PSEUDORANDOM GENERATORS AND LEARNING ALGORITHMS FOR AC^0

MEERA SITHARAM

Abstract. - For any AC^0 function f of n bits, there is a polynomial p such that any $p(\log n)$-wise decomposable distribution "fools" f. In other words, f cannot distinguish between the pseudorandom strings in the distribution and truly random strings. The polynomial p depends only on the size and depth of the circuit computing f.

This subsumes and extends the class of distributions that were previously known to fool AC^0 functions, and partially answers an open question posed by Linial and Nisan in 1990, as to whether every polylog-wise independent distribution fools AC^0 functions.

- Each polylog-wise decomposable distribution serves as a fixed training set of examples for learning (approximately interpolating) all AC^0 functions computed by circuits of some fixed depth and size. Furthermore, small, natural distributions (training sets) exist that yield deterministic learning algorithms that run in time $O(2^{\text{polylog } n})$ for AC^0 functions, where the degree of the polylog depends on the size and depth of the circuit to be learnt.

This improves on the randomized algorithms with the same time complexity given, for example, by Linial, Mansour and Nisan in 1989, where the examples for the training set are picked randomly from specific distributions.

Key words. Circuit complexity; Lower bounds; Fourier transforms.
Subject classifications. 68Q15, 68Q99.

1. Introduction

Basic questions about the class AC^0 have remained open despite its apparent simplicity and extensive study (e.g. [8], [17], [21], [9], [12], [4], survey [3]). In particular, the following natural extension of Hastad's result [9] has not been settled. Consider the set of n bit strings x that satisfy a single parity equation $\langle x, c\rangle = 0(1)$, where $\langle , \rangle$ is the inner product over $\mathbb{F}_2^n$ and c is a constant vector

of polylog n non-zero bits. This set is exactly the support of a single *parity* function involving the polylog n non-zero bits in c; and, by increasing the degree of the polylog, this set fools, i.e, provides good pseudorandom strings, for AC^0 circuits of each depth and size. However, the following natural question remains open.

- $\star$ Consider the product or *and* of several *parity* functions that involve independent vectors all of whose linear combinations have Hamming weight at least polylog n. Here independence is with respect to $\mathbb{F}_2^n$. The support of this product is the set of n bit strings that satisfy several independent parity equations of polylog n bits. Does every such product fool AC^0 circuits of some fixed depth and size? And viceversa: for each AC^0 function f, does such a product (with a sufficiently high polylog degree) fool f?

A more general version of question $\star$ has also appeared in the literature [13] phrased as: do polylog-wise independent distributions fool AC^0 functions? (These distributions include sets that are not linear subspaces of $\mathbb{F}_2^n$).

The two formulations of the question are related by [15] which exhibits a class of *parity* products whose supports provide linear, polylog-wise independent distributions for fooling AC^0 circuits of each fixed depth and size. Here, a distribution is linear if it is closed under sums in $\mathbb{F}_2^n$. The linearity and small size of the distribution given by [15] make it possible to simulate bounded error probabilistic AC^0 circuits in $DTIME[O(2^{\text{polylog } n})]$. A positive answer to $\star$ improves on this by allowing any small, linear, polylog-wise independent distribution to be used as a pseudorandom generator for such a deterministic simulation.

In this paper, we partially answer $\star$: we show that for any AC^0 function f of n bits, there is a polynomial p such that any $p(\log n)$-wise decomposable distribution fools f, and p depends only on the size and depth of the circuit computing f.

A formal definition of polylog-wise decomposability is given in Section 3. To explain intuitively, we first note that a linear, polylog-wise independent distribution is not only the support of an *and* of parities involving independent vectors $c_1, \ldots, c_{n-k}$ as described in $\star$, but also is a subspace of $\mathbb{F}_2^n$ of dimension k that is orthogonal to those vectors. This is because every vector x in the distribution satisfies $\langle x, c_1\rangle = 0, \ldots, \langle x, c_{n-k}\rangle = 0$. Consider any coset $D + a$ of a polylog-wise independent distribution D. (This coset, by definition, contains all vectors $x + a$ for $x \in D$). The coset $D + a$ can thus be characterized as containing $k+1$ independent vectors that are non-orthogonal to exactly (a fixed) half of $c_1, \ldots, c_{n-k}$. In other words, since any linear combination of $c_1, \ldots, c_{n-k}$

is of size at least polylog n, the coset $D + a$ contains $k + 1$ independent vectors that force every non-orthogonal vector to have size at least polylog n; but k could be larger than polylog n. A polylog-wise decomposable distribution, however, stipulates the existence of some polylog n vectors in $D + a$ such that every vector that is not orthogonal to them has size at least polylog n.

