\log\log n)}, n^{\Omega(k/\log k)}\}$ *monomials.*

Proof Let $k \leqslant \sqrt{n}$. By Lemma 3.5, there is a function $f \in \mathrm{MAJ}_k$ and a choice of signed parities $\chi_1, \ldots, \chi_n$ such that $f(\chi_1, \ldots, \chi_n)$ computes CQ on $v = \min\{\Omega(k \log n/\log\log n), \Omega(k \log n/\log k)\}$ variables. Since $\mathbf{L}_\infty(f(\chi_1, \ldots, \chi_n)) = 2^{-v/2}$, any PTF computing $f(\chi_1, \ldots, \chi_n)$ requires $2^{v/2}$ monomials by Theorem 2.3. By Lemma 3.1, the same holds for $f(x_1, \ldots, x_n)$. □

3.3 PTF length of MAJ_2: an $n^{\Omega(\log n/\log\log n)}$ bound

Our lower bound for the PTF length of MAJ_2 exploits two related results in the literature. The first is a lower bound on the degree of any PTF for MAJ_2, due to O'Donnell and Servedio (2003). We additionally amplify the degree requirements by replacing each variable in MAJ_2 by a parity on a separate set of $\approx \log n$ variables. Denote the resulting composition by $\mathrm{MAJ}_2 \circ \mathrm{PARITY}$. The second result we use is a general theorem of Krause and Pudlák (1997) which, given the PTF degree of a function f, states a lower bound on the PTF length of a *related* function f^{op}. We obtain the result of this section by relating the PTF length of MAJ_2 to that of $(\mathrm{MAJ}_2 \circ \mathrm{PARITY})^{\mathrm{op}}$.

The *degree* of a function f, denoted $\deg(f)$, is the minimum degree of any polynomial that strongly represents it. For MAJ_2, we have:

Theorem 3.7 (O'Donnell and Servedio 2003, Theorem 17) *Let* $f(x, y) = \mathrm{MAJ}(x_1, \ldots, x_n) \wedge \mathrm{MAJ}(y_1, \ldots, y_n)$. *Then f has degree* $\Omega(\frac{\log n}{\log\log n})$.

The key to the lower bound in this section is the following link between PTF degree and length requirements.

Definition 3.8 *For $f:\{-1,1\}^n \to \{-1,1\}$, define $f^{\mathrm{op}}:\{-1,1\}^{3n} \to \{-1,1\}$ as*

$$f^{\mathrm{op}}(x_1,\dots,x_n,y_1,\dots,y_n,z_1,\dots,z_n) = f(u_1,\dots,u_n),$$

where $u_i = (\overline{z_i} \wedge x_i) \vee (z_i \wedge y_i)$.

Proposition 3.9 (Krause and Pudlák 1997, Proposition 2.1) *Let $f:\{-1,1\}^n \to \{-1,1\}$ be given. Then f^{op} requires PTF length $2^{\deg(f)}$.*

We need another observation.

Lemma 3.10 *Let $g(x) = f(\bigoplus_{i=1}^k x_{1,i},\dots,\bigoplus_{i=1}^k x_{n,i})$. Then $\deg(g) = k \cdot \deg(f)$.*

Proof Our proof is inspired by the XOR lemma of O'Donnell and Servedio (2003, Theorem 13). The upper bound $k \cdot \deg(f)$ is trivial: take any polynomial of degree $\deg(f)$ that strongly represents f and replace each variable by its corresponding length-k parity on $x_{i,j}$. To prove that $k \cdot \deg(f)$ is also a lower bound on $\deg(g)$, note that f has no strong representation over parities of degree less than $\deg(f)$. By the Theorem of the Alternative, f has a weak representation p_w over parities of degree at least $\deg(f)$. Substituting corresponding parities on $x_{i,j}$ for the variables of p_w yields a weak representation of g; it is nonzero on many assignments to $x_{i,j}$ since p_w is nonzero on at least one assignment to $x_1,\dots,x_n$. The degree of any monomial in the resulting PTF for g is at least $k \cdot \deg(f)$. By the Theorem of the Alternative, g cannot have a strong representation over the parities of degree less than $k \cdot \deg(f)$. We conclude that $\deg(g) \geqslant k \cdot \deg(f)$. □

Combining the above yields the desired bound:

Theorem 1.2 (Restated from Sect. 1.1) *The function* $\mathrm{MAJ}(x_1,\dots,x_n) \wedge \mathrm{MAJ}(y_1,\dots,y_n)$ *requires PTF length $n^{\Omega(\log n/\log\log n)}$.*

