comput. complex. **18** (2009), 219–247
1016-3328/09/020219-29, *published online* 12 June 2009
DOI 10.1007/s00037-009-0274-4

computational complexity

APPROXIMATE INCLUSION-EXCLUSION FOR ARBITRARY SYMMETRIC FUNCTIONS

Alexander A. Sherstov

Abstract. Let $A_1, \ldots, A_n$ be events in a probability space. The *approximate inclusion-exclusion problem,* due to Linial and Nisan (1990), is to estimate $\mathbf{P}[A_1 \cup \cdots \cup A_n]$ given $\mathbf{P}[\bigcap_{i \in S} A_i]$ for $|S| \leq k$. Kahn *et al.* (1996) solved this problem optimally for each k. We study the following more general question: estimate $\mathbf{P}[f(A_1, \ldots, A_n)]$ given $\mathbf{P}[\bigcap_{i \in S} A_i]$ for $|S| \leq k$, where $f\colon \{0,1\}^n \to \{0,1\}$ is a given symmetric function. We solve this general problem for every f 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 analyze, for every nonconstant symmetric $f\colon \{0,1\}^n \to \{0,1\}$ and every $\epsilon \in [2^{-n}, 1/3]$, the least degree $\deg_\epsilon(f)$ of a polynomial that approximates f pointwise within ϵ. We show that $\deg_\epsilon(f) = \tilde{\Theta}(\deg_{1/3}(f) + \sqrt{n \log(1/\epsilon)})$, where $\deg_{1/3}(f)$ is well-known for each f. Previously, the answer for vanishing ϵ was known only for $f = \mathrm{OR}$. We construct the approximating polynomial explicitly for all f and ϵ.

Keywords. Approximate inclusion/exclusion, approximate degree of Boolean functions, approximation by polynomials.

Subject classification. 03D15, 68Q17.

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 $A_1, A_2, \ldots, A_n$:

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

A moment's reflection shows that knowledge of every term in this summation is necessary in general for an exact answer (Linial & Nisan 1990). It is therefore natural to wonder if one can closely approximate $\mathbf{P}[\bigcup A_i]$ using the probabilities of intersections of up to k events, where $k \ll n$. This problem, due to Linial and Nisan (1990), 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 of k. A follow-up article by Kahn, Linial, and Samorodnitsky (1996) 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 the probability that *most* of the events $A_1, \ldots, A_n$ occur, or the probability that an *odd number* of the events from among $A_1, \ldots, A_n$ occur. More generally, let $f\colon \{0,1\}^n \to \{0,1\}$ be a given Boolean function. The problem of interest to us is that of estimating

$$\mathbf{P}\left[f(A_1, \ldots, A_n)\right]$$

given $\mathbf{P}[\bigcap_{i\in S} A_i]$ for $|S| \le k$. Our approach is different from the previous methods (Kahn *et al.* 1996; Linial & Nisan 1990), which are specialized to the case $f = \mathrm{OR}$. Our first contribution is to show that the inclusion/exclusion problem for a given f is exactly equivalent to a classical approximation problem. Specifically, define

$$\delta^*(f,k) = \frac{1}{2}\sup\left\{ \mathop{\mathbf{P}}_{\mathcal{P}_1}\left[f(A_1,\ldots,A_n)\right] - \mathop{\mathbf{P}}_{\mathcal{P}_2}\left[f(B_1,\ldots,B_n)\right] \right\},$$

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

$$\mathop{\mathbf{P}}_{\mathcal{P}_1}\left[\bigcap_{i\in S} A_i\right] = \mathop{\mathbf{P}}_{\mathcal{P}_2}\left[\bigcap_{i\in S} B_i\right] \quad \text{for} \quad |S| \le k\,.$$

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

Theorem 1.1. *Let $f\colon \{0,1\}^n \to \{0,1\}$ be arbitrary and $0 \leq k \leq n$. Then*

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

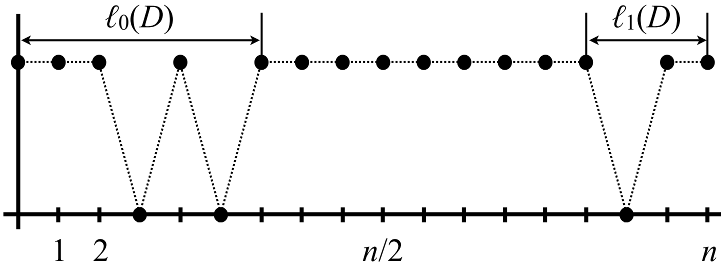
where the minimum is over polynomials $\phi(x_1,\ldots,x_n)$ of degree up to k.

Theorem 1.1 states that the approximate inclusion/exclusion problem for a given f is equivalent to the problem of approximating f by a multivariate polynomial of degree up to k. We are able to solve the approximation problem for all symmetric functions, thereby solving the corresponding inclusion/exclusion problem. The next three subsections give a detailed description of our results.

1.1. Inclusion/exclusion. Let f be a symmetric function, i.e., $f(x) = D(x_1 + \cdots + x_n)$ for some predicate $D\colon \{0,1,\ldots,n\} \to \{0,1\}$. Our results are in terms of the quantities

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

defined as the smallest integers such that D is constant in $[\ell_0(D), n - \ell_1(D)]$. These quantities arise frequently in the study of symmetric functions (Paturi 1992; Razborov 2002). 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. The first question that we settle is precisely how large k needs to be for a good approximation to exist. Recall that the $\tilde{\Theta}$ notation indicates equality within a polylogarithmic factor. We prove:

THEOREM 1.2 (Existence of a good approximation). *Let $f(x) = D(x_1 + \cdots + x_n)$, where $D\colon \{0, 1, \ldots, n\} \to \{0, 1\}$ is a given nonconstant predicate. Put $\ell = \ell_0(D) + \ell_1(D)$. Then*

