Electronic Colloquium on Computational Complexity, Report No. 116 (2007)

Approximate Inclusion-Exclusion for Arbitrary Symmetric Functions*

Alexander A. Sherstov

The Univ. of Texas at Austin, Dept. of Computer Sciences, Austin, TX 78712 USA
sherstov@cs.utexas.edu

July 20, 2007

Abstract. Let $A_1, A_2, \ldots, A_n$ be events in some probability space. The *approximate inclusion-exclusion problem*, due to Linial and Nisan (1990), is to estimate $\mathbf{Pr}[A_1 \cup \cdots \cup A_n]$ given $\mathbf{Pr}[\bigcap_{i \in S} A_i]$ for all $|S| \leqslant k$. Kahn et al. (1996) solve this problem optimally for each k. We study the following more general question: given $\mathbf{Pr}[\bigcap_{i \in S} A_i]$ for all $|S| \leqslant k$, estimate

$$\mathbf{Pr}\Big[\text{the } \textit{number} \text{ of events among } A_1, \ldots, A_n \text{ that hold is in } Z\Big],$$

where $Z \subseteq \{0, 1, \ldots, n\}$ is a given set. (In the Linial-Nisan problem, $Z = \{1, \ldots, n\}$.) We solve this general problem for all Z and k, giving an algorithm that runs in polynomial time and achieves an approximation error that is essentially optimal. We prove this optimal error to be $2^{-\tilde{\Theta}(k^2/n)}$ for k above a certain threshold, and $\Theta(1)$ otherwise.

As part of our solution, we determine, for every predicate $D : \{0, 1, \ldots, n\} \to \{0, 1\}$ and every $\epsilon \in [1/2^n, 1/3]$, the least degree $\deg_\epsilon(D)$ of a polynomial that approximates D pointwise within ϵ. Namely, we show that $\deg_\epsilon(D) = \tilde{\Theta}\left(\deg_{1/3}(D) + \sqrt{n \log(1/\epsilon)}\right)$, where $\deg_{1/3}(D)$ is well-known for each D. Previously, the answer for vanishing ϵ was known only for $D = \mathrm{OR}$ (Kahn et al., 1996). We construct the approximating polynomial explicitly for every D and ϵ.

Our proof departs considerably from Linial and Nisan (1990) and Kahn et al. (1996). Its key ingredient is the *Approximation/Orthogonality Principle*, a certain equivalence of approximation and orthogonality in a Euclidean space, recently proved by the author in the context of quantum lower bounds (Sherstov 2007). Our polynomial constructions feature new uses of the Chebyshev polynomials.

*This work has been previously published as Technical Report #TR-07-34 (July 24, 2007) of the Dept. of Computer Sciences, The University of Texas at Austin.

ISSN 1433-8092

1 Introduction

Let $A_1, A_2, \ldots, A_n$ be events in a probability space. The well-known inclusion-exclusion principle allows one to compute the probability of $A_1 \cup \cdots \cup A_n$ using the probabilities of various intersections of $A_1, A_2, \ldots, A_n$:

$$\mathbf{Pr}[A_1 \cup \cdots \cup A_n] = \sum_i \mathbf{Pr}[A_i] - \sum_{i<j} \mathbf{Pr}[A_i \cap A_j] + \sum_{i<j<k} \mathbf{Pr}[A_i \cap A_j \cap A_k] - \cdots + (-1)^n \mathbf{Pr}[A_1 \cap \cdots \cap A_n].$$

A moment's reflection reveals that knowledge of *every* term in this summation is necessary in general for an exact answer. In this light, it is natural to wonder if one can closely approximate $\mathbf{Pr}[\bigcup A_i]$ using the probabilities of intersections of up to k events, where $k \ll n$. This problem, due to Linial and Nisan [10], is known as *approximate inclusion-exclusion.* Linial and Nisan studied this question and gave near-tight bounds on the least approximation error as a function k. A follow-up article by Kahn, Linial, and Samorodnitsky [5] improved those bounds to optimal.

While $A_1 \cup \cdots \cup A_n$ is an important event, it is certainly not the only one of interest. For example, we might be interested in

$$\mathbf{Pr}\Big[\text{ most of the events } A_1, A_2, \ldots, A_n \text{ hold }\Big],$$

or

$$\mathbf{Pr}\Big[\text{ an odd number of events from among } A_1, A_2, \ldots, A_n \text{ hold }\Big].$$

More generally, we might like to know the likelihood that the *number* of events that hold is in a given subset of $\{0, 1, \ldots, n\}$. Formally, let $D : \{0, 1, \ldots, n\} \to \{0, 1\}$ be an arbitrary predicate. Consider

$$\mathbf{Pr}\Big[\, D(\mathbf{I}[A_1] + \cdots + \mathbf{I}[A_n]) = 1 \,\Big], \tag{1.1}$$

where as usual

$$\mathbf{I}[A_i] \stackrel{\text{def}}{=} \begin{cases} 1 & A_i \text{ holds} \\ 0 & \text{otherwise.} \end{cases}$$

As before, we would like to estimate (1.1) to optimal error given the values of $\mathbf{Pr}[\bigcap_{i \in S} A_i]$ for all S with $|S| \leqslant k$. This new problem is a natural generalization of approximate inclusion-exclusion. Yet the methods of Linial and Nisan [10] and Kahn et al. [5] do not cover this broader question.

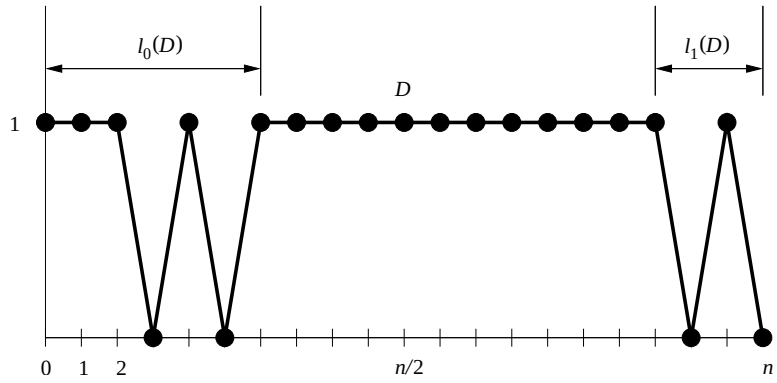
We solve this problem completely for every D and k. More precisely, we give an algorithm that, for every D and k, runs in polynomial time and achieves an approximation error that is essentially optimal. Before we state our results, we introduce some helpful notation.

1.1 Notation

Let $D : \{0, 1, \ldots, n\} \to \{0, 1\}$ be an arbitrary predicate. Define

$$\ell_0(D) \in \{0, 1, \ldots, \lfloor n/2 \rfloor\},$$
$$\ell_1(D) \in \{0, 1, \ldots, \lceil n/2 \rceil\}$$

to be the smallest integers such that D is constant in the range $[\ell_0(D), n - \ell_1(D)]$. The figure below illustrates this definition for a typical predicate D:



The key point is that $\ell_0(D) + \ell_1(D)$ is large if and only if D changes value near the middle of the range. We need another definition.

Definition 1.1. Let $D : \{0, 1, \ldots, n\} \to \{0, 1\}$ and $0 \leqslant k \leqslant n$. Define

$$\delta^*(D, k)$$
$$\stackrel{\text{def}}{=} \sup \left| \mathop{\mathbf{Pr}}_{\mathscr{P}_1}\left[D(\mathbf{I}[A_1] + \cdots + \mathbf{I}[A_n]) = 1\right] - \mathop{\mathbf{Pr}}_{\mathscr{P}_2}\left[D(\mathbf{I}[B_1] + \cdots + \mathbf{I}[B_n]) = 1\right] \right|,$$

where the supremum is taken over all probability spaces $\mathscr{P}_1$ and $\mathscr{P}_2$, over all events $A_1, \ldots, A_n$ in $\mathscr{P}_1$, and over all events $B_1, \ldots, B_n$ in $\mathscr{P}_2$, such that

$$\mathop{\mathbf{Pr}}_{\mathscr{P}_1}\left[\bigcap_{i \in S} A_i\right] = \mathop{\mathbf{Pr}}_{\mathscr{P}_2}\left[\bigcap_{i \in S} B_i\right] \qquad \text{for } |S| \leqslant k.$$

In words, the quantity $\delta^*(D, k)$ in the above definition is the least error achievable in approximating $\mathbf{Pr}[D(\mathbf{I}[A_1] + \cdots + \mathbf{I}[A_n]) = 1]$ in principle, information-theoretically, if unlimited computing power is available.

1.2 Main Result

The first question we settle is precisely how large k needs to be for a good approximation to even *exist*. We prove:

Theorem 1.2 (Existence of a good approximation). *Let $D : \{0,1,\ldots,n\} \to \{0,1\}$. Put $\ell = \ell_0(D) + \ell_1(D)$. Then*

$$\delta^*(D,k) = \begin{cases} \Theta(1) & \text{if} \quad k \leqslant \Theta\left(\sqrt{n\ell}\right), \\ 2^{-\tilde{\Theta}(k^2/n)} & \text{if} \quad \tilde{\Theta}\left(\sqrt{n\ell}\right) \leqslant k \leqslant \Theta(n). \end{cases}$$

Theorem 1.2 tells us that a good approximation exists if and only if $k \geqslant \tilde{\Theta}\left(\sqrt{n\ell}\right)$, where $\ell = \ell_0(D) + \ell_1(D)$. We now give an efficient way to actually *construct* the near-optimal approximation for any given D and k.