Proof Let $f = \mathrm{MAJ}(x_1,\dots,x_t) \wedge \mathrm{MAJ}(x_{t+1},\dots,x_{2t})$. Define a new function $f^{\oplus}:(\{-1,1\}^k)^{2t} \to \{-1,1\}$ as

$$f^{\oplus}(x) = \mathrm{MAJ}\left(\bigoplus_{i=1}^k x_{1,i},\dots,\bigoplus_{i=1}^k x_{t,i}\right) \wedge \mathrm{MAJ}\left(\bigoplus_{i=1}^k x_{t+1,i},\dots,\bigoplus_{i=1}^k x_{2t,i}\right).$$

By Lemma 3.10, $\deg(f^{\oplus}) = k \cdot \deg(f)$. Consider now $f^{\oplus\mathrm{op}}$. For single bits $a,b,c \in \{-1,1\}$, we have $(\overline{c} \wedge a) \vee (c \wedge b) = \frac{1}{2}(1+c)a + \frac{1}{2}(1-c)b$. As a result, $f^{\oplus\mathrm{op}}$ can be computed by the intersection of two PTFs:

$$f^{\oplus\mathrm{op}}(x,y,z) = \left(\prod_{i=1}^k q_{1,i} + \cdots + \prod_{i=1}^k q_{t,i} \geqslant 0\right) \wedge \left(\prod_{i=1}^k q_{t+1,i} + \cdots + \prod_{i=1}^k q_{2t,i} \geqslant 0\right),$$

where $q_{i,j} = (1+z_{i,j})x_{i,j} + (1-z_{i,j})y_{i,j}$.

Therefore, $f^{\oplus\mathrm{op}}$ is computed by the intersection of two PTFs, each with weight at most $4^k t$. Lemma 3.1 implies that if the intersection of two majorities, each on a distinct set of $4^k t$ variables, has a PTF with ℓ monomials, then so does $f^{\oplus\mathrm{op}}$. But by Proposition 3.9, $f^{\oplus\mathrm{op}}$ requires a PTF of length $2^{\deg(f^{\oplus})} = 2^{k\cdot\deg(f)}$. To summarize, the intersection of two majorities, each on $4^k t$ variables, requires a PTF of length $2^{k\cdot\Omega(\log t/\log\log t)}$. The theorem follows by setting $t = \sqrt{n}$ and $k = \frac{1}{4}\log n$. □

Using a rational approximation to the sign function, it is possible to obtain a PTF for $\mathrm{MAJ}(x_1, \dots, x_n) \wedge \mathrm{MAJ}(y_1, \dots, y_n)$ with $n^{O(\log n)}$ monomials (Beigel et al. 1995). Our lower bound of $n^{\Omega(\log n/\log\log n)}$ nearly matches that upper bound.

A key ingredient in our proof of the $n^{\Omega(\log n/\log\log n)}$ lower bound on the PTF length of MAJ_2 was the non-trivial degree lower bound for the same function, due to O'Donnell and Servedio (2003). We could obtain an $n^{\omega(1)}$ lower bound for the PTF length of MAJ_2 by using the simpler $\omega(1)$ lower bound on the degree of MAJ_2 due to Minsky and Papert (1988). That would suffice to show that MAJ_2 does not have a short PTF; the proof would be analogous to that of Theorem 1.2.

Theorems 1.2 and 3.6, established above, immediately imply:

Theorem 1.3 (Restated from Sect. 1.1) *Let $k \leqslant \sqrt{n}$. Then there are (explicitly given) functions in* MAJ_k *that require PTF length $n^{\Omega(k/\log k)}$.*

4 SQ dimension of MAJ$_k$

Recall that the SQ dimension captures the hardness of a concept class. We explicitly construct distributions under which the intersection of n^ϵ majorities, for any constant $0 < \epsilon \leqslant 1/2$, has SQ dimension $2^{\Omega(n^\epsilon)}$. This is an exponential improvement on $n^{\Omega(\log n)}$, the previous best lower bound that was based on computing parity functions by intersections of halfspaces. We additionally prove (Sect. 4.1) that the latter construction could not give a bound better than $n^{\Theta(\log n)}$.

Let $f : \{-1, 1\}^n \to \{-1, 1\}$ be any function. Recall that for a fixed string $y \in \{-1, 1\}^n$, the *y-reflection of f* is the function $f_y(x) = f(x \oplus y)$. A key observation is that any two distinct reflections of a bent function are uncorrelated under the uniform distribution. This result is known in the coding theory literature; for completeness, we give a self-contained proof below.