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

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

THEOREM 1.3 (Efficient approximation scheme). *Let $f(x) = D(x_1 + \cdots + x_n)$, where $D\colon \{0, 1, \ldots, n\} \to \{0, 1\}$ is a given nonconstant predicate. Put $\ell = \ell_0(D) + \ell_1(D)$. Then for every $k \geq \tilde{\Theta}(\sqrt{n\ell})$ there are reals*

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

computable in time poly(n), *such that*

$$\left| \mathbf{P}\left[f(A_1, \ldots, A_n)\right] - \sum_{j=0}^{k} a_j \sum_{S:|S|=j} \mathbf{P}\left[\bigcap_{i\in S} A_i\right] \right| \leq 2^{-\tilde{\Theta}(k^2/n)}$$

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

Theorem 1.3 gives the desired approximation algorithm. Note that it is not necessary to know the individual probabilities $\mathbf{P}[\bigcap_{i\in S} A_i]$; it suffices to know the $k+1$ sums

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

Our proof takes inspiration from the elegant papers of Linial and Nisan (1990) and Kahn *et al.* (1996), who obtained analogues of Theorems 1.2 and 1.3 for the special case $f = \text{OR}$. 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.

First of all, the linear-programming reductions of Linial & Nisan (1990) and Kahn *et al.* (1996) are restricted to $f = \text{OR}$. To handle arbitrary Boolean functions f, we start with a different and more versatile tool (Ioffe

& Tikhomirov 1968), which gives a certain equivalence between approximation and orthogonality in Euclidean space. With some work, this yields Theorem 1.1, which is the desired reduction from the original problem to a question in approximation theory. The new question amounts to determining, for each predicate D and each $\epsilon \in [2^{-n}, 1/3]$, 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 $D = \mathrm{OR}$ (Buhrman *et al.* 1999; Kahn *et al.* 1996). We solve the general case for all D and ϵ, with details to follow next.

1.2. Approximation by polynomials. As just outlined, a key part of our proof is the following result of independent interest. For a predicate $D\colon \{0,1,\dots,n\} \to \{0,1\}$, its *$\epsilon$-approximate degree* $\deg_\epsilon(D)$ is the smallest degree of a univariate real polynomial p that approximates D pointwise to within ϵ:

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

The approximation of predicates is synonymous with the approximation of symmetric Boolean functions. Namely, it is well-known (Minsky & Papert 1988) that $\deg_\epsilon(D)$ is the least degree of a multivariate polynomial $\phi(x_1,\dots,x_n)$ with

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

where $f(x) = D(x_1 + \cdots + x_n)$ is the symmetric Boolean function that corresponds to D. We prove:

THEOREM 1.4 (Approximate degree of predicates). *Let $D\colon \{0,1,\dots,n\} \to \{0,1\}$ be a nonconstant predicate. Let $\epsilon \in [2^{-n}, 1/3]$. Then*

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

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

In words, Theorem 1.4 rather fully characterizes the ℓ_∞-approximation of symmetric Boolean functions. Approximation of Boolean functions by real polynomials in the ℓ_∞ norm is a fundamental subject in complexity theory, and several studies have been made of the approximate degree of selected Boolean functions (Aaronson & Shi 2004; Paturi 1992). In addition to its intrinsic value as a subject in complexity theory, ℓ_∞-approximation has enabled substantial progress on several important problems. Perhaps the most illustrative example is the study of quantum communication and query complexity (Beals *et al.* 2001; Buhrman *et al.* 1999; Razborov 2002; Sherstov 2008a; Shi & Zhu 2007). More recently, ℓ_∞-approximation has played an increasing role in the study of classical communication (Buhrman *et al.* 2007; Sherstov 2008b). Another application is computational learning theory (Kalai *et al.* 2005; Klivans & Servedio 2004; Klivans & Sherstov 2007), where ℓ_∞-approximation has contributed both upper bounds and lower bounds. Finally, the inclusion/exclusion problem (Kahn *et al.* 1996; Linial & Nisan 1990) illustrates the use of ℓ_∞-approximation in algorithm design.

Theorem 1.4 is a broad generalization of several earlier results in the literature. The first of these is due to Paturi (1992), who showed that

$$\deg_{1/3}(D) = \Theta\Big(\sqrt{n\big(\ell_0(D) + \ell_1(D)\big)}\Big) \quad \text{for all} \quad D\,.$$

Unfortunately, Paturi's result and its proof give no insight into the behavior of the ϵ-approximate degree for vanishing ϵ. Another relevant result is due to Kahn *et al.* (1996), who conducted an in-depth study of the predicate $D = \mathrm{OR}$, defined as usual by $\mathrm{OR}(i) = 1 \Leftrightarrow i \geq 1$. Kahn *et al.* showed that

$$\deg_\epsilon(\mathrm{OR}) = \tilde{\Theta}\big(\sqrt{n \log(1/\epsilon)}\big) \qquad (2^{-n} \leq \epsilon \leq 1/3)\,,$$

where the $\tilde{\Theta}$ notation hides $\log n$ factors. Using different techniques, Buhrman *et al.* (1999) gave the final, exact answer for $D = \mathrm{OR}$:

$$\deg_\epsilon(\mathrm{OR}) = \Theta\big(\sqrt{n \log(1/\epsilon)}\big) \qquad (2^{-n} \leq \epsilon \leq 1/3)\,.$$

Thus, our work generalizes the above results to every predicate and every error rate $\epsilon \in [2^{-n}, 1/3]$.

Theorem 1.4 has another, more revealing interpretation. In view of Paturi's work, it can be restated as:

$$(1.5) \qquad \deg_\epsilon(D) = \tilde{\Theta}\big(\deg_{1/3}(D) + \sqrt{n \log(1/\epsilon)}\big) \qquad (2^{-n} \leq \epsilon \leq 1/3)\,,$$

where D is nonconstant. In words, past a certain threshold, the dependence of the ϵ-approximate degree on ϵ is essentially 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.

