Improved Learning of AC^0 Functions

Merrick L. Furst
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
furst@theory.cs.cmu.edu

Jeffrey C. Jackson
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
jcj@cs.cmu.edu

Sean W. Smith
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
smith@theory.cs.cmu.edu

Abstract

Two extensions of the Linial, Mansour, Nisan AC^0 learning algorithm are presented. The LMN method works when input examples are drawn uniformly. The new algorithms improve on theirs by performing well when given inputs drawn from unknown, mutually independent distributions. A variant of the one of the algorithms is conjectured to work in an even broader setting.

1 INTRODUCTION

Linial, Mansour, and Nisan [LMN89] introduced the use of the Fourier transform to accomplish Boolean function learning. They showed that AC^0 functions are well-characterized by their low frequency Fourier spectra and gave an algorithm which approximates such functions reasonably well from uniformly chosen examples. While the class AC^0 is provably weak in that it does not contain modular counting functions, from a learning theory point of view it is fairly rich. For example, AC^0 contains polynomial-size DNF and addition. Thus the LMN learning procedure is potentially powerful, but the restriction that their algorithm be given examples drawn according to a uniform distribution is particularly limiting.

A further limitation of the LMN algorithm is its running time. Valiant's [Val84] learning requirements are widely accepted as a baseline characterization of feasible learning. They include that a learning algorithm should be distribution independent and run in polynomial time. The LMN algorithm runs in quasi-polynomial ($O(2^{\text{poly}\log n})$) time.

In this paper we develop two extensions of the LMN learning algorithm which produce good approximating functions when samples are drawn according to unknown distributions which assign values to the input variables independently. Call such a distribution *mutually independent* since it is the joint probability distribution corresponding to a set of mutually independent random variables [Fel57]. The running times of the new algorithms are dependent on the distribution—the farther the distribution is from uniform the higher the bound—and, as is the case for the LMN algorithm, the time bounds are quasi-polynomial in n.

A variant of one of our algorithms gives a more general learning method which we conjecture produces reasonable approximations of AC^0 functions for a broader class of input distributions. A brief outline of the general algorithm is presented along with a conjecture about a class of distributions for which it might perform well.

Our two learning methods differ in several ways. The *direct* algorithm is very similar to the LMN algorithm and depends on a substantial generalization of their theory. This algorithm is straightforward, but our bound on its running time is very sensitive to the probability distribution on the inputs. The *indirect* algorithm, discovered independently by Umesh Vazirani [Vaz], is more complicated but also relatively simple to analyze. Its time bound is only mildly affected by changes in distribution, but for distributions not too far from uniform it is likely greater than the direct bound. We suggest a possible hybrid of the two methods which may have a better running time than either method alone for certain distributions.

The outline of the paper is as follows. We begin with definitions and proceed to discuss the key idea behind our direct extension: we use an appropriate change of basis for the space of n-bit functions. Next, we prove that under this change AC^0 functions continue to exhibit the low-order spectral property which the LMN result capitalizes on. After giving an overview of the direct algorithm and analyzing its running time, we discuss the indirect approach and compare it with the first. Finally, we indicate some directions for further research.

2 DEFINITIONS AND NOTATION

All sets are subsets of $\{1, \ldots, n\}$, where as usual n represents the number of variables in the function to be learned. Capital letters denote set variables unless otherwise noted. The complement of a set X is indicated by $\bar{X}$, although we abuse the concept of complementation somewhat: in cases where X is specified to be a subset of some other set S, $\bar{X} = S - X$, otherwise $\bar{X} = \{1, \ldots, n\} - X$. Strings of $0/1$ variables are referred to by barred lower case letters (e.g. $\bar{x}$) which may be superscripted to indicate one of a sequence of strings (e.g. $\bar{x}^j$). x_i refers to the ith variable in a string $\bar{x}$. Barred constants (e.g. $\bar{0}$, $\bar{*}$) indicate strings of the given value with length implied by context.

Unless otherwise specified, all functions are assumed to have as domain the set of strings $\{0, 1\}^n$. The range of Boolean functions will sometimes be $\{0,1\}$, particularly when we are dealing with circuit models, but will usually be $\{1, -1\}$ for reasons that should become clear subsequently. Frequently we will write sets as arguments where strings would be expected, e.g. $f(X)$ rather than $f(\bar{x})$ for f a function on $\{0, 1\}^n$. In such cases $f(X)$ is a shorthand for $f(c(X))$ where $c(X)$ is a characteristic function defined by $c_i(X) = 0$ if $i \in X$ and 1 otherwise. Note that the sense of this function is opposite the natural one in which 0 represents set absence.

An AC^0 *function* is a Boolean function which can be computed by a family of acyclic circuits (one circuit for each number n of inputs) consisting of AND and OR gates plus negations only on inputs and satisfying two properties:

- The number of gates in each circuit (its *size*) is bounded by a fixed polynomial in n.
- The maximum number of gates between an input and the output (the circuit *depth*) is a fixed constant.

A *random restriction* $\rho_{p,q}$ is a function which given input $\bar{x}$ maps x_i to $*$ with fixed probability p and assigns 0's and 1's to the other variables according to a probability distribution q. If $\bar{x}$ represents the input to a function f then $\rho_{p,q}$ induces another function $f\lceil\rho$ which has variables corresponding to the stars and has the other variables of f fixed to 0 or 1. This is a generalization of the original definition [FSS81] in which q is the uniform distribution. The subscripts of ρ are generally dropped and their values understood from context.

The function obtained by setting a certain subset S of the variables of f to the values indicated by the characteristic function of a subset $X \subseteq S$ is denoted by $f\lceil S \leftarrow X$ or, when S is implied by context, simply $f\lceil X$. For example, if $S = \{1, 3\}$ and $X = \{3\}$ then $f\lceil X$ is the function f with variable x_1 set to 1 and x_3 to 0.

We will use several parameters of the probability distribution q throughout the sequel. We define $\mu_i = \Pr[x_i = 1]$, where the probability is with respect to q. Another parameter which we will use frequently is

$$\beta = \max_i(1/\mu_i, 1/(1 - \mu_i)).$$

We assume that this value is finite, since infinite β implies some variable is actually a constant and can be ignored by the learning procedure.