Lemma 4.1 (cf. Macwilliams and Sloane 1977, p. 427, Problem 12) *Let $f : \{-1, 1\}^n \to \{-1, 1\}$ be a bent function. Then for any distinct $y, y' \in \{-1, 1\}^n$, $\mathbf{E}_{x\sim U}[f(x \oplus y) \cdot f(x \oplus y')] = 0$.*

Proof For a fixed pair y, y' of distinct strings, we have $y \oplus y' \neq 1^n$. Thus,

$$\begin{aligned}\mathbf{E}_{x\sim U}[f(x\oplus y)f(x\oplus y')] &= \mathbf{E}_x\left[\left(\sum_S \hat{f}(S)\chi_S(x)\chi_S(y)\right)\left(\sum_T \hat{f}(T)\chi_T(x)\chi_T(y')\right)\right]\\ &= \sum_S\sum_T \hat{f}(S)\hat{f}(T)\chi_S(y)\chi_T(y')\cdot\mathbf{E}_x[\chi_S(x)\chi_T(x)]\\ &= \sum_S \hat{f}(S)^2\chi_S(y)\chi_S(y') = \frac{1}{2^n}\sum_S \chi_S(y\oplus y') = 0.\end{aligned}$$

The last equality holds because on every $z \in \{-1, 1\}^n \setminus 1^n$, exactly half of the parities evaluate to -1 and the other half, to 1. □

The following is a simple consequence of Lemma 4.1:

Theorem 4.2 *Let $\mathcal{C}$ denote the concept class of bent functions on n variables. Then* $\mathrm{sqdim}_U(\mathcal{C}) = 2^n$.

Proof Fix a bent function f and consider its 2^n reflections, themselves bent functions. By Lemma 4.1, any two of them are orthogonal. □

Consider a function $h : \{-1,1\}^n \to \{-1,1\}^n$. The *$h$-induced distribution on $\{-1,1\}^n$*, denoted by $h \circ U$, is the distribution given by

$$(h \circ U)(z) = \Pr_{x \sim U}[h(x) = z]$$

for any $z \in \{-1,1\}^n$. Put differently, $h \circ U$ is the uniform distribution over the multiset $h(\{-1,1\}^n)$.

Proposition 4.3 *Let $f, g : \{-1,1\}^n \to \{-1,1\}$ and $h : \{-1,1\}^n \to \{-1,1\}^n$ be arbitrary functions. Then* $\mathbf{E}_{x \sim h \circ U}[f(x) \cdot g(x)] = \mathbf{E}_{x \sim U}[f(h(x)) \cdot g(h(x))]$.

Proof By definition of $h \circ U$, picking a random input according to $h \circ U$ is equivalent to picking $x \in \{-1,1\}^n$ uniformly at random and returning $h(x)$. □

We are ready to prove the claimed SQ lower bound for MAJ_k.

Theorem 1.1 (Restated from page 1) *There are* (*explicitly given*) *distributions on* $\{-1,1\}^n$ *under which*

$$\mathrm{sqdim}(\mathrm{MAJ}_k) = \begin{cases} n^{\Omega(k/\log k)} & \text{if } \log n \leqslant k \leqslant \sqrt{n}, \\ \max\{n^{\Omega(k/\log\log n)}, n^{\Omega(\log k)}\} & \text{if } k \leqslant \log n. \end{cases}$$

Proof Let $k \leqslant \log n$. Fix n monomials $\chi_1, \chi_2, \ldots, \chi_n$ as in Lemma 3.3. Let $v = \Omega(\log n \cdot \log k)$. Then there are 2^v functions $\mathcal{F} = \{f_1, f_2, \ldots, f_{2^v}\} \subset \mathrm{MAJ}_k$, where each $f_i(\chi_1(x), \chi_2(x), \ldots, \chi_n(x))$ computes $\mathrm{IP}(x \oplus y)$ on v variables for a distinct $y \in \{-1,1\}^v$.

Define $h : \{-1,1\}^n \to \{-1,1\}^n$ by

$$h(x) = (\chi_1(x), \chi_2(x), \ldots, \chi_n(x)).$$

Then for every two distinct $f_i, f_j \in \mathcal{F}$,

$$\begin{aligned} &\mathbf{E}_{x \sim h \circ U}[f_i(x) \cdot f_j(x)] \\ &\quad = \begin{cases} \mathbf{E}_{x \sim U}[f_i(\chi_1(x), \ldots, \chi_n(x)) \cdot f_j(\chi_1(x), \ldots, \chi_n(x))] & \text{by Proposition 4.3} \\ 0 & \text{by Lemma 4.1.} \end{cases} \end{aligned}$$

In words, every pair of functions in $\mathcal{F}$ are orthogonal under the distribution $h \circ U$. Therefore, $\mathrm{sqdim}_{h \circ U}(\mathrm{MAJ}_k) \geqslant |\mathcal{F}| = 2^v = n^{\Omega(\log k)}$ for $k \leqslant \log n$. Moreover, the distribution $h \circ U$ has an explicit description: pick a random $x \in \{-1,1\}^n$ and return the n-bit string $(\chi_1(x), \ldots, \chi_n(x))$, where $\chi_1, \ldots, \chi_n$ are the explicitly given monomials from Lemma 3.3. Applying an analogous argument to Lemma 3.5 yields the alternate lower bound $\mathrm{sqdim}(\mathrm{MAJ}_k) = \min\{n^{\Omega(k/\log k)}, n^{\Omega(k/\log\log n)}\}$ for $k \leqslant \sqrt{n}$. □

For completeness, we recall an upper bound on the SQ dimension of MAJ_k. It is an immediate consequence of the results of Blum et al. (1997) and Klivans et al. (2004).