Recent progress. Equation (1.5) in this paper determines the ϵ-approximate degree of every predicate to within logarithmic factors. Ronald de Wolf (2008) has recently improved our bounds to a final, tight answer: $\deg_\epsilon(D) = \Theta(\deg_{1/3}(D) + \sqrt{n\log(1/\epsilon)})$ for every nonconstant D and every $\epsilon \in [2^{-n}, 1/3]$. In view of Theorem 1.1, this automatically leads to sharper bounds for the inclusion/exclusion problem. De Wolf's argument, short and elegant, is based on quantum query complexity.

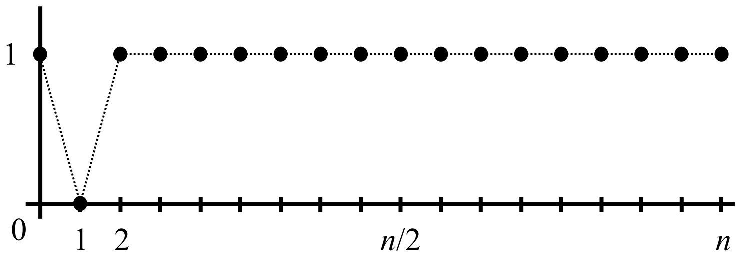
Proof outline. Our proof of Theorem 1.4 combines the work of Paturi and Kahn *et al.* with new uses of Chebyshev polynomials and interpolation techniques. We defer a more technical overview to Section 3. Although it may seem that Theorem 1.4 has a more direct proof, the simpler ideas that come to mind turn out to be useless. For example, an obvious approach is to start with Paturi's $\frac{1}{3}$-approximating polynomial p for the given predicate D and boost its accuracy by composing it with another polynomial, q. Let $\epsilon \in (0, 1/3)$ be the desired accuracy. For this approach to work, the polynomial q must send $[-\frac{1}{3}, \frac{1}{3}] \to [-\epsilon, \epsilon]$ and $[\frac{2}{3}, \frac{4}{3}] \to [1-\epsilon, 1+\epsilon]$. Up to translation/scaling, this is equivalent to requiring that q approximate the sign function within ϵ on the interval $[-1, -1+\alpha] \cup [1-\alpha, 1]$ for some constant $\alpha \in (0, 1)$. Eremenko and Yuditskii (2007) show that the least degree of such a polynomial q is $\Theta(\log(1/\epsilon))$. Taking $p(t)$ to be Paturi's approximating polynomial for the given predicate D, we see that the composition $q(p(t))$ has degree

$$\Theta\Big(\sqrt{n\big(\ell_0(D) + \ell_1(D)\big)}\ \log(1/\epsilon)\Big).$$

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

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

Another tempting strategy is to view a given predicate $D\colon \{0, 1, \ldots, n\} \to \{0, 1\}$ as a continuous (piecewise-linear) function on $[0, n]$ and then apply Jackson's fundamental theorems on uniform approximation (Jackson 1930). Unfortunately, the continuous approximation problem is hard even for the following simple predicate:



It is obvious that an ϵ-approximating polynomial for this continuous function yields (after translation) an ϵ-approximating polynomial of the same degree for $|x|$ on $[-1,1]$. In his classical work, Bernstein (1914) proves that the latter polynomial requires degree $\Omega(1/\epsilon)$. In particular, this approach is entirely useless once $\epsilon \leq \Theta(1/n)$. Yet the predicate in question has an approximator of degree $\tilde{\Theta}(\sqrt{n})$, 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 $\{0,1,\dots,n\}$, rather than the entire interval $[0,n]$.

1.3. 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.* (1994), is among the most realistic ones in computational learning theory. Designing efficient algorithms in this model is difficult even for the simplest concept classes. Nevertheless, progress on proving lower bounds has also been scarce. Some recent lower bounds are Klivans & Sherstov (2007); Tarui & Tsukiji (1999).

A summary of this model is as follows. Let $\mathcal{C}$ 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

$$\big(x^{(1)}, y^{(1)}\big)\,,\quad \big(x^{(2)}, y^{(2)}\big)\,,\quad \dots,\quad \big(x^{(m)}, y^{(m)}\big)\,,$$

independent and identically distributed according to λ. Let

$$\mathsf{opt} = \max_{f\in\mathcal{C}} \left\{ \mathop{\mathbf{P}}_{(x,y)\sim\lambda} \big[f(x)=y\big] \right\}$$

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

$$\mathop{\mathbf{P}}_{(x,y)\sim\lambda} \big[h(x)=y\big] \geq \mathsf{opt} - \epsilon\,,$$

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 hypothesis space is small and well-structured, making it possible to efficiently identify the best approximation to the observed examples.

The 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 $D\colon \{0,1,\dots,n\} \to \{0,1\}$ be a predicate and $f(x) = D(x_1 + \cdots + x_n)$. Let $\epsilon > 0$ be an arbitrary constant. Then there is a distribution λ on $\{0,1\}^n \times \{0,1\}$ such that*

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

and

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

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

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

$$\mathcal{C} = \Big\{ f(x_{i_1}, x_{i_2}, \dots, x_{i_n}) : \quad 1 \leq i_1 < i_2 < \cdots < i_n \leq 2n \Big\}.$$

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

We also show that the bound on the number of variables in Theorem 1.6 is optimal to within a multiplicative constant (see Theorem 5.5). Prior to our work, Tarui and Tsukiji (1999) obtained the special case of Theorem 1.6 for $f = \mathrm{OR}$.

2. Preliminaries

A *Boolean function* is a mapping $\{0,1\}^n \to \{0,1\}$. A *predicate* is a mapping $\{0,1,\dots,n\} \to \{0,1\}$. The notation $[n]$ stands for the set $\{1,2,\dots,n\}$. 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 $\phi\colon X \to \mathbb{R}$, we define

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

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

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

For $S \subseteq [n]$, define $\chi_S\colon \{0,1\}^n \to \{-1,+1\}$ 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 $f\colon \{0,1\}^n \to \mathbb{R}$ has a unique *Fourier representation*

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

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

2.1. Approximation vs. orthogonality Crucial to our work is a classical result from approximation theory (Ioffe & Tikhomirov 1968), recently used by the author in a different context (Sherstov 2008a). This result establishes an equivalence between approximation and orthogonality in Euclidean space. Relevant definitions and statements from Sherstov (2008a) follow.