Theorem 1.3 (Efficient approximation scheme). *Let $D : \{0,1,\ldots,n\} \to \{0,1\}$. Put $\ell = \ell_0(D) + \ell_1(D)$. Then for every $k \geqslant \tilde{\Theta}\left(\sqrt{n\ell}\right)$ there are reals*

$$a_0, a_1, \ldots, a_k,$$

computable in time $\mathsf{poly}(n)$, *such that*

$$\left| \mathbf{Pr}\left[D(\mathbf{I}[A_1] + \cdots + \mathbf{I}[A_n]) = 1\right] \quad - \quad \sum_{j=0}^{k} a_j \sum_{S:|S|=j} \mathbf{Pr}\left[\bigcap_{i \in S} A_i\right] \right| \leqslant 2^{-\tilde{\Theta}(k^2/n)}$$

for any events $A_1, \ldots, A_n$ in any probability space.

Theorem 1.3 gives the desired approximation algorithm. As we see, it is not even necessary to know the individual probabilities $\mathbf{Pr}[\bigcap_{i \in S} A_i]$; it suffices to know the $k+1$ sums

$$\sum_{S:|S|=j} \mathbf{Pr}\left[\bigcap_{i \in S} A_i\right] \qquad (j = 0, 1, \ldots, k).$$

This solves the generalized inclusion/exclusion problem for all predicates. In actuality, our proof works for *arbitrary* Boolean functions, not just predicates. Specifically, fix $f : \{0,1\}^n \to \{0,1\}$ and suppose we wish to approximate

$$\mathbf{Pr}\left[f(\mathbf{I}[A_1], \ldots, \mathbf{I}[A_n]) = 1\right]$$

given $\mathbf{Pr}[\bigcap_{i \in S} A_i]$ for all S with $|S| \leqslant k$. Let $\delta^*(f,k)$ be the best error achievable information-theoretically. In the case of symmetric functions, i.e., $f(x) \equiv D(x_1 + \cdots + x_n)$ for some predicate D, this is precisely the setting of Theorems 1.2 and 1.3. For arbitrary f, we obtain the following result:

Theorem 1.4. *Let $f : \{0,1\}^n \to \{0,1\}$ be arbitrary and $0 \leqslant k \leqslant n$. Then*

$$\delta^*(f,k) = 2 \min_{\phi} \|f - \phi\|_\infty,$$

where the minimum is over multilinear polynomials $\phi(x_1, \ldots, x_n)$ of degree up to k.

Thus, Theorem 1.4 solves the approximate inclusion/exclusion problem for any f whose approximability by polynomials is well understood.

1.3 Other Results

Approximate degree. As part of our proof, we have to show the following result of independent interest. For a predicate $D : \{0,1,\ldots,n\} \to \{0,1\}$, define its *$\epsilon$-approximate degree* $\deg_\epsilon(D)$ to be the smallest degree of a univariate real polynomial $p(t)$ such that

$$\max_{t=0,1,\ldots,n} |D(t) - p(t)| \leqslant \epsilon.$$

This quantity is of inherent significance and has found various applications in theoretical computer science [5, 6, 8–10, 13, 15, etc.], ranging from approximation algorithms and computational learning to complexity theory. Moreover, the main result of this paper depends critically on tight estimates of $\deg_\epsilon(D)$ for all D and ϵ. We prove:

Theorem 1.5 (Approximate degree of predicates). *Let $D : \{0,1,\ldots,n\} \to \{0,1\}$ be a nonconstant predicate. Let $\epsilon \in [1/2^n, 1/3]$. Then*

$$\deg_\epsilon(D) = \tilde{\Theta}\left(\sqrt{n(\ell_0(D) + \ell_1(D))} + \sqrt{n \log(1/\epsilon)}\right),$$

where the $\tilde{\Theta}$ notation suppresses $\log n$ factors. Furthermore, the approximating polynomial for each D and ϵ is given explicitly.

Theorem 1.5 is a broad generalization of two earlier results in the literature. The first of these is due to Paturi [12], who showed that

$$\deg_{1/3}(D) = \Theta\left(\sqrt{n(\ell_0(D) + \ell_1(D))}\right) \qquad \text{for all } D.$$

Unfortunately, Paturi's result and its proof give no insight into the behavior of the ϵ-approximate degree for vanishing ϵ. The other relevant result is due to Kahn et al. [5], who conducted an in-depth study of the predicate $D = \mathrm{OR}$, defined as usual by $\mathrm{OR}(i) = 1 \Leftrightarrow i \geqslant 1$. Kahn et al. showed that

$$\deg_\epsilon(\mathrm{OR}) = \tilde{\Theta}\left(\sqrt{n \log(1/\epsilon)}\right) \qquad (1/2^n \leqslant \epsilon \leqslant 1/3),$$

where the $\tilde{\Theta}$ notation hides $\log n$ factors. Thus, our work generalizes the results of Paturi and Kahn et al. to *every* predicate and *every* error rate $\epsilon \in [1/2^n, 1/3]$.

Theorem 1.5 has another, more revealing and esthetically pleasing interpretation. In view of Paturi's work, it can be restated as:

$$\deg_\epsilon(D) = \tilde{\Theta}\left(\deg_{1/3}(D) + \sqrt{n \log(1/\epsilon)}\right) \qquad (1/2^n \leqslant \epsilon \leqslant 1/3),$$

where D is nonconstant. In words, past a certain threshold, the dependence of the ϵ-approximate degree on ϵ is the same for all nonconstant predicates. This threshold varies from one predicate to another and equals the degree required for a $\frac{1}{3}$-approximation.

Agnostic learning. The proof technique of our main result additionally gives new lower bounds for *agnostic learning.* The agnostic model, due to Kearns et al. [7], is perhaps the most realistic abstraction of learning. Designing efficient algorithms in this model, even for the simplest concept classes, is notoriously difficult. Nevertheless, progress on proving lower bounds has also been scarce. Some recent lower bounds are [9, 17].

A summary of this model is as follows. Let $\mathscr{C}$ be a concept class, i.e., some set of Boolean functions $\{0,1\}^n \to \{0,1\}$. There is an unknown distribution λ on $\{0,1\}^n \times \{0,1\}$, and the learner receives training examples

$$\left(x^{(1)}, y^{(1)}\right), \quad \left(x^{(2)}, y^{(2)}\right), \quad \ldots, \quad \left(x^{(m)}, y^{(m)}\right),$$

independent and identically distributed according to λ. Let

$$\mathsf{opt} \stackrel{\text{def}}{=} \max_{f \in \mathscr{C}} \left\{ \mathop{\mathbf{Pr}}_{(x,y)\sim\lambda} [f(x) = y] \right\}$$

be the error of the function $f^* \in \mathscr{C}$ that best agrees with the training data. The learner needs to produce a hypothesis $h : \{0,1\}^n \to \{0,1\}$ that agrees with the training data almost as well as f^*:

$$\Pr_{(x,y)\sim\lambda}[h(x) = y] \geqslant \mathsf{opt} - \epsilon,$$

where ϵ is an error parameter fixed in advance. As usual, the goal is to find h efficiently.

A natural approach to learning in this and other models is to consider only those hypotheses that depend on few variables. One tests each such hypothesis against the training data and outputs the one with the least error. This technique is attractive in that the resulting hypothesis space is small and well-structured, making it possible to efficiently identify the best approximation to the observed examples.

The central question then becomes, what advantage over random guessing can such hypotheses guarantee? We prove that, when learning symmetric functions, one is forced to use hypotheses that depend on many variables: all others will generally work no better than random guessing.

Theorem 1.6 (Lower bound for agnostic learning). *Let $D : \{0, 1, \ldots, n\} \to \{0, 1\}$ be a predicate and $f(x) \stackrel{\text{def}}{=} D(x_1 + \cdots + x_n)$. Let $\epsilon > 0$ be an arbitrary constant. Then there is a distribution λ on $\{0, 1\}^n \times \{0, 1\}$ such that*

$$\mathop{\mathbf{Pr}}_{(x,y)\sim\lambda}[f(x) = y] \geqslant 1 - \epsilon$$

and

$$\mathop{\mathbf{Pr}}_{(x,y)\sim\lambda}[g(x) = y] = \frac{1}{2}$$

for every $g : \{0, 1\}^n \to \{0, 1\}$ that depends on at most $c\sqrt{n(\ell_0(D) + \ell_1(D))}$ variables, where $c = c(\epsilon)$ is a constant.

We also show that the bound on the number of variables in Theorem 1.6 is optimal to within a multiplicative constant (see Theorem 5.4). Prior to our work, Tarui and Tsukiji [17] obtained the special case of Theorem 1.6 for $f = \mathrm{OR}$. No other lower bounds for symmetric functions were previously known.

To place Theorem 1.6 in the framework of agnostic learning, consider any concept class $\mathscr{C}$ that contains many symmetric functions. For example, we could fix a symmetric function $f : \{0, 1\}^n \to \{0, 1\}$ and consider the concept class $\mathscr{C}$ of $\binom{2n}{n}$ functions, each being a copy of f applied to a separate set of n variables from among $x_1, x_2, \ldots, x_{2n}$:

$$\mathscr{C} = \left\{ f(x_{i_1}, x_{i_2}, \ldots, x_{i_n}) : \quad 1 \leqslant i_1 < i_2 < \cdots < i_n \leqslant 2n \right\}.$$

Theorem 1.6 now supplies scenarios when *some* member of $\mathscr{C}$ matches the training data almost perfectly (to within any $\epsilon > 0$), and yet every hypothesis that depends on few variables is completely useless (i.e., as good as random guessing).

1.4 Our Techniques

The proof of our main result takes inspiration from the elegant papers of Linial and Nisan [10] and Kahn et al. [5], who have studied the special case $D = \mathrm{OR}$. Namely, we adopt the high-level strategy of these works, which is to reduce the original problem via linear-programming duality to a question in approximation theory. Implementing this strategy, however, requires new and stronger techniques. As we will shortly explain, our proof is a substantial departure from [5, 10].

First of all, the linear-programming reduction in [5, 10] does not extend from $D = \mathrm{OR}$ to arbitrary predicates. To overcome this difficulty, we start with a different and more versatile tool, the *Approximation/Orthogonality Principle*. This principle gives a certain equivalence between approximation and orthogonality in a Euclidean space and has been recently proved by the author [15] in the context of quantum lower bounds. With some work, this yields the desired reduction from the original problem to a question in approximation theory. In addition, the proof turns out simpler and more modular than in [5, 10].