In particular, polylog-wise decomposable distributions turn out to be polylog-wise independent. Furthermore, all distributions that are known to fool AC^0 functions, including that given in [15], are properly included in the class of polylog-wise decomposable distributions.

The connection between the learnability of AC^0 functions, and their spectral norms was shown in the innovative paper [12], which gives a randomized algorithm for learning (approximately interpolating) AC^0 functions in time $O(2^{\text{polylog } n})$, where, again, the degree of the polylog depends only on the size and depth of the circuit to be learnt. The examples for the training set in this algorithm are chosen randomly from the uniform distribution. Since then, other randomized algorithms have appeared for learning AC^0 functions with examples picked randomly from other distributions [7], and for learning functions computed by certain types of decision trees [11], [1]. The analyses of all of these learning algorithms estimate appropriate spectral norms (in some cases, the spectrum is itself defined in the sense of distributions), and then rely on applying Chernoff bounds to the difference between the norms of the function f to be learnt, and the function defined only by the randomly sampled values of f.

We show a connection between pseudorandom generators for AC^0 functions, and uniform training sets for *deterministically* learning AC^0 functions. In particular, using basic properties of spectra of Boolean functions, Shannon's sampling theorem [18], and specific properties of AC^0 spectral norms given in [12], we show that every polylog-wise decomposable distribution not only serves as a source of pseudorandom strings for AC^0 functions as discussed earlier, but also as a uniform training set of examples, or a deterministic set of sample points, for all AC^0 functions computed by circuits of some fixed depth and size. Such distributions, when sufficiently small, provide *deterministic* algorithms that learn AC^0 functions in time $O(2^{\text{polylog } n})$, where the degree of the polylog depends only on the size and depth of the AC^0 circuit to be learnt. Since small polylog-wise decomposable distributions exist, such algorithms exist as well.

The paper is organized as follows. Section 2 clarifies notational conventions and provides some basic background. Section 3 defines polylog-wise decomposable distributions, relates them to polylog-wise independent distributions, and

shows that polylog-wise decomposable distributions fool AC^0 functions. Section 4 concludes by showing the relationship between pseudorandom generators and deterministic learning algorithms for AC^0 functions.

2. Background and conventions

Unless otherwise specified, all n-tuples are elements of the finite vector space $\mathbb{F}_2^n$ with the inner product $\langle , \rangle$. The number of non-zero entries in x is denoted $|x|$, the n-tuple $(a, \ldots, a)$ is denoted (a^n), the complement of a bit a is denoted $\bar{a}$, and the tuple $(1^n) + x$ (the bit-wise complement of x) is denoted $\bar{x}$. Two vectors x and y are said to be disjoint if there are no coordinates (indices) where both x and y are 1; and log is $\log_2$.

All functions map from $\mathbb{F}_2^n$ to $\mathbb{C}$, the field of complex numbers, and Boolean functions are viewed as complex functions that happen to be $\{0,1\}$-valued to facilitate use of the Fourier transform. See [5] for a simple and entertaining introduction to Fourier transforms of functions over $\mathbb{Z}_2^n$ and other finite groups. The 2^n characters $(-1)^{\langle ,x\rangle}$ for $x \in \mathbb{Z}_2^n$ form an orthonormal basis for the space of functions from $\mathbb{F}_2^n$ to $\mathbb{C}$, so that any function f can be written as a linear combination of these characters: $f(y) = \sum_{x \in \mathbb{F}_2^n} (-1)^{\langle y,x\rangle} a_x$. The coefficients $a_x \in \mathbb{C}$ define a function $\hat{f}$, which is called the Fourier transform of f; i.e, $\hat{f}(x) =_{def} a_x$. The function $\hat{f}$ can be obtained as $\hat{f}(x) = 1/2^n \sum_{y \in \mathbb{F}_2^n} (-1)^{\langle x,y\rangle} f(y)$. The inverse Fourier transform of a function g is denoted $g^\vee(y) =_{def} \sum_{x \in \mathbb{F}_2^n} (-1)^{\langle y,x\rangle} g(x)$. Thus, $(\hat{f})^\vee = f$.

The support of a function f, denoted by the uppercase F, is that subset of $\mathbb{F}_2^n$ over which f is non-zero. The support of $\hat{f}$ is denoted $\hat{F}$. For any set $B = \{b^1, \ldots, b^e\}$ of independent, non-zero vectors in $\mathbb{F}_2^n$, the set $\mathcal{O}(B)$ is the coset of all odd combinations (sum of an odd number) of vectors in B, and the set $\mathcal{E}(B)$ is the subspace of all even combinations (sum of an even number) of vectors in B.