Let X be a finite set. Consider $\mathbb{R}^X$, the linear space of functions $X \to \mathbb{R}$. For $\phi \in \mathbb{R}^X$, recall the notation $\|\phi\|_\infty = \max_{x\in X} |\phi(x)|$. Then $(\mathbb{R}^X, \|\cdot\|_\infty)$ is a real normed linear space.

Definition 2.1 (Least error). *For $f\colon X \to \mathbb{R}$ and $\Phi \subseteq \mathbb{R}^X$, define*

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

In words, $\epsilon^*(f,\Phi)$ is the least error in an approximation of f by a linear combination of functions in Φ. Since $\operatorname{span}(\Phi)$ has finite dimension, a best approximation to f out of $\operatorname{span}(\Phi)$ always exists (Rivlin 1981, 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). *Let X be a finite set, $f\colon X \to \mathbb{R}$, and $\Phi \subseteq \mathbb{R}^X$. The* modulus of orthogonality *of f with respect to Φ is:*

$$\gamma^*(f, \Phi) = \max_{\psi} \left\{ \sum_{x \in X} f(x)\psi(x) \right\}, \tag{2.3}$$

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

The maximization in (2.3) is over a nonempty compact set that contains $\psi = 0$. Also, the use of "max" instead of "sup" is legitimate because (2.3) maximizes a continuous function over a compact set. To summarize, the modulus of orthogonality is a well-defined nonnegative real number for every function $f\colon X \to \mathbb{R}$.

A key fact is that the least error and the modulus of orthogonality are always equal:

Theorem 2.4. *Let X be a finite set, $\Phi \subseteq \mathbb{R}^X$, and $f\colon X \to \mathbb{R}$. Then*

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

For the reader's convenience, we give a short and elementary proof of Theorem 2.4. In much greater generality, it is a classical result from functional analysis, due to Ioffe and Tikhomirov (1968). Proofs in the context of a Banach space are available in recent textbooks (DeVore & Lorentz 1993, p. 61, Thm. 1.3).

Proof of Theorem 2.4. The theorem holds trivially when $\operatorname{span}(\Phi) = \{0\}$. In the contrary case, let $\phi_1, \ldots, \phi_k$ be a basis for $\operatorname{span}(\Phi)$. Our first observation is that $\epsilon^*(f, \Phi)$ is the optimum of the following linear program in the variables $\epsilon, \alpha_1, \ldots, \alpha_k$:

$$\boxed{\begin{array}{lrl} \text{minimize:} & \epsilon & \\ \text{subject to:} & \left| f(x) - \sum_{i=1}^{k} \alpha_i \phi_i(x) \right| \leq \epsilon & \text{for each } x \in X\,, \\ & \alpha_i \in \mathbb{R} & \text{for each } i\,, \\ & \epsilon \geq 0. & \end{array}}$$

Standard manipulations reveal the dual:

$$\boxed{\begin{array}{lll}
\text{maximize:} & \displaystyle\sum_{x\in X} \beta_x f(x) & \\
\text{subject to:} & \displaystyle\sum_{x\in X} |\beta_x| \leq 1, & \\
 & \displaystyle\sum_{x\in X} \beta_x \phi_i(x) = 0 & \text{for each } i\,, \\
 & \beta_x \in \mathbb{R} & \text{for each } x \in X\,.
\end{array}}$$

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 $\phi_1,\ldots,\phi_k$ form a basis for $\operatorname{span}(\Phi)$, the optimum of the second program is by definition $\gamma^*(f,\Phi)$. □

2.2. Approximation by polynomials. Let $f\colon \{0,1\}^n \to \mathbb{R}$. As we saw above, 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 2.5 (Approximate degree of functions). *Let* $f\colon \{0,1\}^n \to \mathbb{R}$ *and* $\epsilon \geq 0$. *The* ϵ-approximate degree $\deg_\epsilon(f)$ of f *is the minimum integer* k, $0 \leq k \leq n$, *for which there exists* $\phi \in \operatorname{span}\{\chi_S : |S| \leq k\}$ *with*

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

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 $\epsilon \geq 0$. Second, $\deg_\epsilon(f)$ is not substantially affected by the choice of a constant $\epsilon \in (0,1/2)$. More precisely, we have:

Proposition 2.6 (Folklore). *Let* $f\colon \{0,1\}^n \to \{0,1\}$ *be arbitrary,* ϵ *a constant with* $0 < \epsilon < 1/2$. *Then*

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

Proof (Folklore). Assume that $\epsilon \leq 1/3$; the case $\epsilon \in (1/3, 1/2)$ has a closely analogous proof, and we omit it. Put $k = \deg_{1/3}(f)$ and fix $\phi \in \operatorname{span}\{\chi_S : |S| \leq k\}$ such that $\max_{x\in\{0,1\}^n} |f(x) - \phi(x)| \leq 1/3$. By basic approximation theory (Rivlin 1981, Cor. 1.4.1), there exists a univariate polynomial p of degree $O(1/\epsilon)$ that sends $[-\frac{1}{3},\frac{1}{3}] \to [-\epsilon,\epsilon]$ and $[\frac{2}{3},\frac{4}{3}] \to [1-\epsilon, 1+\epsilon]$. Then $p(\phi(x))$ is the sought approximator of f. □

In view of Proposition 2.6, 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 $f\colon \{0,1\}^n \to \{0,1\}$ whose value $f(x)$ 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 $\sigma\colon [n] \to [n]$. 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 $f(x) = D(x_1 + \cdots + x_n)$. To carry our discussion further, we extend the notion of approximation to predicates.