Theorem 4.4 *For every distribution μ on $\{-1,1\}^n$, we have* $\operatorname{sqdim}_\mu(\mathrm{MAJ}_k) \leqslant n^{O(k\cdot\log k\cdot\log n)}$.

Proof Klivans et al. (2004, Theorem 29) show that every $f \in \mathrm{MAJ}_k$ has a PTF of degree $d = O(k\cdot\log k\cdot\log n)$. Thus, every $f \in \mathrm{MAJ}_k$ is a halfspace in terms of the parity functions of degree at most d. It follows that the SQ dimension of MAJ_k is at most the SQ dimension of halfspaces in $\sum_{i=0}^{d}\binom{n}{i} \leqslant n^{O(k\cdot\log k\cdot\log n)}$ dimensions. A seminal paper of Blum et al. (1997) proves that the SQ dimension of halfspaces in D dimensions is at most $\operatorname{poly}(D)$, under all distributions. The claim follows. □

4.1 On the SQ dimension under the uniform distribution

The distributions in Theorem 1.1 are non-uniform. Can we prove a comparable lower bound on the SQ dimension of MAJ_k under the uniform distribution? A natural approach would be to compute different parities with functions in MAJ_k. Since the parities are mutually orthogonal under the uniform distribution, this would yield an SQ lower bound. In what follows, we show that this approach yields at best a trivial $n^{\Omega(\log k)}$ SQ lower bound, even for the much larger class of intersections of unate functions. Specifically, we show that intersections of k unate functions cannot compute PARITY on more than $1 + \log k$ bits.

Proposition 4.5 *Let f be a unate function with orientation σ. If $f(x) = -1$ on some x with $\|x \oplus \sigma\| < n$, then $f(y) = -1$ on some y with* $\mathrm{PARITY}(x) \neq \mathrm{PARITY}(y)$.

Proof Suppose $\|x \oplus \sigma\| < n$. Then $x_i = \sigma_i$ for some i. Let $y = x \oplus e_i$. This ensures that $\mathrm{PARITY}(x) \neq \mathrm{PARITY}(y)$, as desired. Furthermore, $f(y) \leqslant f(x)$ by the unate property, i.e., $f(y) = -1$. □

Theorem 4.6 *To compute* PARITY_n *by the* AND *of unate functions*, 2^{n-1} *unate functions are necessary and sufficient.*

Proof Sufficiency is straightforward: PARITY has a trivial CNF with 2^{n-1} clauses, each of which is a unate function. For the lower bound, consider $\bigwedge f_i = \mathrm{PARITY}$, where each f_i is a unate function with orientation σ_i. By Proposition 4.5, f_i can output "false" only on the input x satisfying $\|x \oplus \sigma_i\| = n$: otherwise f_i would output "false" on two inputs of different parity. Thus, 2^{n-1} unate functions are needed to exclude the 2^{n-1} falsifying assignments to PARITY. □

5 Weakly learning intersections of halfspaces

Section 3 showed that the intersection f of even two majorities does not have a polynomial-length PTF. Thus, there is *some* distribution on $\{-1,1\}^n$ with respect to which the correlation of f with every parity is negligible, i.e., inversely superpolynomial ($1/n^{\omega(1)}$). However, this leaves open the possibility of inverse-polynomial correlation (and thus weak learnability) with respect to the *uniform* distribution. In other words, we would like to know if

$$\mathbf{L}_\infty(h_1 \wedge \cdots \wedge h_k) \geqslant \frac{1}{n^{O(1)}}$$

for a slow enough function $k = k(n)$ and all halfspaces $h_1, \ldots, h_k$.

It is easy to construct an intersection of $k = n^{\omega(1)}$ halfspaces that has only negligible Fourier coefficients (e.g., compute a bent function on $\omega(\log n)$ variables). At the other extreme, Klivans et al. (2004, Theorem 20) have shown that the intersection of $k = O(1)$ halfspaces always has a nonnegligible Fourier coefficient. Thus, we restrict our attention to the range $\omega(1) \leqslant k \leqslant n^{O(1)}$.