To complete the solution, we must still answer the resulting question in approximation theory. This amounts to determining, for each predicate D and each $\epsilon \in [1/2^n, 1/3]$, the least degree of a polynomial that approximates D pointwise within ϵ, and then constructing such a polynomial explicitly. Previously, such a construction was known only for $D = \mathrm{OR}$ (Kahn et al. [5]). We solve the general case by combining interpolation techniques with a new use of the Chebyshev polynomials.

It may seem that Theorem 1.5, the backbone of this paper, should have a more intuitive and more elementary proof. However, the simpler ideas that come to mind turn out to be useless, as we now discuss.

- An obvious approach is to start with Paturi's $\frac{1}{3}$-approximating polynomial $p(t)$ for the given predicate $D(t)$ and boost its accuracy by composing it with another polynomial, $q(t)$. Let $\epsilon \in (0, 1/3)$ be the desired accuracy. For this approach to work, $q(t)$ must satisfy:

$$q\left(\left[-\tfrac{1}{3}, \tfrac{1}{3}\right]\right) \subseteq [-\epsilon, \epsilon], \qquad q\left(\left[\tfrac{2}{3}, \tfrac{4}{3}\right]\right) \subseteq [1-\epsilon, 1+\epsilon].$$

  Up to translation/scaling, this is equivalent to requiring that $q(t)$ approximate the sign function within ϵ on the interval $[-1, -1+\alpha] \cup [1, 1-\alpha]$ for some constant $\alpha \in (0, 1)$. Eremenko and Yuditskii [4] show that the least degree of such a polynomial $q(t)$ is $\Theta(\log(1/\epsilon))$. Taking $p(t)$ to be Paturi's approximating polynomial for the given predicate D, we see that the composition $p(q(t))$ has degree

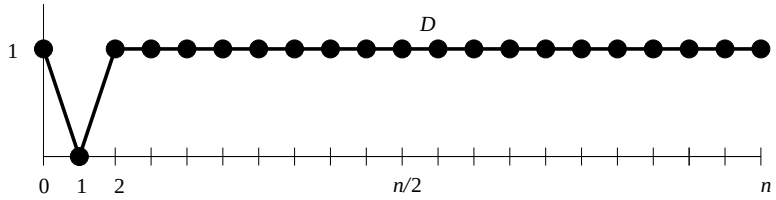
$$\Theta\left(\sqrt{n(\ell_0(D) + \ell_1(D))}\,\log(1/\epsilon)\right).$$

  This is much worse than the optimal bound that we achieve, namely,

$$\tilde{\Theta}\left(\sqrt{n(\ell_0(D) + \ell_1(D))} + \sqrt{n\log(1/\epsilon)}\right).$$

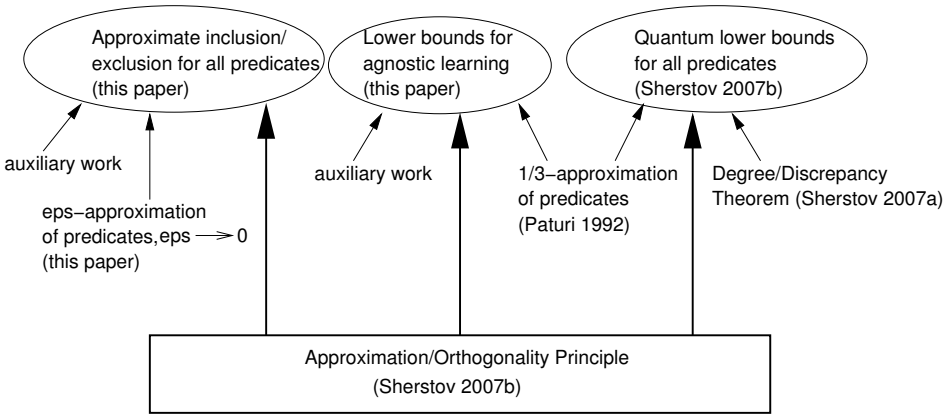
- Another tempting strategy is to view a given predicate $D : \{0, 1, \ldots, n\} \to \{0, 1\}$ as a continuous (piecewise-linear) function on $[0, n]$ and then apply

D. Jackson's fundamental theorems on uniform approximation. Unfortunately, the continuous approximation problem is very hard even for the following simple predicate:



Indeed, an ϵ-approximating polynomial for this continuous function yields (after translation and scaling) an ϵ-approximating polynomial of the same degree for $|x|$ on $[-1,1]$. In his classical work, S.N. Bernstein [2] proves that the latter polynomial requires degree $\Omega(1/\epsilon)$. In particular, this approach is entirely useless once $\epsilon \leqslant \Theta(1/n)$. Yet the predicate in question has an approximator of degree $\Theta(\sqrt{n})$, as we show. Clearly, the key is to exploit the discrete nature of the problem: we are merely seeking an approximation over the finite set of points $\{0,1,\ldots,n\}$, rather than the entire interval $[0,n]$.

We conclude with a broader view of this work. What might be common to approximate inclusion-exclusion, agnostic learning, and quantum communication? These subjects seem quite different at first. One contribution of our paper is to show that, as far as symmetric functions are concerned, these three problems are fundamentally the same mathematical question! Namely, the question is how well a given predicate can be approximated by a univariate polynomial of low degree. We illustrate this equivalence in the following diagram, which shows the skeleton of our proofs:



The three ovals across the top correspond to the main results (two from this paper, one from [15]). The arrows show dependencies in the proofs. What brings the three subjects together is the Approximation/Orthogonality Principle, reviewed in detail in Section 2.2.

1.5 Organization

We start with a thorough review of technical preliminaries in Section 2. We next study the approximation of Boolean predicates by real polynomials in Section 3. Armed with this approximation result, we prove our main theorem in Section 4. Finally, Section 5 reinterprets our technique to give lower bounds for agnostic learning.

2 Preliminaries

This section provides relevant technical background. After some remarks on notation (Section 2.1), we discuss the Approximation/Orthogonality Principle and give its proof for the reader's convenience (Section 2.2). Section 2.3 concludes with some fundamental results about the approximation of Boolean functions by polynomials.

2.1 General

A *Boolean function* is a mapping $\{0,1\}^n \to \{0,1\}$. A *predicate* is a mapping $\{0,1,\ldots,n\} \to \{0,1\}$. The notation $[n]$ stands for the set $\{1,2,\ldots,n\}$. The symbol P_k stands for the set of all univariate real polynomials of degree up to k. For a finite set X and a function $\phi : X \to \mathbb{R}$, we define

$$\|\phi\|_\infty \stackrel{\text{def}}{=} \max_{x \in X} |\phi(x)|.$$

We now recall the Fourier transform on $\{0,1\}^n$. Consider the vector space of functions $\{0,1\}^n \to \mathbb{R}$, equipped with the inner product

$$\langle f, g \rangle \stackrel{\text{def}}{=} \frac{1}{2^n} \sum_{x \in \{0,1\}^n} f(x)g(x).$$

For $S \subseteq [n]$, define $\chi_S : \{0,1\}^n \to \{-1,+1\}$ by

$$\chi_S(x) = (-1)^{\sum_{i \in S} x_i}.$$

Then $\{\chi_S\}_{S\subseteq[n]}$ is an orthonormal basis for the inner product space in question. As a result, every function $f : \{0,1\}^n \to \mathbb{R}$ has a unique *Fourier representation*

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

where $\hat{f}(S) \stackrel{\text{def}}{=} \langle f, \chi_S \rangle$. The reals $\hat{f}(S)$ are called the *Fourier coefficients of* f.

2.2 The Approximation/Orthogonality Principle

Crucial to our work is the *Approximation/Orthogonality Principle*, recently proved by the author [15] in the context of quantum lower bounds. This principle establishes a certain equivalence between approximation and orthogonality in a Euclidean space.

We start with some notation from [15], which will be useful throughout this paper. Let X be a finite set. Consider $\mathbb{R}^X$, the linear space of all functions $X \to \mathbb{R}$. Recall the notation

$$\|\phi\|_\infty \stackrel{\text{def}}{=} \max_{x\in X} |\phi(x)|.$$

Then $(\mathbb{R}^X, \|\cdot\|_\infty)$ is a real normed linear space.

Definition 2.1 (Best error). For $f : X \to \mathbb{R}$ and $\Phi \subseteq \mathbb{R}^X$, let

$$\epsilon^*(f,\Phi) \stackrel{\text{def}}{=} \min_{\phi\in\text{span}(\Phi)} \|f-\phi\|_\infty.$$

In words, $\epsilon^*(f,\Phi)$ is the best error in an approximation of f by a linear combination of functions in Φ. Since span(Φ) has finite dimension, a best approximation to f out of span(Φ) always exists [14, Thm. I.1], justifying our use of "min" instead of "inf" in the above definition.

We now introduce a closely related quantity, $\gamma^*(f,\Phi)$, that measures how well f correlates with a real function that is orthogonal to all of Φ.

Definition 2.2 (Modulus of orthogonality, Sherstov [15]). Let X be a finite set, $f : X \to \mathbb{R}$, and $\Phi \subseteq \mathbb{R}^X$. The *modulus of orthogonality* of f with respect to Φ is:

$$\gamma^*(f,\Phi) \stackrel{\text{def}}{=} \max_{\psi} \left\{ \sum_{x\in X} f(x)\psi(x) \right\}, \tag{2.1}$$

where the maximum is taken over all $\psi : X \to \mathbb{R}$ such that $\sum_{x\in X} |\psi(x)| \leqslant 1$ and $\sum_{x\in X} \phi(x)\psi(x) = 0$ for all $\phi \in \Phi$.

The maximization in (2.1) is over a nonempty set that contains $\psi = 0$. Also, the use of "max" instead of "sup" is legitimate because (2.1) maximizes a continuous function over a compact set. To summarize, the modulus of orthogonality is a well-defined nonnegative real number for every function $f : X \to \mathbb{R}$.