Definition 2.7. *For $D\colon \{0,1,\ldots,n\} \to \{0,1\}$, define the* ϵ-approximate degree $\deg_\epsilon(D)$ *to be the minimum degree of a univariate real polynomial p with*

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

Analyzing the approximate degree of predicates is a much simpler task. 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 follows from the *symmetrization argument* of Minsky and Papert (1988). Before we can state this theorem, we introduce some helpful notation.

Definition 2.8. *For $f\colon \{0,1\}^n \to \{0,1\}$ and $D\colon \{0,1,\ldots,n\} \to \{0,1\}$, define*

$$\begin{aligned} \epsilon^*\big(f, \{\chi_S : |S| \leq k\}\big) &= \min_{\phi \in \operatorname{span}\{\chi_S : |S| \leq k\}} \max_{x \in \{0,1\}^n} |f(x) - \phi(x)|\,, \\ \epsilon^*(D, P_k) &= \min_{p \in P_k} \max_{i=0,1,\ldots,n} |D(i) - p(i)|\,. \end{aligned}$$

Definition 2.8 merely instantiates the symbol $\epsilon^*(\phi, \Phi)$ from Section 2.1 to the special cases $\phi = f$ and $\phi = D$. We have:

Proposition 2.9 (Minsky & Papert 1988). *Let $f\colon \{0,1\}^n \to \{0,1\}$ be a symmetric Boolean function. Let D be the predicate with $f(x) \equiv D(x_1 + \cdots + x_n)$. Then*

$$\epsilon^*\big(f, \{\chi_S : |S| \leq k\}\big) = \epsilon^*(D, P_k) \quad \textit{for all} \quad k = 0, 1, \ldots, n\,.$$

In particular,

$$\deg_\epsilon(f) = \deg_\epsilon(D) \quad \textit{for all} \quad \epsilon \geq 0\,.$$

For a symmetric $f\colon \{0,1\}^n \to \{0,1\}$, put $\ell_0(f) = \ell_0(D)$ and $\ell_1(f) = \ell_1(D)$, where $\ell_0(D), \ell_1(D)$ were defined in the Introduction and D is the predicate for which $f(x) \equiv D(x_1 + \cdots + x_n)$. Using Proposition 2.9 and tools from approximation theory, Paturi (1992) gave an asymptotically tight estimate of $\deg_{1/3}(f)$ for every symmetric Boolean function:

THEOREM 2.10 (Paturi 1992). *Let $f\colon \{0,1\}^n \to \{0,1\}$ be a symmetric function. Then*

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

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

The purpose of this section is to establish Theorem 1.4. We prove the upper and lower bounds in this result separately, as Lemma 3.11 and Lemma 3.14, in the two subsections that follow.

3.1. Upper bound on the approximate degree. Our construction makes heavy use of Chebyshev polynomials, which is not surprising given their fundamental role in approximation. The other key ingredient 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 (Aspnes *et al.* 1994; Kahn *et al.* 1996) 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 stronger continuous setting, $[0, n]$. Kahn *et al.* (1996), who obtained the special case of Theorem 1.4 for $D = \text{OR}$, also used Chebyshev polynomials and interpolation, although in a simpler and much different way.

We start by recalling a few properties of Chebyshev polynomials, whose proofs can be found in any standard textbook on approximation theory, e.g., Cheney (1982); Rivlin (1981).

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

$$T_d(1) = 1 \tag{3.2}$$

$$|T_d(t)| \le 1 \qquad (-1 \le t \le 1) \tag{3.3}$$

$$T_d'(t) \ge d^2 \qquad (t \ge 1) \tag{3.4}$$

$$T_d(1+\delta) \ge \frac{1}{2} \cdot 2^{d\sqrt{2\delta}} \qquad (0 \le \delta \le 1/2) \tag{3.5}$$

$$2 \le T_{\lceil a \rceil}\left(1 + \frac{1}{a^2}\right) \le 7 \qquad (a \ge 1)\,. \tag{3.6}$$

At the heart of our construction is the following technical lemma, which gives an efficient method for approximating a given predicate D everywhere except in the vicinity of points where D changes value.

Lemma 3.7. *Let $\ell \ge 0$, $\Delta \ge 1$, and $d \ge 1$ be integers with $\ell + \Delta \le n/2$. 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

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

Proof. Define

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

One readily verifies the following properties of p_1:

$$\left\{\begin{array}{lll} p_1([0, n-\ell-\Delta]) \subseteq [-1,1] & & \text{by (3.3)}\,; \\ p_1([n-\ell-\Delta, n]) \subseteq [1,7] & & \text{by (3.2), (3.4), (3.6)}\,; \\ p_1'(t) \ge \dfrac{1}{\ell+\Delta} \ \text{ for } \ t \ge n-\ell-\Delta & & \text{by (3.4)}\,; \\ p_1(n-\ell) - p_1(n-\ell-\Delta) \ge \dfrac{\Delta}{\ell+\Delta} & & \text{by previous line}\,; \\ p_1(n-\ell+\Delta) - p_1(n-\ell) \ge \dfrac{\Delta}{\ell+\Delta} & & \text{likewise}\,. \end{array}\right. \tag{3.8}$$

Now consider the polynomial defined by

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

In view of (3.8), this new polynomial satisfies

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

and

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

Finally, define

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

Using (3.5) 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

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

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

Lemma 3.9. *Let $0 \le \ell \le n/2$. Then for any $\epsilon \le 1/3$,*

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

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

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

Our assumptions about ℓ and ϵ imply that $\ell + \Delta \ll n/2$, and thus Lemma 3.7 is applicable. Denote by $p(t)$ the polynomial constructed in Lemma 3.7. Define

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

We claim that the polynomial given by

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

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

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

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

Remark 3.10. *Applying Lemma 3.7 with $\Delta = 1$ and $d = \lceil \log(1/\epsilon)\rceil$ shows that*

$$\deg_\epsilon(\mathrm{EXACT}_\ell) = \deg_\epsilon(\mathrm{EXACT}_{n-\ell}) = O\bigl(\sqrt{n(\ell+1)}\log(1/\epsilon)\bigr)\,,$$

which slightly improves on the bound of Lemma 3.9 when $\epsilon \geq 1/n$. For simplicity and conciseness, however, we prefer to work with Lemma 3.9 for all ϵ.