This section reports our progress on the problem. Section 5.1 studies two generalizations of MAJ_k and proves that the resulting functions have only negligible Fourier coefficients for all $k = \omega(1)$. On the positive side, Sect. 5.2 proves that no combining function of $k \leqslant \sqrt{\log n}$ halfspaces can compute a bent function on $\omega(\log n)$ variables (which would have only negligible Fourier coefficients). Section 5.3 proves a positive result for a specialization of the problem to unate functions and to intersections of read-once functions.

5.1 Negative results for related concept classes

We consider two generalizations of MAJ_k: the XOR of k majorities, and the AND of k unate functions. In both cases, we show that all Fourier coefficients can be negligible whenever $k = \omega(1)$.

Proposition 5.1 *Let $k = k(n)$ be arbitrary with $\omega(1) \leqslant k \leqslant O(\sqrt{n})$. Let $h_1, \dots, h_k$ be majority functions, each on a separate set of n/k variables. Then $\mathbf{L}_\infty(h_1 \oplus \cdots \oplus h_k) \leqslant 1/n^{\omega(1)}$.*

Proof We can assume that $t = n/k$ is an odd integer; otherwise, work with the largest odd integer t less than n/k. If f and g are functions on disjoint variables, then $\mathbf{L}_\infty(f \oplus g) = \mathbf{L}_\infty(f) \cdot \mathbf{L}_\infty(g)$. For t odd, it is well known (O'Donnell 2003) that $\mathbf{L}_\infty(\text{MAJ}_t) = O(1/\sqrt{t})$. The claim follows. □

Thus, the XOR of $\omega(1)$ majorities has negligible Fourier coefficients. We can extend this result to the AND of unate functions:

Theorem 5.2 *There are unate functions $h_1, \dots, h_k$ such that $\mathbf{L}_\infty(\bigwedge h_i) = 1/n^{\omega(1)}$ whenever $k = \omega(1)$.*

Proof Assume $k \leqslant \sqrt{n}$ (otherwise simply set $h_{\sqrt{n}+1} \equiv \cdots \equiv h_k \equiv 1$). Given $k = \omega(1)$, let $t = \log k = \omega(1)$. Let $f = g_1 \oplus \cdots \oplus g_t$, where each g_i is a majority function on a distinct set of n/t variables. By Proposition 5.1, $\mathbf{L}_\infty(f) = 1/n^{\omega(1)}$. At the same time, f is computed by the AND of 2^{t-1} functions $h_1, \dots, h_{2^{t-1}}$, where each h_i is a disjunction on $\{g_1, \neg g_1, \dots, g_t, \neg g_t\}$. Since each g_i is a unate function, so is $\neg g_i$. Then each h_i is a disjunction of unate functions on disjoint variable sets and is thus itself a unate function. In summary, f is computed by the AND of $2^{t-1} \leqslant k$ unate functions. □

5.2 Computing a bent function with halfspaces

Consider a function f of the form $f = g(h_1, h_2, \dots, h_k)$, where each h_i is a halfspace and $g : \{-1, 1\}^k \to \{-1, 1\}$ is an arbitrary combining function. We will give a combinatorial argument that for $k = o(\sqrt{n})$, the function f cannot be bent. Our proof technique is inspired by the analysis of edge slicing in (Saks 1993). An *edge* is a pair of vertices $x, y \in \{-1, 1\}^n$ of the hypercube that differ in exactly one coordinate. It is easy to see that the hypercube contains $2^{n-1}n$ edges. An edge (x, y) is *sliced* by function f if $f(x) \neq f(y)$; otherwise, the edge (x, y) is *unsliced*.

The proof below is based on two observations. First, it is known that a single halfspace slices at most a $\Theta(1/\sqrt{n})$ fraction of the edges. The halfspace $x_1 + x_2 + \cdots + x_n \geqslant 0$ achieves this bound exactly. The second observation is that a bent function slices many edges; in fact, we prove that *every* bent function slices exactly half of the edges. To see why this is intuitively satisfying, note that a random Boolean function is likely to be nearly bent. At the same time, when the vertices of the hypercube are randomly labeled $+1$ or -1, one would expect about half of the edges to be sliced. To summarize, bent functions slice many edges, while a single halfspace slices few. We combine these two facts to prove that no function on $o(\sqrt{n})$ halfspaces can compute a bent function.

Theorem 5.3 (O'Neil 1971) *A halfspace slices at most $\frac{1}{2}n\binom{n}{n/2} = \Theta(2^n\sqrt{n})$ edges of the n-cube.*

Theorem 5.3 proves the first observation. To prove the second, we first relate the number of edges sliced by a Boolean function to its Fourier spectrum.