Theorem 2.3 (Approximation/Orthogonality Principle, Sherstov [15]). *Let X be a finite set, $\Phi \subseteq \mathbb{R}^X$, and $f : X \to \mathbb{R}$. Then*

$$\epsilon^*(f, \Phi) = \gamma^*(f, \Phi).$$

Proof. Let $\phi_1, \ldots, \phi_k : X \to \mathbb{R}$ be a basis for span(Φ). Our first observation is that $\epsilon^*(f, \Phi)$ is the optimum of the following linear program in the variables $\epsilon, \alpha_1, \ldots, \alpha_k$:

minimize:	ϵ		
subject to:			
	$\left\lvert f(x) - \sum_{i=1}^{k} \alpha_i \phi_i(x) \right\rvert \leqslant \epsilon$		for each $x \in X$,
	$\alpha_i \in \mathbb{R}$		for each i,
	$\epsilon \geqslant 0.$		

Standard manipulations reveal the dual:

maximize:	$\sum_{x \in X} \beta_x f(x)$		
subject to:			
	$\sum_{x \in X} \lvert\beta_x\rvert \leqslant 1,$		
	$\sum_{x \in X} \beta_x \phi_i(x) = 0$		for each i,
	$\beta_x \in \mathbb{R}$		for each $x \in X$.

Both programs are clearly feasible and thus have the same finite optimum. We have already observed that the optimum of first program is $\epsilon^*(f, \Phi)$. Since $\phi_1, \phi_2, \ldots, \phi_k$ form a basis for span(Φ), the optimum of the second program is by definition $\gamma^*(f, \Phi)$. □

2.3 Approximation by Polynomials

Let $f : \{0,1\}^n \to \mathbb{R}$. As we saw in Section 2.1, any such function f has an *exact* representation as a linear combination of χ_S, where $S \subseteq [n]$. A fundamental question to ask is how closely f can be *approximated* by a linear combination of functions χ_S with $|S|$ small.

Definition 2.4 (Approximate degree of functions). Let $f : \{0,1\}^n \to \mathbb{R}$ and $\epsilon \geqslant 0$. The *ϵ-approximate degree* $\deg_\epsilon(f)$ *of* f is the minimum integer k, $0 \leqslant k \leqslant n$, for which there exists $\phi \in \operatorname{span}(\{\chi_S\}_{|S| \leqslant k})$ with

$$\max_{x \in \{0,1\}^n} |f(x) - \phi(x)| \leqslant \epsilon.$$

We will be primarily interested in the approximate degree of Boolean functions. As a first observation, $\deg_\epsilon(f) = \deg_\epsilon(\neg f)$ for all such functions and all $\epsilon \geqslant 0$. Second, $\deg_\epsilon(f)$ is not substantially affected by the choice of a constant $\epsilon \in (0, 1/2)$. More precisely, we have:

Proposition 2.5 (Folklore). *Let $f : \{0,1\}^n \to \{0,1\}$ be arbitrary, ϵ a constant with $0 < \epsilon < 1/2$. Then*

$$\deg_\epsilon(f) = \Theta(\deg_{1/3}(f)).$$

Proof (folklore). Assume that $\epsilon \leqslant 1/3$; the case $\epsilon \in (1/3, 1/2)$ has a closely analogous proof, and we omit it. Let $k \stackrel{\text{def}}{=} \deg_{1/3}(f)$. We have to show that $\deg_\epsilon(f) = O(k)$. For this, fix $\phi \in \operatorname{span}(\{\chi_S\}_{|S| \leqslant k})$ with $\max_{x \in \{0,1\}^n} |f(x) - \phi(x)| \leqslant 1/3$. By basic approximation theory (see Rivlin [14, Cor. 1.4.1]), there exists a univariate polynomial p of degree $O(1/\epsilon)$ with

$$p\left(\left[-\tfrac{1}{3}, \tfrac{1}{3}\right]\right) \subseteq [-\epsilon, \epsilon], \qquad p\left(\left[\tfrac{2}{3}, \tfrac{4}{3}\right]\right) \subseteq [1-\epsilon, 1+\epsilon].$$

Then clearly $p(\phi(x))$ is the sought approximator of f. □

In view of Proposition 2.5, the convention is to work with $\deg_{1/3}(f)$ by default. Determining this quantity for a given Boolean function f can be difficult. There is, however, a family of Boolean functions whose approximate degree is analytically manageable. This is the family of *symmetric* Boolean functions, i.e., functions $f : \{0,1\}^n \to \{0,1\}$ whose value is uniquely determined by $x_1 + \cdots + x_n$. Equivalently, a Boolean function f is symmetric if and only if

$$f(x_1, x_2, \ldots, x_n) = f(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)})$$

for all inputs $x \in \{0,1\}^n$ and all permutations $\sigma : [n] \to [n]$. Note that there is a one-to-one correspondence between predicates and symmetric Boolean functions. Namely, one associates a predicate D with the symmetric function

$$f(x) \stackrel{\text{def}}{=} D(x_1 + \cdots + x_n).$$

To carry our discussion further, we extend the notion of approximation to predicates.

Definition 2.6 (Approximate degree of predicates). For a predicate $D : \{0,1,\ldots,n\} \to \{0,1\}$, define its *$\epsilon$-approximate degree* $\deg_\epsilon(D)$ to be the minimum degree of a univariate real polynomial p with

$$\max_{i=0,1,\ldots,n} |D(i) - p(i)| \leqslant \epsilon.$$

Analyzing the approximate degree of predicates is a much simpler task and, indeed, a basic question in approximation theory. It is therefore fortunate that the ϵ-approximate degree of a symmetric function is the same as the ϵ-approximate degree of its associated predicate. This equivalence is known as the *symmetrization argument* of Minsky and Papert [11]. Before we can state this theorem, we introduce some important notation.

Definition 2.7. For $f : \{0,1\}^n \to \{0,1\}$ and $D : \{0,1,\ldots,n\} \to \{0,1\}$, define

$$\epsilon^*(f, \{\chi_S\}_{|S| \leqslant k}) \stackrel{\text{def}}{=} \min_{\phi \in \operatorname{span}(\{\chi_S\}_{|S| \leqslant k})} \max_{x \in \{0,1\}^n} |f(x) - \phi(x)|,$$

$$\epsilon^*(D, P_k) \stackrel{\text{def}}{=} \min_{p \in P_k} \max_{i=0,1,\ldots,n} |D(i) - p(i)|.$$

Definition 2.7 merely instantiates the symbol $\epsilon^*(\phi, \Phi)$ from Section 2.2 to the special cases $\phi = f$ and $\phi = D$. We have:

Proposition 2.8 (Symmetrization argument, Minsky and Papert [11]). *Let $f : \{0,1\}^n \to \{0,1\}$ be a symmetric Boolean function. Let D be the predicate with $f(x) \equiv D(x_1 + \cdots + x_n)$. Then*

$$\epsilon^*(f, \{\chi_S\}_{|S| \leqslant k}) = \epsilon^*(D, P_k) \qquad \text{for all } k = 0,1,\ldots,n. \tag{2.2}$$

In particular,

$$\deg_\epsilon(f) = \deg_\epsilon(D) \qquad \text{for all } \epsilon \geqslant 0. \tag{2.3}$$

Proof sketch (Minsky and Papert [11]). It is clear that (2.2) implies (2.3), so we focus on the former. Since $f(x) = D(x_1 + \cdots + x_n)$, we immediately have

$$\epsilon^*(f, \{\chi_S\}_{|S|\leqslant k}) \leqslant \epsilon^*(D, P_k),$$

and it remains to prove the reverse inequality. Fix $\phi \in \text{span}(\{\chi_S\}_{|S|\leqslant k})$ for which $\|\phi - f\|_\infty = \epsilon^*(f, \{\chi_S\}_{|S|\leqslant k})$. Define $\phi' : \{0,1\}^n \to \mathbb{R}$ by

$$\phi'(x) \stackrel{\text{def}}{=} \frac{1}{n!} \sum_{\sigma \in S_n} \phi(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)}).$$

On the one hand,

$$\|f - \phi'\|_\infty \leqslant \|f - \phi\|_\infty = \epsilon^*(f, \{\chi_S\}_{|S|\leqslant k}). \tag{2.4}$$

On the other hand, one can use the uniqueness of the Fourier representation to show that

$$\phi'(x) = p(x_1 + \cdots + x_n)$$

for some $p \in P_k$. But then

$$\|f - \phi'\|_\infty = \|D - p\|_\infty \geqslant \epsilon^*(D, P_k). \tag{2.5}$$

The sought conclusion follows from (2.4) and (2.5). □

Using Proposition 2.8 and tools from approximation theory, Paturi [12] gave an asymptotically tight estimate of $\deg_{1/3}(f)$ for every symmetric Boolean function f. The estimates are in terms of the quantities $\ell_0(f)$ and $\ell_1(f)$, defined next.

Definition 2.9 (Razborov [13]). Let $D : \{0, 1, \ldots, n\} \to \{0, 1\}$. Define

$$\ell_0(D) \in \{0, 1, \ldots, \lfloor n/2 \rfloor\},$$
$$\ell_1(D) \in \{0, 1, \ldots, \lceil n/2 \rceil\}$$

to be the smallest integers such that D is constant in the range $[\ell_0(D), n - \ell_1(D)]$. For a symmetric function $f : \{0,1\}^n \to \{0,1\}$, define $\ell_0(f) = \ell_0(D)$ and $\ell_1(f) = \ell_1(D)$, where D is the predicate for which $f(x) \equiv D(x_1 + \cdots + x_n)$.

See Section 1 for a pictorial illustration of this definition. We are ready to state Paturi's fundamental theorem.