It is convenient to define a set-based notation for probabilities also. For example, if $X = \{2\}$ and it has been specified that $X \subseteq \{1, 2\}$ then we will write $q(X)$ for $q(x_1 = 1 \wedge x_2 = 0)$. In general, if X is specified to be a subset of some set S then $q(X)$ represents the marginal probability that the variables indicated by S take on the values specified by $c(X)$, and if S is not specified then $q(X)$ is just $q(c(X))$. Thus for mutually independent q,

$$q(X) = \prod_{i \in X}(1 - \mu_i) \prod_{i \in \bar{X}} \mu_i.$$

3 AN ORTHONORMAL BASIS FOR BOOLEAN FUNCTIONS SAMPLED UNDER MUTUALLY INDEPENDENT DISTRIBUTIONS

3.1 RATIONALE

Given some element v of a vector space and a basis for this space, a discrete Fourier transform expresses v as the coefficients of the linear combination of basis vectors representing v. We are interested in learning Boolean functions on n bits, which can be represented as Boolean vectors of length 2^n. Linial et al. used as the basis for this space the characters of the group $\mathbf{Z}_2^n$ in $\mathbf{R}$. The characters are given by the 2^n functions

$$\chi_A(X) = -1^{|A \cap X|}.$$

Each χ_A is simply a polynomial of degree $|A|$, and $\{\chi_A \mid |A| \leq k\}$ spans the space of polynomials of degree not more than k. With an inner product defined by

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

and norm $\|f\| = \sqrt{\langle f, f\rangle}$ the characters form an orthonormal basis for the space of real-valued functions on n bits. The Fourier coefficients of a function f with respect to this basis are simply $\hat{f}_A = \langle f, \chi_A\rangle$, the projections of f on the basis vectors. Given a sufficient number m of uniformly selected examples $(\bar{x}^j, f(\bar{x}^j))$, $\sum_{j=1}^m f(\bar{x}^j)\chi_A(\bar{x}^j)/m$ is a good estimate of $\hat{f}_A$ [LMN89].

What happens if the examples are chosen according to some nonuniform distribution q? If we blindly applied the LMN learning algorithm we would actually be calculating an approximation to $\sum_{\bar{x} \in \{0,1\}^n} f(\bar{x})\chi_A(\bar{x})q(\bar{x})$ for each A. That is, we would be calculating the expected value of the product $f\chi_A$ with respect to q rather than with respect to the uniform distribution. This leads to the following observation: if we modify the definition of the inner product to be with respect to the distribution q on the inputs and modify the basis to be orthonormal under this new inner product then the estimated coefficients should on average be close to the appropriate values in this new basis.

More precisely, define the inner product to be

$$\langle f, g\rangle_q = \sum_{\bar{x} \in \{0,1\}^n} f(\bar{x})g(\bar{x})q(\bar{x})$$

and define the norm to be[1]

$$\|f\|_q = \sqrt{\langle f, f\rangle_q}.$$

Let the basis vectors ψ_A be defined by orthonormalizing the χ_A with respect to this inner product (such a basis will

[1] Typically, a subscript indicates that the norm is to be considered one of the Hölder p-norms. We are using the subscript to mean something different here.

be referred to as *orthonormal in the q-norm* or *orthonormal with respect to q*). Then we expect that if a large enough number of samples m are drawn according to q,

$$\tilde{f}_A = \sum_{j=1}^{m} f(\bar{x}^j)\psi_A(\bar{x}^j)/m$$

is a good approximation to $\hat{f}_A$, the projection of f onto ψ_A.

The main result of Linial et al. [LMN89] is that, for any AC^0 function f, the Fourier coefficients of f with respect to the χ_A's become small for large $|A|$. Thus the LMN learning algorithm consists simply of estimating the "low-order" coefficients. We show that essentially the same property of AC^0 functions also holds for the coefficients of a particular ψ basis orthonormal with respect to a mutually independent q. Thus, as with LMN learning, we can obtain a good approximation to an AC^0 function f by estimating low-order coefficients of f relative to the transformed basis. Our learning procedure differs in that the estimated coefficients are with respect to a basis which must also be estimated.

It will be convenient to have a name for a basis which is orthonormal with respect to a mutually independent distribution as opposed to an arbitrary distribution. We will refer to such a basis as a *ϕ basis* and reserve ψ for bases orthonormal with respect to an arbitrary q. From now on a Fourier coefficient $\hat{f}_A$ will be assumed to be the coefficient of the basis vector ϕ_A unless otherwise noted.

3.2 PROPERTIES OF THE ϕ BASIS

Let σ_i be the standard deviation of the ith variable x_i when samples are selected according to q and note that μ_i as previously defined represents the mean. Let $z_i = (x_i - \mu_i)/\sigma_i$; that is, z_i is the normalized variable corresponding to x_i. Then, due to the mutual independence of q, one possible ϕ basis is given by Bahadur [Bah61]:

$$\phi_A = \prod_{i \in A} z_i.$$

This basis will be referred to as *the* ϕ basis; it is the basis which would be obtained by a Gram-Schmidt orthonormalization (with respect to the q-norm) of the χ basis performed in order of increasing $|A|$.

The ϕ basis has a number of properties which make our generalization of the LMN result possible. First, due to the nature of Gram-Schmidt orthonormalization, $\{\phi_A \mid |A| \leq k\}$ spans the same space as $\{\chi_A \mid |A| \leq k\}$, so linear combinations of such ϕ's are simply polynomials of degree not more than k. Also, it follows immediately from the above representation of ϕ_A that for all A, S, and $X \subseteq \bar{S}$, $Y \subseteq S$,

$$\phi_A(X \cup Y) = \phi_{A \cap \bar{S}}(X)\phi_{A \cap S}(Y).$$

Likewise it follows that for any S and $A, B \subseteq S$,

$$\sum_{X \subseteq S} \phi_A(X)\phi_B(X)\, q(X) = \begin{cases} 1 & \text{if } A = B \\ 0 & \text{otherwise.} \end{cases}$$

Another useful property is that for all A and B,

$$\phi_A(B)\sqrt{q(B)} = \phi_B(A)\sqrt{q(A)}.$$

This follows from the above representation of ϕ_A after noting that $\sigma_i = \sqrt{\mu_i(1-\mu_i)}$. Finally, Parseval's identity gives

$$\|f\|_q^2 = \sum_A \hat{f}_A{}^2$$

.