We now prove the sought upper bound for an arbitrary predicate by repeatedly applying Lemma 3.9.

Lemma 3.11 (Upper bound on the approximate degree). *Let $D\colon \{0,1,\ldots,n\} \to \{0,1\}$. Then for any $\epsilon \leq 1/3$,*

$$\deg_\epsilon(D) \leq O\Bigl(\sqrt{n\bigl(\ell_0(D) + \ell_1(D)\bigr)}\,\log n + \sqrt{n\log(1/\epsilon)\log n}\Bigr)\,.$$

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 3.9. Put

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

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

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

where the third inequality follows by Lemma 3.9. □

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.* (1996) have already obtained a near-tight lower bound.

THEOREM 3.12 (Kahn *et al.* 1996, Thm. 2.1 and its proof). *For every polynomial p of degree $k = 0, 1, \dots, n-1$,*

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

Theorem 3.12 has the following immediate corollary:

COROLLARY 3.13. *Let ϵ be given with $2^{-\Theta(n\log n)} \le \epsilon \le 1/3$. Then*

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

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

LEMMA 3.14 (Lower bound on the approximate degree). *Let $D\colon \{0,1,\dots,n\} \to \{0,1\}$ be a nonconstant predicate. Then for each ϵ with $2^{-\Theta(n\log n)} \le \epsilon \le 1/3$,*

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

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

$$\deg_\epsilon(D) \geq \Omega\left(\sqrt{\frac{n\log(1/\epsilon)}{\log n}}\right). \tag{3.15}$$

Abbreviate $\ell = \ell_0(D)$ and assume w.l.o.g. that $\ell \geq 1$ (otherwise work with $\ell = \ell_1(D)$). We can additionally assume that $\ell \leq n/5$ since otherwise the claim follows trivially from Theorem 2.10. Consider the predicate EXACT_0 on $\lfloor n/5 \rfloor$ bits. By Corollary 3.13,

$$\deg_\epsilon(\text{EXACT}_0) \geq \Omega\left(\sqrt{\frac{n\log(1/\epsilon)}{\log n}}\right). \tag{3.16}$$

On the other hand,

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

so that

$$\deg_\epsilon(\text{EXACT}_0) \leq \deg_\epsilon(D)\,. \tag{3.17}$$

Equations (3.16) and (3.17) imply (3.15), 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 $f\colon \{0,1\}^n \to \{0,1\}$. Our discussion will revolve around the quantity $\delta^*(f,k)$, whose definition we restate from the Introduction.

Definition 4.1. *Let $f\colon \{0,1\}^n \to \{0,1\}$ and $0 \leq k \leq n$. Define*

$$\delta^*(f,k) = \frac{1}{2}\sup\left\{\mathop{\mathbf{P}}_{\mathcal{P}_1}\left[f(A_1,\dots,A_n)\right] - \mathop{\mathbf{P}}_{\mathcal{P}_2}\left[f(B_1,\dots,B_n)\right]\right\},$$

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

$$\mathop{\mathbf{P}}_{\mathcal{P}_1}\left[\bigcap_{i\in S} A_i\right] = \mathop{\mathbf{P}}_{\mathcal{P}_2}\left[\bigcap_{i\in S} B_i\right] \quad \textit{for} \quad |S| \leq k\,. \tag{4.2}$$

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

Definition 4.3 (Induced distribution). *Let $E_1, \ldots, E_n$ be events in a probability space $\mathcal{P}$. The* distribution on $\{0,1\}^n$ induced by $\mathcal{P}, E_1, \ldots, E_n$ *is defined as*

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

Proposition 4.4. *Let $E_1, \ldots, E_n$ be events in a probability space $\mathcal{P}$. Let μ be the distribution on $\{0,1\}^n$ induced by $\mathcal{P}, E_1, \ldots, E_n$. Then for every $g\colon \{0,1\}^n \to \{0,1\}$,*

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

Proof.

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

□

For a set $S \subseteq [n]$, define $\mathrm{AND}_S\colon \{0,1\}^n \to \{0,1\}$ by $\mathrm{AND}_S(x) = \bigwedge_{i\in S} x_i = \prod_{i\in S} x_i$. In particular, $\mathrm{AND}_\emptyset \equiv 1$.

Lemma 4.5. *Let $f\colon \{0,1\}^n \to \{0,1\}$ and $0 \le k \le n$. Then*

$$\delta^*(f,k) = \frac{1}{2}\max_{\alpha,\beta}\left\{\mathop{\mathbf{E}}_{x\sim\alpha}\left[f(x)\right] - \mathop{\mathbf{E}}_{x\sim\beta}\left[f(x)\right]\right\}, \tag{4.6}$$

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 $|S| \le k$.

Proof. Fix probability spaces $\mathcal{P}_1, \mathcal{P}_2$, events $A_1, \ldots, A_n$ in $\mathcal{P}_1$, and events $B_1, \ldots, B_n$ in $\mathcal{P}_2$, such that (4.2) holds. Let α and β be the distributions on $\{0,1\}^n$ induced by $\mathcal{P}_1, A_1, \ldots, A_n$ and $\mathcal{P}_2, B_1, \ldots, B_n$, respectively. Then by Proposition 4.4,

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

and

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

Letting δ stand for the right-hand side of (4.6), we conclude that $\delta^*(f,k) \le \delta$.