Theorem 2.10 (Paturi [12]). *Let $f : \{0,1\}^n \to \{0,1\}$ be a symmetric function. Then*

$$\deg_{1/3}(f) = \Theta\left(\sqrt{n(\ell_0(f) + \ell_1(f))}\right).$$

In words, Theorem 2.10 states that the $\frac{1}{3}$-approximate degree is $\Omega(\sqrt{n})$ for every nonconstant predicate, and is higher for those predicates that change value near the middle of the range $\{0, 1, \ldots, n\}$.

3 Best Approximation by Polynomials

This section marks the beginning of our proof. The goal here is to determine, within a logarithmic factor, the approximate degree of every predicate. Specifically, we prove the following theorem:

Theorem 1.5 (Restated from p. 4). *Let $D : \{0, 1, \ldots, n\} \to \{0, 1\}$ be a nonconstant predicate. Let $\epsilon \in [1/2^n, 1/3]$. Then*

$$\deg_\epsilon(D) = \tilde{\Theta}\left(\sqrt{n(\ell_0(D) + \ell_1(D))} \;+\; \sqrt{n \log(1/\epsilon)}\right),$$

where the $\tilde{\Theta}$ notation suppresses $\log n$ factors. Furthermore, the approximating polynomial for each D and ϵ is given explicitly.

We prove the upper and lower bounds in this result separately, as Lemma 3.4 and Lemma 3.6, in the two subsections that follow.

3.1 Upper Bound on the Approximate Degree

Our construction makes heavy use of the Chebyshev polynomials, which is not surprising given their fundamental role in approximation. The other key ingredient in our proof is *interpolation*, which here amounts to multiplying an imperfect approximator $p(t)$ by another polynomial $q(t)$ that zeroes out p's mistakes. This interpolation technique is well-known [1, 5] and is vital to exploiting the discrete character of the problem: we are interested in approximation over the discrete set of points $\{0, 1, \ldots, n\}$ rather than the more difficult continuous setting, $[0, n]$. Kahn et al. [5], who obtained the special case of Theorem 1.5 for $D = \mathrm{OR}$, also used the Chebyshev polynomials and interpolation, although in a simpler and much different way.

We start by recalling a few properties of the Chebyshev polynomials, whose proofs can be found in any standard textbook on approximation theory, e.g., [3, 14].

Fact 3.1 (Chebyshev polynomials). *The d^{th} Chebyshev polynomial, $T_d(t)$, has degree d and satisfies the following properties:*

$$T_d(1) = 1 \tag{3.1}$$

$$|T_d(t)| \leqslant 1 \qquad (-1 \leqslant t \leqslant 1) \tag{3.2}$$

$$T_d'(t) \geqslant d^2 \qquad (t \geqslant 1) \tag{3.3}$$

$$T_d(1 + \delta) \geqslant \tfrac{1}{2} \cdot 2^{d\sqrt{2\delta}} \qquad (0 \leqslant \delta \leqslant 1/2) \tag{3.4}$$

$$2 \leqslant T_{\lceil a \rceil}\left(1 + \tfrac{1}{a^2}\right) \leqslant 7 \qquad (a \geqslant 1) \tag{3.5}$$

At the heart of our construction is the following technical lemma.

Lemma 3.2. *Let $\ell \geqslant 0$, $\Delta \geqslant 1$, and $d \geqslant 1$ be integers with $\ell + \Delta \leqslant n/2$. Then there is an (explicitly given) polynomial $p(t)$ of degree at most $22(d+1)\sqrt{n(\ell+\Delta)}/\Delta$ with*

$$p(n-\ell) = 1$$

and

$$|p(t)| \leqslant 2^{-d} \qquad \text{for } t \in [0,n] \setminus (n-\ell-\Delta, n-\ell+\Delta).$$

Proof. Let

$$p_1(t) \stackrel{\text{def}}{=} T_{\left\lceil \sqrt{\frac{n-\ell-\Delta}{\ell+\Delta}} \right\rceil}\left(\frac{t}{n-\ell-\Delta}\right).$$

One readily verifies the following properties of p_1:

$$\left.\begin{array}{ll} p_1([0, n-\ell-\Delta]) \subseteq [-1,1] & \text{by (3.2);} \\ p_1([n-\ell-\Delta, n]) \subseteq [1,7] & \text{by (3.1), (3.3), (3.5);} \\ p_1'(t) \geqslant \dfrac{1}{\ell+\Delta} \text{ for } t \geqslant n-\ell-\Delta & \text{by (3.3);} \\ p_1(n-\ell) - p_1(n-\ell-\Delta) \geqslant \dfrac{\Delta}{\ell+\Delta} & \text{by previous line;} \\ p_1(n-\ell+\Delta) - p_1(n-\ell) \geqslant \dfrac{\Delta}{\ell+\Delta} & \text{likewise.} \end{array}\right\} \tag{3.6}$$

Now consider the polynomial

$$p_2(t) \stackrel{\text{def}}{=} \left(\frac{p_1(t) - p_1(n-\ell)}{8}\right)^2.$$

In view of (3.6), this new polynomial satisfies

$$p_2(n-\ell) = 0$$

and

$$p_2(t) \in \left[\frac{\Delta^2}{64(\ell+\Delta)^2}, 1\right] \qquad \text{for } \ t \in [0,n] \setminus (n-\ell-\Delta, n-\ell+\Delta).$$

Finally, let

$$p_3(t) \stackrel{\text{def}}{=} T_{\left\lceil \frac{8(d+1)(\ell+\Delta)}{\sqrt{2}\Delta} \right\rceil}\left(1 + \frac{\Delta^2}{64(\ell+\Delta)^2} - p_2(t)\right).$$

Using (3.4) and the properties of p_2, one sees that $p(t) = p_3(t)/p_3(n-\ell)$ is the desired polynomial. □

There are a large number of distinct predicates on $\{0, 1, \ldots, n\}$. To simplify the analysis, we would like to work with a small family of predicates that have simple structure yet allow us to efficiently express any other predicate. A natural choice is the family of predicates EXACT_ℓ for $\ell = 0, 1, \ldots, n$, where

$$\mathrm{EXACT}_\ell(t) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } t = \ell, \\ 0 & \text{otherwise.} \end{cases}$$

For a moment, we shall focus on an explicit construction for EXACT_ℓ.

Lemma 3.3. *Let $0 \leqslant \ell \leqslant n/2$. Then for any $\epsilon \leqslant 1/3$,*

$$\deg_\epsilon(\mathrm{EXACT}_\ell) = \deg_\epsilon(\mathrm{EXACT}_{n-\ell}) = O\left(\sqrt{n(\ell+1)}\log n + \sqrt{n\log(1/\epsilon)\log n}\right).$$

Proof. The first equality in the statement of the lemma is obvious, and we concentrate on the second. We may assume that $\ell \leqslant n/\log^2 n$ and $\log(1/\epsilon) \leqslant n/\log n$, since otherwise the claim is trivial. Set

$$\Delta \stackrel{\text{def}}{=} \left\lceil \frac{\log(1/\epsilon)}{\log n} \right\rceil, \qquad d \stackrel{\text{def}}{=} 3\Delta\lceil \log n \rceil.$$

Our assumptions about ℓ and ϵ imply that $\ell + \Delta \ll n/2$, and thus Lemma 3.2 is applicable. Denote by $p(t)$ the polynomial constructed in Lemma 3.2. Let

$$q(t) \stackrel{\text{def}}{=} \prod_{\substack{i=-(\Delta-1),\ldots,(\Delta-1) \\ i\neq 0}} (t - (n - \ell + i)).$$

We claim that the polynomial

$$r(t) \stackrel{\text{def}}{=} \frac{1}{q(n-\ell)} \cdot p(t)q(t)$$

is the sought approximation to $\mathrm{EXACT}_{n-\ell}$. Indeed, it is easy to verify that $r(t)$ has the desired degree. For $t \in \{0, 1, \ldots, n\} \setminus \{n - \ell - (\Delta - 1), \ldots, n - \ell + (\Delta - 1)\}$,

$$|r(t) - \mathrm{EXACT}_{n-\ell}(t)| = |r(t)| \leqslant n^{2(\Delta-1)} \cdot \frac{1}{2^d} \leqslant \epsilon.$$

Since $r(t) = \mathrm{EXACT}_{n-\ell}(t)$ for all remaining t, the proof is complete. □

We now prove the sought upper bound for an arbitrary predicate by repeatedly applying Lemma 3.3.

Lemma 3.4 (Upper bound on the approximate degree). *Let* $D : \{0,1,\ldots,n\} \to \{0,1\}$. *Then for any* $\epsilon \leqslant 1/3$,

$$\deg_\epsilon(D) \leqslant O\Big(\sqrt{n(\ell_0(D)+\ell_1(D))}\,\log n + \sqrt{n\log(1/\epsilon)\log n}\Big).$$

Moreover, the approximating polynomial is given explicitly.

Proof. Without loss of generality, we can assume that $D(\lceil n/2\rceil) = 0$ (otherwise, work with the negation of D). For $\ell = 0,1,\ldots,n$, let $p_\ell(t)$ denote the polynomial that approximates $\mathrm{EXACT}_\ell(t)$ pointwise to within ϵ/n, as constructed in Lemma 3.3. Put

$$p(t) \stackrel{\text{def}}{=} \sum_{\ell\,:\,D(\ell)=1} p_\ell(t).$$

Then clearly $p(t)$ approximates D pointwise to within ϵ. It remains to place an upper bound on the degree of p:

$$\begin{aligned}
\deg_\epsilon(D) &\leqslant \deg p \\
&\leqslant \max_{\substack{\ell\,:\,D(\ell)=1,\\ \ell<\lceil n/2\rceil}} \{\deg p_\ell\} \quad + \quad \max_{\substack{\ell\,:\,D(\ell)=1,\\ \ell>\lceil n/2\rceil}} \{\deg p_{n-\ell}\} \\
&\leqslant O\Big(\Big(\sqrt{n\ell_0(D)} + \sqrt{n\ell_1(D)}\Big)\log n + \sqrt{n\log(n/\epsilon)\log n}\Big) \\
&\leqslant O\Big(\sqrt{n(\ell_0(D)+\ell_1(D))}\log n + \sqrt{n\log(1/\epsilon)\log n}\Big),
\end{aligned}$$

where the third inequality follows by Lemma 3.3. □

3.2 Lower Bound on the Approximate Degree

Our lower bounds follow by a reduction to EXACT_0, the simplest nonconstant predicate, for which Kahn et al. [5] have already proven a tight lower bound.