4 THE DROPOFF LEMMA

As noted above, Linial et al. have shown that the sum of squares of coefficients of high-degree terms (the *high-order power spectrum*) of AC^0 functions becomes exponentially small as order increases when the coefficients are relative to the χ basis. In this section we show that this also holds for coefficients relative to the ϕ basis. We do this by generalizing the series of lemmas used in their proof.

Essentially, we prove that the following facts hold for Fourier coefficients relative to the ϕ basis:

1. Random restrictions of AC^0 functions have small minterms and maxterms with high probability as long as the distribution function q is mutually independent.
2. All the high-order Fourier coefficients of a function with small minterms and maxterms are zero.
3. The coefficients of an AC^0 function are closely related to the coefficients of its restrictions.
4. Probabilistic arguments can be used to tie the above facts together and show that the high-order coefficients of an AC^0 function must be small.

We present the proof of the Dropoff Lemma in this order.

4.1 RANDOM RESTRICTIONS

The linchpin of the Linial et al. result is Hastad's Switching Lemma [Has86]. This lemma states that when restriction $\rho_{p,q}$ of a suitable CNF function has uniform q then with high probability the minterms of the restricted function are small; a similar statement about a DNF function and maxterms follows immediately. Thus such a randomly restricted CNF formula can with high probability be rewritten as a DNF formula having a small number of variables in every term. We generalize this to a lemma which holds for any mutually independent q. For uniform q, $\beta = 2$ and Lemma 1 reduces to a restatement of Hastad's Lemma.

Lemma 1 *Let f be a CNF formula with at most t variables in any clause and let random restriction $\rho_{p,q}$ have mutually independent q with parameters μ_i and β as defined previously. Then*

$$\Pr[f\lceil\rho \text{ has a minterm of size} > s] \leq (\beta pt/\ln \Phi_g)^s$$

where $\Phi_g = (1+\sqrt{5})/2$, the golden ratio.

Proof Sketch: Our proof involves a lengthy reworking of the Boppana and Sipser proof of Hastad's lemma [BS90]. We give here only the details of our extension to a key inequality in their proof; the remainder of our proof is straightforward.

Let C be an OR of variables, none of which are negated, and let Y be a subset of these variables. Let q be a mutually independent distribution and let $\rho_{p,q}$ be a random restriction defined on the variables in C. Then we show that

$$\Pr\left[\rho(Y) = \bar{*} \;\middle|\; C\lceil\rho \not\equiv 1\right] \leq (\beta p)^{|Y|}.$$

Indeed, because q is mutually independent,

$$\Pr\left[\rho(Y) = \bar{*} \;\middle|\; C\lceil\rho \not\equiv 1\right] = \prod_{y_i \in Y} \frac{\Pr[\rho(y_i) = *]}{\Pr[\rho(y_i) \neq 1]}.$$

From the definition of β it follows that for all i, $\mu_i \leq (\beta - 1)/\beta$, so

$$\forall i \;\; \Pr[\rho(y_i) \neq 1] \geq 1 - \frac{\beta - 1}{\beta}(1 - p).$$

Thus, noting that $\beta \geq 2$,

$$\begin{aligned}\frac{\Pr[\rho(y_i) = *]}{\Pr[\rho(y_i) \neq 1]} &\leq \frac{p}{1 - \frac{\beta-1}{\beta}(1-p)} \\ &\leq \beta p.\end{aligned}$$

□

Because it is also true that for all i, $(1 - \mu_i) \leq (\beta - 1)/\beta$, an analog to Lemma 1 can easily be proved for the case where C in the above proof is an AND. This gives:

Lemma 2 *Let f be a DNF formula with at most t variables in any term and let random restriction $\rho_{p,q}$ have mutually independent q. Then*

$$\Pr[f\lceil\rho \textit{ has a maxterm of size} > s] \leq (\beta pt/\ln \Phi_g)^s.$$

Finally we can state the principal result of this section, which is obtained by successively applying Lemmas 1 and 2 to the lowest levels of a circuit (cf. [BS90, Theorem 3.6]):

Lemma 3 *Let f be a Boolean function computed by a circuit of size M and depth d, and let $\rho_{p,q}$ have mutually independent probability distribution q with parameter β as defined previously. Then*

$$\Pr[f\lceil\rho \textit{ has a minterm or maxterm of size} > s] \leq M2^{-s}$$

where $\Pr[*] = (2\beta s/\ln \Phi_g)^{-d}$.

4.2 SMALL MINTERMS AND MAXTERMS MEAN VANISHING HIGH-ORDER SPECTRUM

Here we begin to relate the above results to the Fourier spectrum of AC^0 functions. We show that if a function has only small minterms and maxterms then its high-order Fourier coefficients—even with respect to certain ψ bases—vanish.

Lemma 4 *Let q be any (not necessarily mutually independent) probability distribution, let ψ_A be a basis orthonormal in the q-norm such that $\{\psi_A \mid |A| \leq k\}$ spans the same space as $\{\chi_A \mid |A| \leq k\}$, and let $\hat{f}_A$ be the Fourier coefficient of ψ_A. Then if all of the minterms and maxterms of a Boolean function f have size at most $\sqrt{k}$, $\hat{f}_A = 0$ for all $|A| > k$.*

Proof: It is not hard to see that if the $\sqrt{k}$ condition is met then f can be computed by a decision tree of depth no more than k [BI87]. Linial et al. [LMN89] further show that the Fourier coefficients of the χ basis for such an f satisfy the lemma. This means that f is in the space spanned by $\{\chi_A \mid |A| \leq k\}$. Since by definition the low-order ψ_A span the same space, our lemma follows immediately. □

Finally, note that the ϕ basis meets the criteria of the lemma due to the nature of the Gram-Schmidt orthonormalization process which defines it.