It remains to show that $\delta^*(f,k) \ge \delta$. Given a probability distribution μ on $\{0,1\}^n$, there is an obvious discrete probability space $\mathcal{P}$ and events $E_1,\dots,E_n$ in it that induce μ: simply let $\mathcal{P} = \{0,1\}^n$ 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 4.4) and show that $\delta^*(f,k) \ge \delta$. □

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

Theorem 1.1 (Restated from p. 221). *Let* $f\colon \{0,1\}^n \to \{0,1\}$ *be arbitrary and* $0 \le k \le n$. *Then*

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

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

Proof. In view of Theorem 2.4, it suffices to prove that

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

The remainder of the proof establishes this equality.

To rephrase Lemma 4.5,

$$\delta^*(f,k) = \frac{1}{2}\max_{\alpha,\beta}\left\{\sum_{x\in\{0,1\}^n}\big[\alpha(x)-\beta(x)\big]f(x)\right\}, \tag{4.7}$$

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

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

Let α,β be distributions for which the maximum is attained in (4.7). Setting $\psi = (\alpha-\beta)/2$, we see that $\sum_{x\in\{0,1\}^n}|\psi(x)| \le 1$ and thus $\delta^*(f,k) \le \gamma^*(f,\Phi)$.

It remains to show that $\gamma^*(f,\Phi) \le \delta^*(f,k)$. Suppose first that $\gamma^*(f,\Phi) = 0$. Since $\delta^*(f,k) \ge 0$ always, 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 (2.3). 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 (4.7) shows that $\gamma^*(f,\Phi) \leq \delta^*(f,k)$, as desired. □

Theorem 1.1, 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 1.1 holds for *every* function $f\colon \{0,1\}^n \to \{0,1\}$. For the special case of symmetric functions, we have already estimated the least error achievable by a polynomial of a given degree k. By combining these estimates with Theorem 1.1, we now prove the main result of the paper.

Theorem 4.8 (Restatement of Theorems 1.2 and 1.3). *Let* $f\colon \{0,1\}^n \to \{0,1\}$ *be a nonconstant symmetric function. Put* $\ell = \ell_0(f) + \ell_1(f)$. *Then*

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

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

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

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

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

$$\begin{aligned} \delta^*(f,k) &= \epsilon^*(f,\Phi) && \text{by Theorem 1.1}\\ &= \epsilon^*\big(f, \{\chi_S : |S| \leq k\}\big) && \text{since } \operatorname{span}(\Phi) = \operatorname{span}\{\chi_S : |S| \leq k\}\\ &= \epsilon^*(D, P_k) && \text{by Proposition 2.9}\,. \end{aligned} \tag{4.9}$$

By Theorem 2.10 and Lemmas 3.11 and 3.14,

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

In view of (4.9), this proves the claim regarding $\delta^*(f,k)$.

We now turn to the claim regarding $a_0, a_1, \ldots, a_k$. For $k \ge \Theta(\sqrt{n\ell}\log n)$, Lemma 3.11 gives an explicit univariate polynomial $p(t)$ of degree at most k such that

$$(4.10) \qquad |f(x) - p(x_1+\cdots+x_n)| \le 2^{-\Theta\left(\frac{k^2}{n\log n}\right)} \quad \text{for all} \quad x\in\{0,1\}^n\,.$$

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

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

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

$$\begin{aligned} \mathop{\mathbf{E}}_{x\sim\mu}\big[p(x_1+\cdots+x_n)\big] &= \mathop{\mathbf{E}}_{x\sim\mu}\left[\sum_{j=0}^{k} a_j \sum_{|S|=j}\prod_{i\in S} x_i\right] \\ &= \sum_{j=0}^{k} a_j \sum_{|S|=j} \mathop{\mathbf{E}}_{x\sim\mu}\left[\prod_{i\in S} x_i\right] \\ &= \sum_{j=0}^{k} a_j \sum_{|S|=j} \mathbf{P}\left[\bigcap_{i\in S} A_i\right] \qquad \text{by Proposition 4.4}\,, \end{aligned}$$

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 polynomial in n. Therefore, the quantity $\mathbf{E}_{x\sim\mu}[p(x_1+\cdots+x_n)]$ has the desired representation. It remains to verify that it approximates $\mathbf{P}[f(A_1,\ldots,A_n)]$ as claimed:

$$\begin{aligned} \left|\mathbf{P}\left[f(A_1,\ldots,A_n)\right] - \mathop{\mathbf{E}}_{x\sim\mu}\big[p(x_1+\cdots+x_n)\big]\right| &= \left|\mathop{\mathbf{E}}_{x\sim\mu}\big[f(x)-p(x_1+\cdots+x_n)\big]\right| \\ &\le 2^{-\Theta\left(\frac{k^2}{n\log n}\right)}, \end{aligned}$$

where the equality holds by Proposition 4.4 and the inequality by (4.10). □

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 subject of our study.

Definition 5.1. *Let $f\colon \{0,1\}^n \to \{0,1\}$ and $0 \le k \le n$. Define*

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

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

$$\mathop{\mathbf{P}}_{(x,y)\sim\lambda} \big[g(x)=y\big] = \frac{1}{2} \tag{5.2}$$

for every $g\colon \{0,1\}^n \to \{0,1\}$ 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

$$\Gamma^*\Big(f,\ \Theta\Big(\sqrt{n\big(\ell_0(f)+\ell_1(f)\big)}\Big)\Big) \ge 1-\epsilon$$

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 least error and modulus of orthogonality, quantities for which we have developed considerable intuition.

Lemma 5.3. *Let λ be a distribution on $\{0,1\}^n \times \{0,1\}$. Then for every $f\colon \{0,1\}^n \to \{0,1\}$,*

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

Proof.