Lemma 5.4 *Every Boolean function f slices exactly $2^{n-1}\sum_S |S|\hat{f}(S)^2$ edges.*

Proof The probability $p(f)$ that a random edge is sliced by f is

$$p(f) = \mathbf{E}_{i\in[n]}[\mathbf{E}_{x\in\{-1,1\}^n}[\mathbf{I}[f(x) \neq f(x \oplus e_i)]]].$$

Note that $\mathbf{E}_{x\in\{-1,1\}^n}[\mathbf{I}[f(x) \neq f(x \oplus e_i)]] = \mathbf{Inf}_i(f)$, the *influence* of variable x_i on f. As a result,

$$p(f) = \mathbf{E}_{i\in[n]}[\mathbf{Inf}_i(f)] = \frac{1}{n}\sum_{i\in[n]}\mathbf{Inf}_i(f) = \frac{1}{n}\sum_{S\subseteq[n]}|S|\hat{f}(S)^2.$$

The last equation is based on the well-known equality $\sum_{i\in[n]}\mathbf{Inf}_i(f) = \sum_{S\subseteq[n]}|S|\hat{f}(S)^2$ (see Bshouty and Tamon 1996, Lemma 4.1). Since the total number of edges is $2^{n-1}n$, we see that f slices $p(f)\cdot 2^{n-1}n = 2^{n-1}\sum_{S\subseteq[n]}|S|\hat{f}(S)^2$ edges. □

As a special case, PARITY$_n = x_1x_2\ldots x_n$ slices every edge ($2^{n-1}n$), while the constant function $f = 1$ slices no edges. For bent functions, we obtain the following corollary:

Corollary 5.4.1 *Every bent function slices exactly $2^{n-2}n$ edges.*

Proof Every bent function f satisfies $\hat{f}(S)^2 = 1/2^n$ for all $S \subseteq [n]$. By Lemma 5.4, the number of edges sliced by f is:

$$2^{n-1}\sum_{S\subseteq[n]}|S|\hat{f}(S)^2 = \frac{1}{2}\sum_{S\subseteq[n]}|S| = \frac{1}{2}\sum_{k=0}^{n}k\binom{n}{k} = n2^{n-2}.$$

□

Our sought result is a straightforward consequence of Theorem 5.3 and Corollary 5.4.1.

Theorem 5.5 *Let $f = g(h_1, h_2, \ldots, h_k)$, where each $h_i : \{-1, 1\}^n \to \{-1, 1\}$ is a halfspace and $g : \{-1, 1\}^k \to \{-1, 1\}$ is an arbitrary Boolean function. If $k = o(\sqrt{n})$, then f is not bent.*

Proof For f to slice an edge, at least one of $h_1, h_2, \dots, h_k$ must slice it. By Theorem 5.3, a single halfspace can slice at most $\Theta(2^n\sqrt{n})$ edges. Since every bent function slices exactly $2^{n-2}n$ edges (by Corollary 5.4.1), $k = \Omega(\sqrt{n})$ halfspaces are necessary for f to be bent. □

For a general combining function, the $\Omega(\sqrt{n})$ bound of Theorem 5.5 is not far off. For example, the XOR of n halfspaces can compute the bent function IP_n. Also, the majority of $2n$ halfspaces can implement any symmetric function on n bits (Bruck 1990) and, therefore, can implement the bent function CQ_n. It is less clear how tight the $\Omega(\sqrt{n})$ bound is for AND. In particular, the AND of $2^{\Omega(n)} \gg \sqrt{n}$ unate functions (and thus, halfspaces) is needed to compute CQ_n. (The proof is a straightforward generalization of our argument in Theorem 4.6.) In light of the $2^{\Omega(n)}$ complexity of CQ_n, it is plausible that AND is weaker than other combining functions and the lower bound for AND can be improved.

5.3 Read-once intersections and unate functions

Given the intersection $f = h_1 \wedge \dots \wedge h_k$ of functions on disjoint variable sets, we can exploit their independence in analyzing the spectrum of f.

Lemma 5.6 *Let $f = h_1 \wedge h_2 \wedge \dots \wedge h_k$, where the h_i are arbitrary Boolean functions on disjoint variable sets. Then $\mathbf{L}_\infty(f) \geqslant \frac{1}{3}\max_i\{\mathbf{L}_\infty^+(h_i)\}$.*

Proof It suffices to prove that $\mathbf{L}_\infty(f) \geqslant \mathbf{L}_\infty^+(h_k)/3$. Let $p_i = \Pr_x[h_i(x) = 1]$ and $p = \Pr_x[f(x) = 1]$. The independence of the h_i implies that $p = \prod p_i$. We have:

$$f = h_1 \wedge h_2 \wedge \dots \wedge h_k = -1 + \frac{1}{2^{k-1}} \cdot \prod_{i\in[k]}(1 + h_i) = -1 + \frac{1}{2^{k-1}} \sum_{A\subseteq[k]} h_A,$$

where $h_A \overset{\text{def}}{=} \prod_{i\in A} h_i$. Therefore for all $S \neq \emptyset$,

$$\hat{f}(S) = \frac{1}{2^{k-1}} \sum_{A\subseteq[k]} [\widehat{h_A}(S)]. \tag{5.1}$$