4.3 RELATING COEFFICIENTS OF A FUNCTION AND ITS RESTRICTIONS

Putting together the results thus far, we know that with high probability the random restrictions of an AC^0 function have zero-valued high-order Fourier coefficients. Now we show a key relationship between the coefficients of arbitrary functions and their restrictions when the coefficients are relative to a ϕ basis.

We begin with a lemma which allows a rewriting of the definition of a Fourier coefficient and follow with the coefficient-relating result.

Lemma 5 *Let f be any real-valued n-bit function, S any set, and ϕ_A orthonormal with respect to mutually independent q. Then for any subset A,*

$$\hat{f}_A = \sum_{X \subseteq \bar{S}} \widehat{f\lceil X}_{A \cap S} \phi_{A \cap \bar{S}}(X)\, q(X)$$

Proof:

$$\begin{aligned}\hat{f}_A &= \sum_Z f(Z)\phi_A(Z)\, q(Z) \\ &= \sum_{X \subseteq \bar{S}, Y \subseteq S} f\lceil X(Y)\phi_{A\cap\bar{S}}(X)\phi_{A\cap S}(Y) q(X) q(Y) \\ &= \sum_{X \subseteq \bar{S}} \left[\sum_{Y \subseteq S} f\lceil X(Y)\phi_{A\cap S}(Y)\, q(Y)\right] \cdot \\ &\qquad \phi_{A\cap\bar{S}}(X)\, q(X)\end{aligned}$$

□

Lemma 6 *Let f, S and ϕ_A be as above. Then for any $B \subseteq S$,*

$$\sum_{C \subseteq \bar{S}} \hat{f}^2_{B \cup C} = \sum_{X \subseteq \bar{S}} \widehat{f\lceil X}_B^2\, q(X)$$

Proof:

$$\sum_{C\subseteq\bar{S}} \hat{f}_{B\cup C}^2 = \sum_{C\subseteq\bar{S}} \left[\sum_{X\subseteq\bar{S}} \widehat{f\lceil X}_B \phi_C(X)\, q(X)\right]^2$$
$$= \sum_{C,X,Y\subseteq\bar{S}} \widehat{f\lceil X}_B \widehat{f\lceil Y}_B \phi_C(X)\phi_C(Y)\, q(X)q(Y)$$
$$= \sum_{X,Y} \widehat{f\lceil X}_B \widehat{f\lceil Y}_B \sqrt{q(X)q(Y)} \cdot \sum_{C} \phi_C(X)\phi_C(Y)\sqrt{q(X)q(Y)}$$
$$= \sum_{X,Y} \widehat{f\lceil X}_B \widehat{f\lceil Y}_B \sqrt{q(X)q(Y)} \cdot \sum_{C} \phi_X(C)\phi_Y(C)\, q(C)$$

from which the Lemma follows by the orthonormality of ϕ on subsets. □

4.4 BOUNDING HIGH ORDER POWER SPECTRUM

We now use a series of probabilistic arguments to tie the above lemmas together into the desired result. Although the proofs are very similar to those in [LMN89], we include them for completeness. We begin with an easily proved bound on the high-order spectrum of any function.

Lemma 7 *Let f be any real-valued n-bit function, p a real in $(0,1)$, k an integer, and p and k chosen such that $pk > 8$. Let $\hat{f}_A$ be a Fourier coefficient with respect to any orthonormal basis. Then*

$$\sum_{|A|>k} \hat{f}_A{}^2 \leq 2E_S\left[\sum_{|A\cap S|>pk/2} \hat{f}_A{}^2\right]$$

where S is a subset chosen at random such that each variable appears in it independently with probability p.

Proof: Clearly $E_S[|A\cap S|] \geq pk$ for every $|A| > k$. Thus Chernoff bounds tell us that

$$\Pr_S[|A\cap S| \leq pk/2] \leq e^{-pk/8} \leq \tfrac{1}{2}.$$

Thus for each A at least half of the S's will satisfy $|A\cap S| > pk/2$. □

To this point we have not been much concerned with the form of the output of the function f; in fact, several of the previous lemmas hold for any real-valued function. For the sequel, we specify that f is Boolean and, furthermore, maps to either 1 or -1. It is then the case by Parseval's identity that

$$\sum_{A\subseteq\{1,\ldots,n\}} \hat{f}_A{}^2 = 1$$

and thus that the sum of any subset of the squared coefficients of such a Boolean function is bounded by unity. With this fact in hand we can prove the following bound on the summation of the previous lemma:

Lemma 8 *Let f be a function from $\{0,1\}^n$ to $\{1,-1\}$, S any set, t an integer in $[0,n]$, and q a mutually independent distribution defining the basis for $\hat{f}_A$. Then*

$$\sum_{|A\cap S|>t} \hat{f}_A{}^2 \leq \Pr_X[f\lceil X \textit{ has a minterm or maxterm of size} > \sqrt{t}\,]$$

where X is an assignment to the variables in $\bar{S}$ chosen according to q.

Proof:

$$\sum_{|A\cap S|>t} \hat{f}_A{}^2 = \sum_{B\subseteq S,|B|>t} \sum_{C\subseteq\bar{S}} \hat{f}_{B\cup C}{}^2$$
$$= \sum_{B\subseteq S,|B|>t} E_{X\subseteq\bar{S}}[\widehat{f\lceil X}_B^2]$$
$$= E_X\left[\sum_{|B|>t} \widehat{f\lceil X}_B^2\right]$$

where the second line follows from Lemma 6 and expectation is with respect to q. Now since the terms in the expectation are never larger than unity, it is clearly bounded above by the probability that $\sum_{|B|>t} \widehat{f\lceil X}_B^2$ is nonzero. But then application of Lemma 4 completes the proof. □

We can now prove the main result.