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

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

Theorem 5.4. *Let $f\colon \{0,1\}^n \to \{0,1\}$ and $0 \le k \le n$. Then*

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

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

Proof. By Theorem 2.4, 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 (5.2) holds. Setting $g = 0$ gives:

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

Lemma 5.3 now leads to the following convenient characterization of $\Gamma^*(f,k)$:

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

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

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

for every function $g\colon \{0,1\}^n \to \{0,1\}$ that depends on k or fewer variables. With this new characterization, it is not difficult to show that $\Gamma^*(f,k) = \frac{1}{2} + \gamma^*(f,\Phi)$. The argument is closely analogous to the one in Theorem 1.1, and we do not repeat it here. □

Theorem 5.4 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. 227). *Let $D\colon \{0,1,\dots,n\} \to \{0,1\}$ be a predicate and $f(x) = D(x_1 + \cdots + x_n)$. Let $\epsilon > 0$ be an arbitrary constant. Then there is a distribution λ on $\{0,1\}^n \times \{0,1\}$ such that*

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

and

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

for every $g\colon \{0,1\}^n \to \{0,1\}$ 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 5.4, we need only show that $\epsilon^*(f,\Phi) \geq \frac{1}{2} - \epsilon$, where $\Phi = \{\chi_S : |S| \leq c\sqrt{n(\ell_0(f)+\ell_1(f))}\}$ for a suitably small constant c. But this is immediate from Proposition 2.6 and Theorem 2.10. □

Theorem 1.6 is best possible, as we now show.

Theorem 5.5 (On the tightness of Thm. 1.6). *Let $f\colon \{0,1\}^n \to \{0,1\}$ 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*

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

for every $g\colon \{0,1\}^n \to \{0,1\}$ 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

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

Proof. To rephrase the theorem, we need to show that $\Gamma^*(f,k) \leq 1-\epsilon$, where $k = C\sqrt{n(\ell_0(f)+\ell_1(f))}$. In view of Theorem 5.4, this is equivalent to the inequality $\epsilon^*(f,\ \{\chi_S : |S| \leq k\}) \leq \frac{1}{2}-\epsilon$. The latter is certainly true for a large enough constant C, by Proposition 2.6 and Theorem 2.10. □

Remark 5.6. *Let f be an arbitrary symmetric function. Theorem 5.5 tells us that if all hypotheses that depend on at most $k = \Theta(\sqrt{n(\ell_0(f)+\ell_1(f))})$ 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*

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

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 5.5, except in place of Paturi's result we would use our bounds on the approximate degree (Theorem 1.4) that work in the full range* $[2^{-n}, 1/3]$.

Acknowledgements

I would like to thank Sanjeev Arora, Anna Gál, Adam Klivans, Avi Wigderson, Ronald de Wolf, and the anonymous reviewers for their helpful feedback on this work. I am also thankful to Ronald for a stimulating discussion on how quantum arguments can contribute short and simple proofs in classical approximation theory. This research was supported by Adam Klivans' NSF CAREER Award and NSF Grant CCF-0728536.

References

S. Aaronson & Y. Shi (2004). Quantum lower bounds for the collision and the element distinctness problems. *J. ACM* **51**(4), 595–605.

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

R. Beals, H. Buhrman, R. Cleve, M. Mosca & R. de Wolf (2001). Quantum lower bounds by polynomials. *J. ACM* **48**(4), 778–797.

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

H. Buhrman, R. Cleve, R. de Wolf & C. Zalka (1999). Bounds for small-error and zero-error quantum algorithms. In *Proc. of the 40th Symposium on Foundations of Computer Science (FOCS)*, 358–368.

H. Buhrman, N. K. Vereshchagin & R. de Wolf (2007). On computation and communication with small bias. In *Proc. of the 22nd Conf. on Computational Complexity (CCC)*, 24–32.

E. W. Cheney (1982). *Introduction to Approximation Theory.* Chelsea Publishing, New York, 2nd edition.

R. A. DeVore & G. G. Lorentz (1993). *Constructive Approximation*, volume 303. Springer-Verlag, Berlin.

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

A. D. Ioffe & V. M. Tikhomirov (1968). Duality of convex functions and extremum problems. *Russ. Math. Surv.* **23**(6), 53–124.

D. Jackson (1930). *The Theory of Approximation.* Amer. Math. Soc., Colloquium Publications, Vol. XI, New York.

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

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

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

A. R. Klivans & R. A. Servedio (2004). Learning DNF in time $2^{\tilde{O}(n^{1/3})}$. *J. Comput. Syst. Sci.* **68**(2), 303–318.

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

N. Linial & N. Nisan (1990). Approximate inclusion-exclusion. *Combinatorica* **10**(4), 349–365.

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

R. Paturi (1992). On the degree of polynomials that approximate symmetric Boolean functions. In *Proc. of the 24th Symposium on Theory of Computing (STOC)*, 468–474.

A. A. Razborov (2002). Quantum communication complexity of symmetric predicates. *Izvestiya of the Russian Academy of Science, Mathematics* **67**, 145–159.

T. J. Rivlin (1981). *An Introduction to the Approximation of Functions.* Dover Publications, New York.

A. A. Sherstov (2008a). The pattern matrix method for lower bounds on quantum communication. In *Proc. of the 40th Symposium on Theory of Computing (STOC)*, 85–94.

A. A. Sherstov (2008b). The unbounded-error communication complexity of symmetric functions. In *Proc. of the 49th Symposium on Foundations of Computer Science (FOCS)*, 384–393.

Y. Shi & Y. Zhu (2007). Quantum communication complexity of block-composed functions. Available at arxiv.org/abs/0710.0095.

J. Tarui & T. Tsukiji (1999). Learning DNF by approximating inclusion-exclusion formulae. In *Proc. of the 14th Conf. on Computational Complexity (CCC)*, 215–221.

R. de Wolf (2008). A note on quantum algorithms and the minimal degree of ϵ-error polynomials for symmetric functions. *Quantum Information and Computation* **8**(10), 943–950.

Manuscript received 24 August 2008

Alexander A. Sherstov
Department of Computer Sciences
The University of Texas at Austin
1 University Station C0500
Austin, TX 78712-0233, USA
sherstov@cs.utexas.edu
http://www.cs.utexas.edu/~sherstov