Let $S \neq \emptyset$ be a subset of the variables on which h_k is defined. Because $h_1, \dots, h_k$ are on disjoint sets of variables, we see that

$$\widehat{h_A}(S) = \begin{cases} \widehat{h_k}(S)\prod_{i\in A\setminus\{k\}} \widehat{h_i}(\emptyset) & \text{if } k \in A, \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

Substituting (5.2) in (5.1) yields:

$$|\hat{f}(S)| = \left| \frac{\widehat{h_k}(S)}{2^{k-1}} \sum_{A\subseteq[k-1]} \prod_{i\in A} \widehat{h_i}(\emptyset) \right| = \left| \frac{\widehat{h_k}(S)}{2^{k-1}} \prod_{i\in[k-1]} (1 + \widehat{h_i}(\emptyset)) \right|$$
$$= |\widehat{h_k}(S)| \prod_{i\in[k-1]} p_i \geqslant |\widehat{h_k}(S)| \cdot p = |\widehat{h_k}(S)| \cdot \frac{\hat{f}(\emptyset) + 1}{2}.$$

The above derivation uses the identity $p_i = (1 + \widehat{h_i}(\emptyset))/2$. We have shown that

$$|\hat{f}(S)| \geqslant |\widehat{h_k}(S)| \cdot \frac{\hat{f}(\emptyset) + 1}{2}.$$

Consider two cases. If $\hat{f}(\emptyset) \geqslant -1/3$, we obtain $\mathbf{L}_\infty(f) \geqslant |\hat{f}(S)| \geqslant |\widehat{h_k}(S)|/3$. If $\hat{f}(\emptyset) < -1/3$, we have $\mathbf{L}_\infty(f) \geqslant |\hat{f}(\emptyset)| > 1/3 \geqslant |\widehat{h_k}(S)|/3$. In either case,

$$\mathbf{L}_\infty(f) \geqslant \frac{1}{3}|\widehat{h_k}(S)|.$$

Since the choice of $\widehat{h_k}(S)$ was arbitrary from among the nonconstant Fourier coefficients of h_k, we have $\mathbf{L}_\infty(f) \geqslant \mathbf{L}_\infty^+(h_k)/3$. □

Lemma 5.6 states that if at least one of $h_1, \ldots, h_k$ has a large *nonconstant* Fourier coefficient, then $f = h_1 \wedge \cdots \wedge h_k$ will have a large Fourier coefficient as well. Somewhat surprisingly, the claim holds for any k, although the read-once requirement effectively restricts $k \leqslant n$.

We can improve on Lemma 5.6 by considering unate functions in MAJ_k instead of intersections of general read-once functions. We obtain weak learnability in this case by appealing to the benign Fourier properties of unate functions. Analyses of the max-norm of unate functions seem to be folklore, with surveys appearing in (Bshouty and Tamon 1996; Saks 1993). For completeness, we provide a proof below.

Theorem 5.7 *For any unate function $f : \{-1,1\}^n \to \{-1,1\}$,*

$$\mathbf{L}_\infty(f) \geqslant \max\{|\hat{f}(\emptyset)|, |\hat{f}(\{1\})|, \ldots, |\hat{f}(\{n\})|\} \geqslant \frac{1}{n+1}.$$

Proof For a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$, let $f|_{x_i=a}$ denote the subfunction of f with the ith variable set to a. It is easy to see that for all f,

$$\mathbf{E}[f|_{x_i=1} - f|_{x_i=-1}] = 2\hat{f}(\{i\}), \tag{5.3}$$

and

$$\mathbf{E}[(f|_{x_i=1} - f|_{x_i=-1})^2] = 4\sum_{A:i\in A} \hat{f}(A)^2. \tag{5.4}$$

W.l.o.g. assume that f is monotone rather than unate; this does not affect the *absolute* values of f's Fourier coefficients. Then we have $\mathbf{E}[(f|_{x_i=1} - f|_{x_i=-1})^2] = 2\mathbf{E}[f|_{x_i=1} - f|_{x_i=-1}]$. Substituting (5.3) and (5.4) into the latter equality yields:

$$\hat{f}(\{i\}) = \sum_{A:i\in A} \hat{f}(A)^2.$$

Summing over i, we obtain $\sum_i \hat{f}(\{i\}) = \sum_A |A|\hat{f}(A)^2 \geqslant 1 - \hat{f}(\emptyset)^2$, from which we conclude that $\hat{f}(\emptyset)^2 + \sum_i \hat{f}(\{i\}) \geqslant 1$. The claim follows. □

As a corollary, we obtain the following result.