Lemma 9 (Dropoff Lemma) *Let f be a function from $\{0,1\}^n$ to $\{1,-1\}$ computed by a circuit of depth d and size M, and let k be any integer. Then*

$$\sum_{|A|>k} \hat{f}_A{}^2 \leq M2^{1-k^{\frac{1}{d+2}}/5\beta}.$$

Proof: From the previous two lemmas,

$$\sum_{|A|>k} \hat{f}_A{}^2 \leq 2E_S \Pr_X[f\lceil\bar{S}\leftarrow X \text{ has a minterm or maxterm of size} > \sqrt{pk/2}\,].$$

But this latter value is just

$$2\Pr_\rho[f\lceil\rho \text{ has a minterm or maxterm of size} > \sqrt{pk/2}\,]$$

and by Lemma 3 is bounded above by $2M2^{-\sqrt{pk/2}}$ as long as $p \leq (2\beta\sqrt{pk/2}/\ln\Phi_g)^{-d}$. Some simplifying arithmetic gives the result. □

5 DIRECT LEARNING

As alluded to earlier, our direct learning algorithm, like that of Linial et al., depends on the spectral property of AC^0 functions proved above. That is, since the high-order Fourier coefficients relative to a ϕ basis are small, we need only estimate low-order coefficients in order to derive a close approximation to the desired function. As shown below, the linear combination of the low-order ϕ basis vectors defined by these coefficients is a function which is close to the true function in the sense that the norm of the difference between the functions is small. Furthermore, the sign of this approximating function will with high probability match the true function, where the the probability is relative to the input distribution q.

Actually, since we assume that only input/output pairs are given, the distribution q must also be estimated and hence the function is learned relative to an approximate basis. In spite of this we are able to prove a bound on the running time of our algorithm which is similar to the bound on LMN learning. More specifically, let $f \triangle \tilde{f}$ denote the probability that $f(\bar{x}) \neq \tilde{f}(\bar{x})$ when the input $\bar{x}$'s are drawn according to a mutually independent probability distribution q. Our algorithm, given parameters ϵ and δ, produces an $\tilde{f}$ such that $\Pr[f \triangle \tilde{f} > \epsilon] \leq \delta$ when the algorithm is given access to a sufficient number of examples of the true function f drawn according to q. The algorithm runs in time and number of examples quasi-polynomial in n and $1/\epsilon$, exponential in the parameter β of q, and polynomial in $\log(1/\delta)$.

In the sections that follow we first give the algorithm and then prove the bound on its running time.

5.1 THE DIRECT ALGORITHM

Algorithm 1 *Given m examples $(\bar{x}^j, f(\bar{x}^j))$ of a function $f : \{0,1\}^n \to \{1,-1\}$ and an integer $k \leq n$, determine $\tilde{f}$ as follows:*

1. *Compute $\mu'_i = \frac{1}{m}\sum_{j=1}^{m} x_i^j$ for $1 \leq i \leq n$.*
2. *Define $z'_i = (x_i - \mu'_i)/\sqrt{\mu'_i(1-\mu'_i)}$.*
3. *Define $\phi'_A = \prod_{i \in A} z'_i$.*
4. *Compute $\tilde{f}'_A = \frac{1}{m}\sum_{j=1}^{m} f(\bar{x}^j)\phi'_A(\bar{x}^j)$ for $|A| \leq k$ and 0 otherwise. If $|\tilde{f}'_A| > 1$ let $\tilde{f}'_A = \text{sign}(\tilde{f}'_A)$, where* sign *is the obvious function with range $\{-1,0,1\}$.*
5. *Define $g(\bar{x}) = \sum \tilde{f}'_A \phi'_A(\bar{x})$.*
6. *Define $\tilde{f}(\bar{x}) = \text{sign}(g(\bar{x}))$.*

We intend primes ($'$) to indicate values that are based on an estimated probability distribution rather than the true one. A twiddle ($\tilde{}$) indicates that the value includes other estimates. When a Fourier coefficient is based on an estimated distribution it is written with the twiddle replacing the usual hat ($\hat{}$). Thus $\tilde{f}'_A$ rather than $\tilde{\hat{f}}'_A$ is used to represent an estimate of the A^{th} Fourier coefficient of f relative to the estimated basis ϕ'_A.

Notice that the restriction on the magnitude of $\tilde{f}'_A$ can only bring this estimated coefficient closer to the true coefficient in the ϕ' basis, since all of the coefficients of a Boolean function must be no larger than 1 in magnitude. The restriction also plays a helpful role in the lemmas to follow.

5.2 BOUNDS FOR ϵ/δ LEARNING

Here we derive upper bounds on the values of m and k required for the above algorithm to achieve specified error bounds on an input distribution with a given β. Our first step is to generalize a lemma of Linial et al. [LMN89] to the case of arbitrary distributions q.

Lemma 10 *Let f be a function mapping $\{0,1\}^n$ to $\{1,-1\}$ and g an arbitrary function on the same domain. Let q be an arbitrary probability distribution on $\{0,1\}^n$, let the Fourier coefficients be relative to the basis ψ_A defined by q, and let probabilities be with respect to q. Then*

$$\Pr[f(\bar{x}) \neq \text{sign}(g(\bar{x}))] \leq \sum (\hat{f}_A - \hat{g}_A)^2.$$

Proof: $\Pr[f(\bar{x}) \neq \text{sign}(g(\bar{x}))] \leq \Pr[|f(\bar{x}) - g(\bar{x})| > 1] \leq E[(f(\bar{x}) - g(\bar{x}))^2] \leq \sum (f(\bar{x}) - g(\bar{x}))^2 q(\bar{x}) = \|f - g\|_q^2$ and the lemma follows from Parseval's identity and the linearity of the Fourier transform. □

Now for the remainder of this section let q be a mutually independent distribution on the inputs to the algorithm and let unprimed Fourier coefficients be with respect to the ϕ basis defined by q. Then in the notation of our learning algorithm, Lemma 10 says that $f \triangle \tilde{f} \leq \sum (\hat{f}_A - \hat{g}_A)^2$. So our goal becomes finding an m and k such that with probability at least $1 - \delta$ the algorithm produces a g satisfying $\sum (\hat{f}_A - \hat{g}_A)^2 \leq \epsilon$. While the details of this calculation are a bit messy, the basic idea is not. Allocate half of the ϵ error to each of two jobs: taking care of the error in the coefficients corresponding to sets smaller than k and larger than k. The Dropoff Lemma is used to bound the error in the latter and will also fix k. Chernoff bound arguments will give the value of m needed to bound the error due to estimating the low-order coefficients and basis functions.