Theorem 3.5 (Kahn, Linial, and Samorodnitsky [5, Thm. 2.1 and its proof]). *Let* $0 \leqslant k \leqslant n-1$. *Then for every polynomial* p *of degree* k,

$$\max_{i=0,1,\ldots,n} |\mathrm{EXACT}_0(i) - p(i)| \geqslant n^{-\Theta(k^2/n)}.$$

Theorem 3.5 has the following immediate corollary:

Corollary 3.5.1. *Let* $2^{-\Theta(n\log n)} \leqslant \epsilon \leqslant 1/3$. *Then*

$$\deg_\epsilon(\mathrm{EXACT}_0) \geqslant \Omega\left(\sqrt{\frac{n\log(1/\epsilon)}{\log n}}\right).$$

We are now in a position to prove the desired lower bound on the approximate degree of any given predicate.

Lemma 3.6 (Lower bound on the approximate degree). *Let $D : \{0,1,\ldots,n\} \to \{0,1\}$ be a nonconstant predicate. Then for $2^{-\Theta(n \log n)} \leqslant \epsilon \leqslant 1/3$,*

$$\deg_\epsilon(D) \geqslant \Omega\left(\sqrt{n(\ell_0(D) + \ell_1(D))} + \sqrt{\frac{n \log(1/\epsilon)}{\log n}} \right).$$

Proof. In view of Paturi's result (Theorem 2.10), it suffices to show that

$$\deg_\epsilon(D) \geqslant \Omega\left(\sqrt{\frac{n \log(1/\epsilon)}{\log n}} \right). \tag{3.7}$$

Abbreviate $\ell = \ell_0(D)$. We can assume that $\ell \leqslant n/5$ since otherwise the claim follows trivially from Theorem 2.10. Consider the predicate EXACT_0 on $\lfloor n/5 \rfloor$ bits. By Corollary 3.5.1,

$$\deg_\epsilon(\mathrm{EXACT}_0) \geqslant \Omega\left(\sqrt{\frac{n \log(1/\epsilon)}{\log n}} \right) \tag{3.8}$$

On the other hand,

$$\mathrm{EXACT}_0(t) = (1 - 2D(\ell)) \cdot D(t + \ell - 1) + D(\ell),$$

so that

$$\deg_\epsilon(\mathrm{EXACT}_0) \leqslant \deg_\epsilon(D). \tag{3.9}$$

Equations (3.8) and (3.9) imply (3.7), thereby completing the proof. □

4 Approximating a Function of Events

We now turn to the proof of our main results, Theorems 1.2 and 1.3. Fix an arbitrary function $f : \{0,1\}^n \to \{0,1\}$. For events $A_1, \ldots, A_n$ in a probability space $\mathscr{P}$, let

$$\mathbf{Pr}[f(A_1,\ldots,A_n)] \stackrel{\text{def}}{=} \mathbf{Pr}\left[\bigcup_{x : f(x)=1} \left(\bigcap_{i : x_i = 0} \overline{A_i} \quad \bigcap_{i : x_i = 1} A_i \right) \right].$$

Suppose that $\mathbf{Pr}[\bigcap_{i \in S} A_i]$ is given for each S with $|S| \leqslant k$. Our goal here is show how to use this information to efficiently construct a near-optimal approximation to $\mathbf{Pr}[f(A_1,\ldots,A_n)]$. Our discussion will revolve around the quantity $\delta^*(f,k)$, defined next.

Definition 4.1. Let $f : \{0,1\}^n \to \{0,1\}$ and $0 \leqslant k \leqslant n$. Define

$$\delta^*(f,k) \stackrel{\text{def}}{=} \sup \left\{ \mathop{\mathbf{Pr}}_{\mathscr{P}_1}[f(A_1,\ldots,A_n)] - \mathop{\mathbf{Pr}}_{\mathscr{P}_2}[f(B_1,\ldots,B_n)] \right\},$$

where the supremum is taken over all probability spaces $\mathscr{P}_1$ and $\mathscr{P}_2$, over all events $A_1,\ldots,A_n$ in $\mathscr{P}_1$, and over all events $B_1,\ldots,B_n$ in $\mathscr{P}_2$, such that

$$\mathop{\mathbf{Pr}}_{\mathscr{P}_1}\left[\bigcap_{i\in S} A_i\right] = \mathop{\mathbf{Pr}}_{\mathscr{P}_2}\left[\bigcap_{i\in S} B_i\right] \qquad \text{for } |S| \leqslant k. \tag{4.1}$$

In words, $\delta^*(f,k)$ is the best error achievable in approximating $\mathbf{Pr}[f(A_1,\ldots,A_n)]$ in principle, information-theoretically, if unlimited computing power is available.

For a symmetric function $f(x) \equiv D(x_1 + \cdots + x_n)$, the notation we have established in this section relates as follows to the notation of the Introduction:

$$\mathbf{Pr}[f(A_1,\ldots,A_n)] = \mathbf{Pr}\left[D(\mathbf{I}[A_1] + \cdots + \mathbf{I}[A_n]) = 1\right],$$

$$\delta^*(f,k) = \delta^*(D,k).$$

We need the more general notation because much of the development in this section takes place in the setting of arbitrary functions $f : \{0,1\}^n \to \{0,1\}$, even though our ultimate results are for *symmetric* functions. This approach makes the proof cleaner and more modular, in addition to yielding partial results for *nonsymmetric* functions.

Our immediate goal is to understand the quantitative behavior of $\delta^*(f,k)$. To this end, we will show that the arbitrary probability spaces in the definition of $\delta^*(f,k)$ can in fact be restricted to probability distributions on $\{0,1\}^n$.

Definition 4.2 (Induced distribution). Let $E_1,\ldots,E_n$ be events in a probability space $\mathscr{P}$. The *distribution on* $\{0,1\}^n$ *induced by* $\mathscr{P}, E_1,\ldots,E_n$ is defined as

$$\mu(x) \stackrel{\text{def}}{=} \mathbf{Pr}\left[\bigcap_{i:x_i=0} \overline{E_i} \quad \bigcap_{i:x_i=1} E_i\right].$$

Proposition 4.3. *Let $E_1,\ldots,E_n$ be events in a probability space $\mathscr{P}$. Let μ be the distribution on $\{0,1\}^n$ induced by $\mathscr{P}, E_1,\ldots,E_n$. Then for every $g : \{0,1\}^n \to \{0,1\}$,*

$$\mathbf{Pr}[g(E_1,\ldots,E_n)] = \mathop{\mathbf{E}}_{x\sim\mu}[g(x)].$$

Proof:

$$\begin{aligned}\mathbf{Pr}[g(E_1,\ldots,E_n)] &= \sum_{x\in\{0,1\}^n} g(x)\cdot\mathbf{Pr}\left[\bigcap_{i:x_i=0}\overline{E_i}\quad\bigcap_{i:x_i=1}E_i\right] = \sum_{x\in\{0,1\}^n} g(x)\mu(x)\\ &= \mathop{\mathbf{E}}_{x\sim\mu}[g(x)].\end{aligned}$$

□

At this point, we are ready to simplify $\delta^*(f,k)$ as promised. For a set $S\subseteq[n]$, define $\mathrm{AND}_S:\{0,1\}^n\to\{0,1\}$ by

$$\mathrm{AND}_S(x) \stackrel{\text{def}}{=} \bigwedge_{i\in S} x_i = \prod_{i\in S} x_i.$$

In particular, $\mathrm{AND}_\emptyset\equiv 1$.

Lemma 4.4. *Let $f:\{0,1\}^n\to\{0,1\}$ and $0\leqslant k\leqslant n$. Then*

$$\delta^*(f,k) = \max_{\alpha,\beta}\left\{\mathop{\mathbf{E}}_{x\sim\alpha}[f(x)] - \mathop{\mathbf{E}}_{x\sim\beta}[f(x)]\right\}, \tag{4.2}$$

where the maximum is taken over all probability distributions α,β on $\{0,1\}^n$ such that $\mathbf{E}_{x\sim\alpha}[\mathrm{AND}_S(x)] = \mathbf{E}_{x\sim\beta}[\mathrm{AND}_S(x)]$ for $|S|\leqslant k$.

Proof. Fix probability spaces $\mathscr{P}_1,\mathscr{P}_2$, events $A_1,\ldots,A_n$ in $\mathscr{P}_1$, and events $B_1,\ldots,B_n$ in $\mathscr{P}_2$, such that (4.1) holds. Let α and β be the distributions on $\{0,1\}^n$ induced by $\mathscr{P}_1,A_1,\ldots,A_n$ and $\mathscr{P}_2,B_1,\ldots,B_n$, respectively. Then by Proposition 4.3,

$$\mathop{\mathbf{E}}_{x\sim\alpha}[f(x)] - \mathop{\mathbf{E}}_{x\sim\beta}[f(x)] \quad=\quad \mathop{\mathbf{Pr}}_{\mathscr{P}_1}[f(A_1,\ldots,A_n)] - \mathop{\mathbf{Pr}}_{\mathscr{P}_2}[f(B_1,\ldots,B_n)]$$

and

$$\mathop{\mathbf{E}}_{x\sim\alpha}[\mathrm{AND}_S(x)] \quad=\quad \mathop{\mathbf{E}}_{x\sim\beta}[\mathrm{AND}_S(x)] \qquad \text{for } |S|\leqslant k.$$

Letting δ stand for the right-hand side of (4.2), we conclude that $\delta^*(f,k)\leqslant\delta$.