Theorem 5.8 *Let $f = g(h_1, \ldots, h_k)$, where $g : \{-1,1\}^k \to \{-1,1\}$ is a monotone function (e.g.,* AND *or* MAJ*) and the functions $h_i : \{-1,1\}^n \to \{-1,1\}$ are unate with a common orientation (e.g., halfspaces with a common orientation or halfspaces on disjoint sets of variables). Then f is unate and $\mathbf{L}_\infty(f) \geqslant 1/(n+1)$.*

References

Alekhnovich, M., Braverman, M., Feldman, V., Klivans, A., & Pitassi, T. (2004). Learnability and automatizability. In *Proceedings of the 45th annual symposium on foundations of computer science (FOCS).*

Aspnes, J., Beigel, R., Furst, M. L., & Rudich, S. (1994). The expressive power of voting polynomials. *Combinatorica*, *14*(2), 135–148.

Beigel, R., Reingold, N., & Spielman, D. A. (1995). PP is closed under intersection. *Journal of Computer and System Sciences*, *50*(2), 191–202.

Blum, A., Frieze, A., Kannan, R., & Vempala, S. (1997). A polynomial time algorithm for learning noisy linear threshold functions. *Algorithmica*, *22*(1/2), 35–52.

Blum, A., Furst, M., Jackson, J., Kearns, M., Mansour, Y., & Rudich, S. (1994). Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Proceedings of the 26th annual ACM symposium on theory of computing (STOC)* (pp. 253–262). New York: ACM.

Blum, A., Kalai, A., & Wasserman, H. (2003). Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM*, *50*(4), 506–519.

Bruck, J. (1990). Harmonic analysis of polynomial threshold functions. *SIAM Journal on Discrete Mathematics*, *3*(2), 168–177.

Bshouty, N. H., & Tamon, C. (1996). On the Fourier spectrum of monotone functions. *Journal of the ACM*, *43*(4), 747–770.

Feldman, V., Gopalan, P., Khot, S., & Ponnuswami, A. K. (2006). New results for learning noisy parities and halfspaces. In *Proceedings of the 47th annual symposium on foundations of computer science (FOCS)* (pp. 563–574).

Jackson, J. C. (1995). *The harmonic sieve: a novel application of Fourier analysis to machine learning theory and practice*. PhD thesis, Carnegie Mellon University.

Kearns, M. (1993). Efficient noise-tolerant learning from statistical queries. In *Proceedings of the 25th annual ACM symposium on theory of computing (STOC)* (pp. 392–401). New York: ACM.

Klivans, A., O'Donnell, R., & Servedio, R. (2004). Learning intersections and thresholds of halfspaces. *Journal of Computer and System Sciences*, *68*(4), 808–840.

Klivans, A., & Servedio, R. (2004). Learning intersections of halfspaces with a margin. In *Proceedings of the 17th annual conference on learning theory* (pp. 348–362).

Klivans, A. R., & Sherstov, A. A. (2006). Cryptographic hardness for learning intersections of halfspaces. In *Proceedings of the 47th annual symposium on foundations of computer science (FOCS)* (pp. 553–562), October 2006.

Krause, M., & Pudlák, P. (1997). On the computational power of depth-2 circuits with threshold and modulo gates. *Theoretical Computer Science*, *174*(1–2), 137–156.

Kushilevitz, E., & Mansour, Y. (1993). Learning decision trees using the Fourier spectrum. *SIAM Journal on Computing*, *22*(6), 1331–1348.

Kwek, S., & Pitt, L. (1998). PAC learning intersections of halfspaces with membership queries. *Algorithmica*, *22*(1/2), 53–75.

Macwilliams, F. J., & Sloane, N. J. A. (1977). *The theory of error correcting codes*. Amsterdam: North-Holland.

Minsky, M. L., & Papert, S. A. (1988). *Perceptrons: expanded edition*. Cambridge: MIT.

O'Donnell, R. (2003). *Computational applications of noise sensitivity*. PhD thesis, Massachusetts Institute of Technology.

O'Donnell, R., & Servedio, R. A. (2003). New degree bounds for polynomial threshold functions. In *Proceedings of the 25th annual ACM symposium on theory of computing (STOC)* (pp. 325–334). New York: ACM.

O'Neil, P. (1971). Hyperplane cuts of an n-cube. *Discrete Mathematics*, *1*(2), 193–195.

Saks, M. E. (1993). Slicing the hypercube. *Surveys in Combinatorics*, *1993*, 211–255.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, *27*(11), 1134–1142.

Vempala, S. (1997). A random sampling based algorithm for learning the intersection of halfspaces. In *Proceedings of the 38th annual symposium on foundations of computer science (FOCS)* (pp. 508–513).

Yang, K. (2005). New lower bounds for statistical query learning. *Journal of Computer and System Sciences*, *70*(4), 485–509.