Because $\tilde{f}'_A = 0$ for all $|A| > k$ and $\{\phi'_A \mid |A| \leq k\}$ spans the same space as the corresponding set of ϕ_A, $\hat{g}_A$ must also vanish for all $|A| > k$. Thus the following lemma, which follows immediately from the Dropoff Lemma, gives the required bound on $\sum_{|A|>k} (\hat{f}_A - \hat{g}_A)^2$.

Lemma 11 *Let f be a Boolean function with range $\{1,-1\}$ with corresponding $\{0,1\}$- valued function computed by a circuit of depth d and size M. Then*

$$k \geq \left[5\beta \log_2\left(\frac{4M}{\epsilon}\right)\right]^{d+2} \Rightarrow \sum_{|A|>k} \hat{f}_A{}^2 \leq \frac{\epsilon}{2}.$$

The bound on the error in low-order coefficients is a bit more involved. There are really two sources of error: the estimate of the basis functions and the estimate of the coefficients relative to these functions. It seems simplest to consider these sources of error separately. First, define $\tilde{f}_A = \sum f(\bar{x}^j)\phi_A(\bar{x}^j)/m$ and, as in the definition of $\tilde{f}'_A$, restrict the magnitude of this value to 1. That is, $\tilde{f}_A$ represents the coefficient which would be estimated if the true basis function was known. Since there are at most n^k low-order coefficients for $n > 1$, an m satisfying

$$\Pr\left[\exists\, |A| \le k \text{ s.t. } |\hat{f}_A - \tilde{f}_A| > \sqrt{\frac{\epsilon}{8n^k}}\right] \le \frac{\delta}{2} \qquad (1)$$

and

$$\Pr\left[\exists\, |A| \le k \text{ s.t. } |\tilde{f}_A - \hat{g}_A| > \sqrt{\frac{\epsilon}{8n^k}}\right] \le \frac{\delta}{2}, \qquad (2)$$

guarantees that $\Pr[\sum_{|A|\le k}(\hat{f}_A - \hat{g}_A)^2) > \epsilon/2] \le \delta$. Here $|\hat{f}_A - \tilde{f}_A|$ represents the error due to estimating the Fourier coefficients given perfect knowledge of the input distribution, $|\tilde{f}_A - \hat{g}_A|$ the error due to estimating the distribution. An inequality of Hoeffding [Hoe63] will be very useful for finding the required m.

Lemma 12 (Hoeffding) *Let X_i be independent random variables all with mean $E[X]$ such that for all i, $a \le X_i \le b$. Then for any $\lambda > 0$,*

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} X_i - E[X]\right| \ge \lambda\right] \le 2e^{-2\lambda^2 m/(b-a)^2}.$$

For the moment we remove the unity restriction on $\tilde{f}_A$'s magnitude. Then $E[\tilde{f}_A]$ is $\hat{f}_A$, and using Lemma 12 to find an m satisfying (1) requires only that the bounds on $\tilde{f}_A$ be determined. By its definition, the magnitude of $\tilde{f}_A$ is bounded by $\phi_{\max} = \max|\phi_A(\bar{x})|$, where the maximum is over all possible $|A| \le k$ and $\bar{x}$. It is not hard to show that for all i, $|z_i| \le \sqrt{\beta - 1}$, so $\phi_{\max} \le (\beta-1)^{k/2}$. Then by Lemma 12, $m \ge 16n^k(\beta-1)^k\epsilon^{-1}\ln(4n^k/\delta)$ guarantees that for any given $|A| \le k$, $\Pr[|\hat{f}_A - \tilde{f}_A| > \sqrt{\epsilon/8n^k}] \le \delta/2n^k$. Hence such an m satisfies (1). Note finally that restricting the magnitude of $\tilde{f}_A$ can only improve the likelihood that $\tilde{f}_A$ is sufficiently near $\hat{f}_A$, since $\hat{f}_A \in [-1, 1]$.

Finding an m satisfying (2) is more involved. First, rewrite $\hat{g}_A$ as $\sum_{|B|\le k}\tilde{f}'_B\langle\phi_A, \phi'_B\rangle$ (all inner products in this section are with respect to q) and let $\Delta\phi = \max_{|A|\le k,\bar{x}}|\phi_A(\bar{x}) - \phi'_A(\bar{x})|$. Then some algebra shows that for $|A|, |B| \le k$, $|\langle\phi_A, \phi'_B\rangle - \langle\phi_A, \phi_B\rangle| \le \phi_{\max}\Delta\phi$. It follows that for all $|A| \le k$

$$\begin{aligned}
|\tilde{f}_A - \hat{g}_A| &\le |\tilde{f}_A - \tilde{f}'_A\langle\phi_A, \phi'_A\rangle| + \\
&\quad |\sum_{|B|\le k, B\ne A}\tilde{f}'_B\langle\phi_A, \phi'_B\rangle| \\
&\le |\tilde{f}_A - \tilde{f}'_A| + |\tilde{f}'_A(\langle\phi_A, \phi_A\rangle - \langle\phi_A, \phi'_A\rangle)| \\
&\quad +(n^k - 1)\phi_{\max}\Delta\phi \\
&\le \Delta\phi + n^k\phi_{\max}\Delta\phi.
\end{aligned}$$

Actually, careful consideration of the cases A empty and A nonempty shows that the bound can be tightened to simply $n^k\phi_{\max}\Delta\phi$.

Clearly the magnitude of $\Delta\phi$ depends on the error in the estimate of μ'_i. Intuitively, by driving the relative error in μ'_i small we drive $\Delta\phi$ small. Thus define c_μ as the smallest value such that for all i, $|\mu_i - \mu'_i| \le c_\mu \min(\mu_i, 1-\mu_i)$; that is, c_μ is the least upper bound on the relative error. Then by considering the cases $\mu_i \le \frac{1}{2}$ and $\mu_i \ge \frac{1}{2}$ it can be shown that for all i, both the ratios $|z'_i/z_i|$ and $|z_i/z'_i|$ are bounded above by $2c_\mu + 1$ as long as $c_\mu \le \frac{1}{2}$. Hence the largest possible value of the ratio ϕ'_A/ϕ_A for $|A| \le k$ is $(2c_\mu+1)^k$, and thus $\Delta\phi \le [(2c_\mu+1)^k - 1]\phi_{\max}$. Finally, use of the identity $x \le 1/j \Rightarrow (x+1)^j - 1 \le 2jx$ for j a positive integer shows that if $2c_\mu \le 1/k$ then $\Delta\phi \le 4c_\mu k\phi_{\max}$.