It remains to show that $\delta^*(f,k)\geqslant\delta$. Given a probability distribution μ on $\{0,1\}^n$, there is an obvious discrete probability space $\mathscr{P}$ and events $E_1,\ldots,E_n$ in it that induce μ: simply let $\mathscr{P}=\{0,1\}^n$ with E_i defined to be the event that $x_i=1$, where $x\in\{0,1\}^n$ is distributed according to μ. This allows us to reverse the argument of the previous paragraph (again using Proposition 4.3) and show that $\delta^*(f,k)\geqslant\delta$. □

With $\delta^*(f,k)$ thus simplified, we relate it to a quantity that is easy to estimate.

Theorem 1.4 (Restated from p. 4). *Let* $f : \{0,1\}^n \to \{0,1\}$ *be arbitrary and* $0 \leqslant k \leqslant n$. *Then*

$$\delta^*(f,k) = 2\,\epsilon^*(f,\Phi),$$

where $\Phi = \{\mathrm{AND}_S : |S| \leqslant k\}$.

Proof. In view of the Approximation/Orthogonality Principle (Theorem 2.3), it suffices to prove that

$$\delta^*(f,k) = 2\,\gamma^*(f,\Phi).$$

The remainder of the proof is establishes this equality.

To rephrase Lemma 4.4,

$$\delta^*(f,k) = \max_{\alpha,\beta} \left\{ \sum_{x\in\{0,1\}^n} [\alpha(x) - \beta(x)]\, f(x) \right\}, \tag{4.3}$$

where the maximum is over distributions α and β on $\{0,1\}^n$ such that

$$\sum_{x\in\{0,1\}^n} [\alpha(x) - \beta(x)]\, \mathrm{AND}_S(x) = 0 \qquad \text{for } |S| \leqslant k.$$

Let α, β be distributions for which the maximum is attained in (4.3). Setting $\psi = (\alpha - \beta)/2$, we see that $\sum_{x\in\{0,1\}^n} |\psi(x)| \leqslant 1$ and thus $\delta^*(f,k) \leqslant 2\,\gamma^*(f,\Phi)$.

It remains to show that $\gamma^*(f,\Phi) \leqslant \delta^*(f,k)/2$. Suppose first that $\gamma^*(f,\Phi) = 0$. Since $\delta^*(f,k) \geqslant 0$ always and $\delta^*(f,k) \leqslant 2\,\gamma^*(f,\Phi) = 0$ by the first part of the proof, the theorem is true in this case.

Finally, suppose that $\gamma^*(f,\Phi) > 0$ and let ψ be a real function for which the maximum is achieved in (2.1). Then necessarily $\sum_{x\in\{0,1\}^n} |\psi(x)| = 1$. Since ψ is orthogonal to the constant function $1 \in \Phi$, we also have $\sum_{x\in\{0,1\}^n} \psi(x) = 0$. The last two sentences allow us to write

$$\psi = \frac{1}{2}(\alpha - \beta),$$

where α and β are suitable probability distributions over $\{0,1\}^n$. Then (4.3) shows that $\gamma^*(f,\Phi) \leqslant \delta^*(f,k)/2$, as desired. □

Theorem 1.4, which we have just proved, is the crux of our argument. It shows that $\delta^*(f,k)$ measures how well f can be approximated by a multivariate polynomial in $x_1, \ldots, x_n$ of degree k. Observe that Theorem 1.4 holds for *every* function $f : \{0,1\}^n \to \{0,1\}$. For the special case of symmetric functions, we have already obtained (Section 3) tight estimates of the best error achievable by a polynomial of a given degree k. By combining these estimates with Theorem 1.4, we now prove the main result of the paper.

Theorem 4.5 (Restatement of Theorems 1.2 and 1.3). *Let $f : \{0,1\}^n \to \{0,1\}$ be a nonconstant symmetric function. Put $\ell = \ell_0(f) + \ell_1(f)$. Then*

$$\begin{aligned}
&\delta^*(f,k) = \Theta(1) && \text{if} \qquad k \leqslant \Theta\left(\sqrt{n\ell}\right), \\
&\delta^*(f,k) \in \left[2^{-\Theta\left(\frac{k^2 \log n}{n}\right)}, 2^{-\Theta\left(\frac{k^2}{n \log n}\right)}\right] && \text{if} \qquad \Theta\left(\sqrt{n\ell}\log n\right) \leqslant k \leqslant \Theta(n).
\end{aligned}$$

Furthermore, for every $k \geqslant \Theta\left(\sqrt{n\ell}\log n\right)$, there are reals $a_0, a_1, \ldots, a_k$, computable in time poly(n)*, such that*

$$\left| \mathbf{Pr}[f(A_1,\ldots,A_n)] \quad - \quad \sum_{j=0}^{k} a_j \sum_{S:|S|=j} \mathbf{Pr}\left[\bigcap_{i\in S} A_i\right] \right| \leqslant 2^{-\Theta\left(\frac{k^2}{n\log n}\right)}$$

for any events $A_1, \ldots, A_n$ in any probability space $\mathscr{P}$.

Proof. By hypothesis, $f(x) \equiv D(x_1 + \cdots + x_n)$ for a suitable nonconstant predicate $D : \{0,1,\ldots,n\} \to \{0,1\}$. Put $\Phi = \{\mathrm{AND}_S : |S| \leqslant k\}$. We have:

$$\begin{aligned}
\delta^*(f,k) &= 2\,\epsilon^*(f,\Phi) && \text{by Theorem 1.4} \\
&= 2\,\epsilon^*(f,\{\chi_S\}_{|S|\leqslant k}) && \text{since } \mathrm{span}(\Phi) = \mathrm{span}(\{\chi_S\}_{|S|\leqslant k}) \\
&= 2\,\epsilon^*(D,P_k) && \text{by Proposition 2.8.} \qquad (4.4)
\end{aligned}$$

By Theorem 2.10 and Lemmas 3.4 and 3.6,

$$\epsilon^*(D,P_k) \in \begin{cases} \Theta(1) & \text{if} \qquad k \leqslant \Theta\left(\sqrt{n\ell}\right), \\ \left[2^{-\Theta\left(\frac{k^2\log n}{n}\right)}, 2^{-\Theta\left(\frac{k^2}{n\log n}\right)}\right] & \text{if} \qquad \Theta\left(\sqrt{n\ell}\log n\right) \leqslant k \leqslant \Theta(n). \end{cases}$$

In view of (4.4), this proves the claim regarding $\delta^*(f,k)$.

We now turn to the claim regarding $a_0, a_1, \ldots, a_k$. For $k \geqslant \Theta\left(\sqrt{n\ell}\log n\right)$, Lemma 3.4 gives an explicit univariate polynomial $p(t)$ of degree at most k such that

$$|f(x) - p(x_1 + \cdots + x_n)| \leqslant 2^{-\Theta\left(\frac{k^2}{n\log n}\right)} \qquad \text{for all } x \in \{0,1\}^n. \qquad (4.5)$$

Fix a probability space $\mathscr{P}$ and events $A_1, \ldots, A_n$ in it. Let μ be the distribution on $\{0,1\}^n$ induced by $\mathscr{P}, A_1, \ldots, A_n$. We claim that the quantity

$$\mathop{\mathbf{E}}_{x\sim\mu}[p(x_1 + \cdots + x_n)]$$

is the desired approximator of $\mathbf{Pr}[f(A_1,\ldots,A_n)]$. Indeed,

$$\begin{aligned}\mathop{\mathbf{E}}_{x\sim\mu}[p(x_1+\cdots+x_n)] &= \mathop{\mathbf{E}}_{x\sim\mu}\left[\sum_{j=0}^{k} a_j \sum_{|S|=j}\prod_{i\in S} x_i\right] = \sum_{j=0}^{k} a_j \sum_{|S|=j}\mathop{\mathbf{E}}_{x\sim\mu}\left[\prod_{i\in S} x_i\right] \\ &\overset{\text{Prop. 4.3}}{=} \sum_{j=0}^{k} a_j \sum_{|S|=j}\mathbf{Pr}\left[\bigcap_{i\in S} A_i\right],\end{aligned}$$

where the reals $a_0, a_1, \ldots, a_k$ are uniquely determined by the polynomial p, itself explicitly given. It is also clear that $a_0, a_1, \ldots, a_k$ can be computed from the coefficients of p in time $\mathsf{poly}(n)$. Thus, the quantity $\mathbf{E}_{x\sim\mu}[p(x_1+\cdots+x_n)]$ has the desired representation. It remains to verify that it approximates $\mathbf{Pr}[f(A_1,\ldots,A_n)]$ as claimed:

$$\begin{aligned}\left|\mathbf{Pr}[f(A_1,\ldots,A_n)] - \mathop{\mathbf{E}}_{x\sim\mu}[p(x_1+\cdots+x_n)]\right| &\overset{\text{Prop. 4.3}}{=} \left|\mathop{\mathbf{E}}_{x\sim\mu}[f(x)-p(x_1+\cdots+x_n)]\right| \\ &\overset{(4.5)}{\leqslant} 2^{-\Theta\left(\frac{k^2}{n\log n}\right)}.\end{aligned}$$

□

5 Lower Bounds for Agnostic Learning

We now use the proof technique of the previous section to obtain new lower bounds for agnostic learning (Theorem 1.6). The following definition formalizes the object of our study.

Definition 5.1. Let $f : \{0,1\}^n \to \{0,1\}$ and $0 \leqslant k \leqslant n$. Define

$$\Gamma^*(f,k) \overset{\text{def}}{=} \max_{\lambda}\left\{\mathop{\mathbf{Pr}}_{(x,y)\sim\lambda}[f(x)=y]\right\},$$

where the maximum is taken over all distributions λ over $\{0,1\}^n \times \{0,1\}$ such that

$$\mathop{\mathbf{Pr}}_{(x,y)\sim\lambda}[g(x)=y] = \frac{1}{2} \tag{5.1}$$

for every $g : \{0,1\}^n \to \{0,1\}$ that depends on k or fewer variables.