Thus $c_\mu \le \sqrt{\epsilon/128k^2n^{3k}(\beta-1)^{2k}}$ implies that for all $|A| \le k$, $|\tilde{f}_A - \hat{g}_A| \le \sqrt{\epsilon/8n^k}$. Let c_d represent the desired bound on c_μ. Then since $1/\beta = \min(\mu_i, 1-\mu_i)$ it follows that m such that $\Pr[\exists\, i \text{ s.t. } |\mu_i - \mu'_i| > c_d/\beta] \le \delta/2$ satisfies (2). Noting that $E[\mu'_i] = \mu_i$, that $0 \le \mu'_i \le 1$, and that there are n different μ'_i, application of Lemma 12 shows that the required m is bounded by $64k^2n^{3k}\beta^{2k+2}\epsilon^{-1}\ln(4n/\delta)$. This is always larger than the value required for (1), so we have

Theorem 1 *For any positive ϵ and δ and any mutually independent distribution on the inputs, Algorithm 1 produces a function $\tilde{f}$ with the property that* $\Pr[f\triangle\tilde{f} > \epsilon] \le \delta$ *whenever*

$$k \ge \left[5\beta\log_2\left(\frac{4M}{\epsilon}\right)\right]^{d+2},$$

$$m \ge 64k^2n^{3k}\beta^{2k+2}\frac{1}{\epsilon}\ln\left(\frac{4n}{\delta}\right).$$

Thus for fixed ϵ, δ, and β, the algorithm requires $O(2^{\text{poly}\log n})$ examples to adequately approximate AC^0 functions.

The bounds for LMN learning on uniform distributions are similar. The LMN k bound is polylogarithmic in M and $1/\epsilon$ ($O(\log^{d+3}(M/\epsilon))$), and the m bound is quasi-polynomial in n and $1/\epsilon$ ($O(n^{2k}/\epsilon)$) and logarithmic in $1/\delta$.

6 INDIRECT LEARNING

6.1 OVERVIEW

Our approach to learning AC^0 functions sampled according to mutually independent distributions results in a straightforward deterministic algorithm, but the analysis is quite involved. We, and independently Umesh Vazirani [Vaz], have noticed a clever randomized approach which would be somewhat more difficult to implement but admits a simpler analysis. Observe first that for any given value μ in $(0, 1)$ it is easy to construct a small fixed-depth circuit which, given inputs drawn uniformly, produces 1's with probability approximately μ. Thus for any given mutually independent probability distribution on n-bit strings a set of n disjoint circuits can be constructed which given uniform inputs will produce as output each n-bit string with approximately the desired probability. Conversely, a randomized inverse of each of these small circuits can be constructed such that mutually independent inputs to the inverses will produce a nearly uniform output.

With this background, the indirect *uniform construction* approach falls out naturally. We are given a set of input/output pairs $(\bar{x}, f(\bar{x}))$ where the $\bar{x}$'s are drawn according to a mutually independent distribution q. The unknown function f is computed by some AC^0 circuit F. Also, there exists a set of disjoint AC^0 circuits C_i which, given uniform $\bar{y}$'s, produce as output $\bar{x}$'s with distribution close to q. Call the randomized inverses of these circuits C_i^{-1}. Then there is another AC^0 circuit G consisting of the obvious composition of the C_i's and F such that, if G computes function g then for all $\bar{x}$, $g(C_i^{-1}(\bar{x})) = f(\bar{x})$. Since the $C_i^{-1}(\bar{x})$ are almost uniformly distributed, a variant of LMN learning can be used to obtain a good approximation to g and therefore indirectly to f.

6.2 ANALYSIS OF UNIFORM CONSTRUCTION

Clearly if the circuits C_i produce exactly the desired probability distribution q on their outputs then the LMN theory applies immediately to uniform construction, since the C_i^{-1} will produce an exactly uniform distribution for learning g. Considering the forms of the bounds on k and m for LMN learning, the analysis for this case reduces to determining the size and depth of the circuit G and the length of its input. This in turn reduces to determining how long the string $\bar{y}$ generated by the C_i^{-1} is and ascertaining the size and depth of the C_i.

Although many possible forms of the C_i could be considered, we will assume that simple depth 2 DNF circuits are used in order to minimize the increase in depth of G over F. With such circuits any μ of the form $\sum_{j=1}^{l} a_j 2^{-j}$, where $a_j \in \{0,1\}$, can be easily constructed using at most l variables and $l+1$ gates. The idea is to create a circuit with one AND for each j such that $a_j = 1$, to have that AND produce 1's with probability 2^{-j}, and to insure that at most one AND produces a 1 on each input. Such a circuit is easy to construct; for example, the circuit computing $x_1 \vee (\bar{x}_1 \wedge x_2) \vee (\bar{x}_1 \wedge \bar{x}_2 \wedge x_3 \wedge x_4)$ has four variables, one OR, two AND's, and produces 1's with probability $13/16$.

Thus in the case of exact representation of q by depth 2 C_i there must be some value l such that for each variable x_i, $x_i = \sum_{j=1}^{l} a_j 2^{-j}$. Therefore G has at most nl variables and $n(l+1)$ more gates than F and has depth $d+2$.

Of course, even if q is known exactly it may not be desirable or even possible to represent it exactly with the C_i. In this case the LMN theory must be extended a bit to cover the case of nearly uniform distributions. Call a distribution r on $\bar{x}$ *γ-uniform* if for all $\bar{x}$, $|U(\bar{x}) - r(\bar{x})| \leq \gamma/2^n$, where U is the uniform distribution $U(\bar{x}) = 2^{-n}$. Then the probabilities of the occurrence of some event with respect to these distributions can be related in a simple way. In particular, for any Boolean f and approximating g,

$$\begin{aligned} & \Pr_r[f(\bar{x}) \neq \mathrm{sign}(g(\bar{x}))] \\ \leq \; & \Pr_U[f(\bar{x}) \neq \mathrm{sign}(g(\bar{x}))](1+\gamma). \end{aligned}$$

Also, the expected value of a Fourier coefficient computed using examples drawn from a γ-uniform rather than truly uniform distribution will differ from the true coefficient by no more than γ. Finally, as would be expected, the convergence of the C_i^{-1} to a uniform distribution as variables are added is extremely rapid once each C_i has at least $\log \beta$ variables, that is, once the probability of a 1 for each C_i is in the vicinity of the appropriate value.

Putting these facts together with an analysis similar to that used in proving Theorem 1 shows that if each of the C_i has a polylogarithmic number of variables and a similar number of gates then the distribution r induced by the C_i^{-1} will be near enough uniform for an adequate g to be learned. Specifically, let $l = 2\max[2k^2, \log_2 \beta] + 2$ be the number of variables input to each C_i, where k satisfies the LMN bound modified to reflect the increase of 2 in depth d and the logarithmic dependence of circuit size M on l. Then the uniform construction method satisfies specified ϵ/δ bounds as long as the number of examples is at least $64(nl)^k 2^{2l} \epsilon^{-1} \ln(4n^k/\delta)$.

7 COMPARISON OF APPROACHES

The primary advantage of our direct approach to learning AC^0 functions is probably its potential application to non-independent distributions. While it is not at all clear how a technique like uniform construction could be used on an arbitrary distribution, our direct algorithm offers the hope of wider applicability, as discussed in the next section. Also, the ϕ basis and its properties have proved useful in extending another learning result from uniform to mutually independent distributions [Bel91].

Another significant area of difference between the approaches is the use of randomness. Uniform construction is a random algorithm in terms of both learning and the function learned, while our direct algorithm and the function learned are deterministic.

In terms of expected running times, both algorithms are quasi-polynomial. Uniform construction would seem to have a distinct advantage when β for the underlying distribution is large. On the other hand, for moderate β the direct approach should be faster due to the increase in circuit depth which uniform construction must contend with.

An interesting implementation possibility is a hybrid of the two approaches. Variables with means far from uniform would be handled via uniform construction methods—be expanded by an appropriate C_i^{-1}—and those closer to uniform would be unchanged. Then the direct learning algorithm rather than LMN would be applied to the resulting strings, which would now be nearly independent rather than nearly uniform. If only a few variables are far from uniform then increasing the depth of the circuit at these few points might not affect overall circuit depth. Thus the hybrid approach potentially avoids the primary sources of run time blowup in the individual methods.

8 OPEN QUESTIONS

An averaging argument added to a fundamental idea of Valiant and Vazirani [VV85] shows that for every AC^0 function f and every distribution q on the inputs there is a low-degree polynomial which is a close approximation to f with respect to q [BRS90, Tar91]. Unfortunately, this is only an existence proof which does not give rise immediately to a computationally feasible algorithm for finding such polynomials. The obvious question is to find such an algorithm.

Given an unknown distribution q and examples of a function f drawn according to q we can use something like an approximate Gram-Schmidt process to orthogonalize, relative to q, a low-degree basis. We can then estimate the low-degree coefficients of function f. We conjecture that for many natural distributions this will be a good approximation. For what distributions is this true? It is not true for all distributions; Smolensky [Smo] has produced a counterexample.

It is natural to define AC^0 distributions to be those obtained in the following way. Transform uniformly drawn input variables $\bar{y}$ to new variables $\bar{x}$ via an AC^0 circuit C. The induced distribution on the $\bar{x}$ is called AC^0. Does the above variant work for AC^0 distributions?

Acknowledgements

Thanks to Alan Frieze for a helpful suggestion regarding the error analysis in Theorem 1. This research was supported in part by an unrestricted grant from AT&T. The third author received support from an ONR Graduate Fellowship and NSF Grant CCR-8858087.

References

[Bah61] R. R. Bahadur. A representation of the joint distribution of responses to n dichotomous items. In Herbert Solomon, editor, *Studies in Item Analysis and Prediction*, pages 158–168. Stanford University Press, Stanford, California, 1961.

[Bel91] Mihir Bellare. The spectral norm of finite functions. Technical Report TR-495, MIT-LCS, February 1991.

[BI87] M. Blum and R. Impagliazzo. Generic oracles and oracle classes. In *28th Annual Symposium on Foundations of Computer Science*, pages 118–126, 1987.

[BRS90] Richard Beigel, Nick Reingold, and Daniel Spielman. The perceptron strikes back. Technical Report YALEU/DCS/TR-813, Yale University, 1990.

[BS90] Ravi B. Boppana and Michael Sipser. The complexity of finite functions. In *Handbook of Theoretical Computer Science*, volume A. MIT Press/Elsevier, 1990.

[Fel57] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. John Wiley & Sons, second edition, 1957.

[FSS81] M. Furst, J. Saxe, and M. Sipser. Parity, circuits, and the polynomial time hierarchy. In *22nd Annual Symposium on Foundations of Computer Science*, pages 260–270, 1981.

[Has86] J. Hastad. *Computational Limitations for Small Depth Circuits*. PhD thesis, MIT Press, 1986.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal*, 58:13–30, 1963.

[LMN89] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, Fourier transform, and learnability. In *30th Annual Symposium on Foundations of Computer Science*, pages 574–579, 1989.

[Smo] Roman Smolensky. Personal communication.

[Tar91] J. Tarui. Randomized polynomials, threshold circuits, and the polynomial hierarchy. In C. Choffrut and M. Jantzen, editors, *8th Annual Symposium on Theoretical Aspects of Computer Science Proceedings*, pages 238–250, 1991.

[Val84] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.

[Vaz] Umesh Vazirani. Personal communication.

[VV85] L. G. Valiant and V. V. Vazirani. NP is as easy as detecting unique solutions. In *Proceedings of the Seventeenth Annual ACM Symposium on Theory of Computing*, pages 458–463, 1985.