Observe that the maximization in Definition 5.1 is over a nonempty compact set that contains the uniform distribution. Our goal will be to show that

$$\Gamma^*\left(f,\ \Theta\left(\sqrt{n(\ell_0(f)+\ell_1(f))}\right)\right) \geqslant 1-\epsilon$$

for every symmetric function f and every constant $\epsilon > 0$. In other words, even though the training examples agree with f to within ϵ, no hypothesis that depends

on few variables can match the data better than random. Our strategy will be to relate $\Gamma^*(f,k)$ to the best error and modulus of orthogonality, quantities for which have developed considerable intuition.

Lemma 5.2. *Let λ be a distribution on $\{0,1\}^n \times \{0,1\}$. Then for every $f : \{0,1\}^n \to \{0,1\}$,*

$$\mathop{\mathbf{Pr}}_{(x,y)\sim\lambda}[f(x) = y] = \mathop{\mathbf{Pr}}_{(x,y)\sim\lambda}[y = 0] + \sum_{x\in\{0,1\}^n} (\lambda(x,1) - \lambda(x,0)) f(x).$$

Proof:

$$\begin{aligned}
\mathop{\mathbf{Pr}}_{(x,y)\sim\lambda}[f(x) = y] &= \mathop{\mathbf{Pr}}_{(x,y)\sim\lambda}[f(x) = y = 0] + \mathop{\mathbf{Pr}}_{(x,y)\sim\lambda}[f(x) = y = 1] \\
&= \sum_x \lambda(x,0)(1 - f(x)) + \sum_x \lambda(x,1) f(x) \\
&= \sum_x (\lambda(x,1) - \lambda(x,0)) f(x) + \sum_x \lambda(x,0) \\
&= \sum_x (\lambda(x,1) - \lambda(x,0)) f(x) + \mathop{\mathbf{Pr}}_{(x,y)\sim\lambda}[y = 0]. \qquad \square
\end{aligned}$$

We are now in a position to express $\Gamma^*(f,k)$ in terms of a quantity that is easy to estimate.

Theorem 5.3. *Let $f : \{0,1\}^n \to \{0,1\}$ and $0 \leqslant k \leqslant n$. Then*

$$\Gamma^*(f,k) = \frac{1}{2} + \epsilon^*(f,\Phi),$$

where $\Phi = \{\chi_S : |S| \leqslant k\}$.

Proof. By the Approximation/Orthogonality Principle (Theorem 2.3), it suffices to show that

$$\Gamma^*(f,k) = \frac{1}{2} + \gamma^*(f,\Phi).$$

Let λ be a distribution on $\{0,1\}^n \times \{0,1\}$ for which (5.1) holds. Setting $g = 0$ gives:

$$\mathop{\mathbf{Pr}}_{(x,y)\sim\lambda}[y = 0] = \frac{1}{2}.$$

Lemma 5.2 now yields the following convenient characterization of $\Gamma^*(f,k)$:

$$\Gamma^*(f,k) = \frac{1}{2} + \max_\lambda \left\{ \sum_x (\lambda(x,1) - \lambda(x,0)) f(x) \right\},$$

where the maximum is over all distributions λ on $\{0,1\}^n \times \{0,1\}$ such that

$$\sum_x (\lambda(x,1) - \lambda(x,0))g(x) = 0$$

for every function $g : \{0,1\}^n \to \{0,1\}$ that depends on k or fewer variables. With this new characterization, it is not difficult to show that $\Gamma^*(f,k) = 1/2 + \gamma^*(f,\Phi)$. The argument is closely analogous to the one we gave in Theorem 1.4, and we do not repeat it here. □

Theorem 5.3 is the backbone of this section and holds for arbitrary functions. In view of Paturi's work, it yields our sought result for symmetric functions.

Theorem 1.6 (Restated from p. 6). *Let $D : \{0,1,\ldots,n\} \to \{0,1\}$ be a predicate and $f(x) \stackrel{\text{def}}{=} D(x_1 + \cdots + x_n)$. Let $\epsilon > 0$ be an arbitrary constant. Then there is a distribution λ on $\{0,1\}^n \times \{0,1\}$ such that*

$$\mathop{\mathbf{Pr}}_{(x,y)\sim\lambda}[f(x) = y] \geqslant 1 - \epsilon$$

and

$$\mathop{\mathbf{Pr}}_{(x,y)\sim\lambda}[g(x) = y] = \frac{1}{2}$$

for every $g : \{0,1\}^n \to \{0,1\}$ that depends on at most $c\sqrt{n(\ell_0(D) + \ell_1(D))}$ variables, where $c = c(\epsilon)$ is a constant.

Proof. In view of Theorem 5.3, we need only show that

$$\epsilon^*(f,\Phi) \geqslant \frac{1}{2} - \epsilon,$$

where $\Phi = \{\chi_S : |S| \leqslant c\sqrt{n(\ell_0(f) + \ell_1(f))}\}$ for a suitably small constant c. But this is immediate from Proposition 2.5 and Paturi's result (Theorem 2.10). □

Theorem 1.6 is best possible, as we now show.

Theorem 5.4 (On the tightness of Thm. 1.6). *Let $f : \{0,1\}^n \to \{0,1\}$ be a symmetric function and $\epsilon \in (0,1/2)$ be a given constant. Let λ be a distribution on $\{0,1\}^n \times \{0,1\}$ with*

$$\mathop{\mathbf{Pr}}_{(x,y)\sim\lambda}[g(x) = y] = \frac{1}{2}$$

for every $g : \{0,1\}^n \to \{0,1\}$ that depends on at most $C\sqrt{n(\ell_0(f) + \ell_1(f))}$ variables, where $C = C(\epsilon)$ is a large enough constant. Then

$$\mathop{\mathbf{Pr}}_{(x,y)\sim\lambda}[f(x) = y] \leqslant 1 - \epsilon.$$

Proof. To rephrase the theorem, we need to show that

$$\Gamma^*(f,k) \leqslant 1-\epsilon,$$

where $k = C\sqrt{n(\ell_0(f)+\ell_1(f))}$. In view of Theorem 5.3, this is equivalent to

$$\epsilon^*\Big(f,\ \{\chi_S : |S| \leqslant k\}\Big) \leqslant \frac{1}{2} - \epsilon.$$

The latter is certainly true for a large enough constant C, by Proposition 2.5 and Paturi's result (Theorem 2.10). □

Remark 5.5. Let f be an arbitrary symmetric function. Theorem 5.4 tells us that if all hypotheses that depend on at most $k = \Theta\Big(\sqrt{n(\ell_0(f)+\ell_1(f))}\Big)$ variables have zero advantage over random guessing, then the function f itself cannot be a *high-accuracy* classifier. What if we additionally know that all hypotheses that depend on at most K variables, where

$$K \gg \Theta\Big(\sqrt{n(\ell_0(f)+\ell_1(f))}\Big),$$

have zero advantage over random guessing? It turns out that in this case, the function f itself cannot have considerable *advantage* over random guessing (let alone be a *high-accuracy* classifier). The proof is entirely analogous to that of Theorem 5.4, except in place of Paturi's result we would use our our near-tight bounds on the approximate degree (Theorem 1.5) that work in the broader range $[1/2^n, 1/3]$. Such statements seem to be of lesser interest, and we do not formulate them into theorems.

References

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

[2] S. N. Bernstein. Sur la meilleure approximation de $|x|$ par des polynômes de degrés donnés. *Acta Math.*, 37:1–57, 1914.

[3] E. W. Cheney. *Introduction to Approximation Theory*. Chelsea Publishing, New York, 2nd edition, 1982.

[4] A. Eremenko and P. Yuditskii. Uniform approximation of sgn(x) by polynomials and entire functions. *J. d'Analyse Mathématique*, 101:313–324, 2007.

[5] J. Kahn, N. Linial, and A. Samorodnitsky. Inclusion-exclusion: Exact and approximate. *Combinatorica*, 16(4):465–477, 1996.

[6] A. T. Kalai, A. R. Klivans, Y. Mansour, and R. A. Servedio. Agnostically learning halfspaces. In *Proc. of the 46th Symposium on Foundations of Computer Science (FOCS)*, pages 11–20, 2005.

[7] M. J. Kearns, R. E. Schapire, and L. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2–3):115–141, 1994.

[8] A. R. Klivans and R. Servedio. Learning DNF in time $2^{\tilde{O}(n^{1/3})}$. In *Proc. of the 33rd Symposium on Theory of Computing (STOC)*, pages 258–265, 2001.

[9] A. R. Klivans and A. A. Sherstov. A lower bound for agnostically learning disjunctions. In *Proc. of the 20th Conf. on Learning Theory (COLT)*, pages 409–423, 2007.

[10] N. Linial and N. Nisan. Approximate inclusion-exclusion. *Combinatorica*, 10(4):349–365, 1990.

[11] M. L. Minsky and S. A. Papert. *Perceptrons: expanded edition*. MIT Press, Cambridge, MA, USA, 1988.

[12] R. Paturi. On the degree of polynomials that approximate symmetric Boolean functions. In *Proc. of the 24th Symposium on Theory of Computing*, pages 468–474, 1992.

[13] A. A. Razborov. Quantum communication complexity of symmetric predicates. *Izvestiya: Mathematics*, 67(1):145–159, 2003.

[14] T. J. Rivlin. *An Introduction to the Approximation of Functions*. Dover Publications, New York, 1981.

[15] A. A. Sherstov. A discrepancy-based proof of Razborov's quantum lower bounds. Technical Report TR-07-33, The Univ. of Texas at Austin, Dept. of Computer Sciences, July 2007.

[16] A. A. Sherstov. Separating AC^0 from depth-2 majority circuits. In *Proc. of the 39th Symposium on Theory of Computing (STOC)*, pages 294–301, 2007.

[17] J. Tarui and T. Tsukiji. Learning DNF by approximating inclusion-exclusion formulae. In *Proc. of the 14th Conf. on Computational Complexity*, page 215, 1999.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de/