FACT 2.1. *For functions f and g the following hold.*

(i) Parseval's identity holds:

$$1/2^n \sum_u f(u)^2 \;=\; \sum_u \hat{f}(u)^2;$$

(ii) The convolution identity holds:

$$\widehat{fg}(u) = (\hat{f} \circ \hat{g})(u) =_{def} \sum_w \hat{f}(w)\hat{g}(u-w);$$

and

$$(1/2^n)(\widehat{f \circ g})(u) = (\hat{f}\hat{g})(u).$$

(iii) The value of the transform at (0^n) *is the expected value of the function:*

$$\hat{f}(0^n) = (1/2^n)\sum_u f(u).$$

(iv) If f *is Boolean, then* $\hat{f}(0^n) = (1/2^n)|F|$, *and* $f(0^n) = \sum_u \hat{f}(u)$.

(v) If $range(f) = \{0,1\}$, $range(g) = \{-1,+1\}$, *and* $f = (g+1)/2$, *then*

$$\hat{f}(0^n) = 1/2\ (\hat{g}(0^n) + 1),$$

and

$$\forall u : |u| > 0 \quad \hat{f}(u) = 1/2\ \hat{g}(u).$$

(vi) If f *is Boolean, then*

$$\begin{array}{rl} & F \text{ is a subspace of } \mathbb{F}_2^n \\ \iff & \hat{f} \text{ is constant on its support and} \\ & \hat{F} \text{ is a subspace of } \mathbb{F}_2^n \\ \Rightarrow & |F||\hat{F}| = 2^n. \end{array}$$

The last statement is equivalent to saying that the dimension of $\hat{F}$ *and the dimension of* F *total* n.

PROOF. We give a short proof of (iv). The reader is referred to [5] for the remainder. We show the forward direction of the double implication. The reverse direction is symmetric. F being a subspace of $\mathbb{F}_2^n$ of dimension k, is equivalent to saying that there are $n-k$ independent vectors $b^1, \ldots, b^{n-k}$ in $\mathbb{F}_2^n$ such that each $y \in F$ satisfies $\langle y, b^i \rangle = 0$ for $1 \le i \le n-k$. (It follows that for every $x \in span\{b^1, \ldots, b^{n-k}\}$ and for every $y \in F$, $\langle y, x \rangle = 0$; and for every $x \notin span\{b^1, \ldots, b^{n-k}\}$, $\langle y, x \rangle = 0$ for exactly half the y's in F). Thus for every $x \in span\{b^1, \ldots, b^{n-k}\}$,

$$\hat{f}(x) = 1/2^n \sum_{y \in \mathbb{F}_2^n} f(y)(-1)^{\langle y,x \rangle} = 1/2^n \sum_{y \in \mathbb{F}_2^n} f(y) = 1/2^{n-k};$$

and for every for every $x \notin span\{b^1, \ldots, b^{n-k}\}$, $\hat{f}(x) = 0$. Thus $\hat{f}$ is constant (i.e, $1/2^{n-k}$) on its support, $\hat{F} = span\{b^1, \ldots, b^{n-k}\}$, and since $\hat{F}$ has dimension

$n-k$, while F was assumed to have dimension k, the last consequence follows. □

Finally, by circuits we mean rooted dags whose vertices are of unbounded fan-in $\wedge$ and $\vee$ gates, to which the input bits may be fed in negated at the leaves. The following are known spectral properties of Boolean functions computed by circuits.

THEOREM 2.2. *Suppose that a Boolean function f of n variables can be computed by a circuit of size M and depth d. Then*

(i) ([12]) $\sum_{v:|v|\geq e} \hat{f}(v)^2 \leq M2^{\frac{-e^{(1/d+3)}}{4}}$.

(ii) ([9], rephrased)

$$\frac{1}{2^n} \mid \sum_{\substack{x\in\mathbb{F}_2^n \\ |x| \ even}} f(x) - \sum_{\substack{x\in\mathbb{F}_2^n \\ |x| \ odd}} f(x)| = |\hat{f}(1^n)|$$

$$\leq M2^{-n^{1/d}}.$$

This easily implies that the parity function of n bits cannot be computed by AC^0 circuits (of constant depth and size polynomial in n).

3. Decomposable distributions

The main result of this section, Theorem 3.8, shows that if f is computed by an AC^0 circuit of some fixed size and depth, and if $S \subseteq \mathbb{F}_2^n$ is a polylog-wise decomposable distribution, then the sum

$$|\frac{1}{|S|}\sum_{x\in S} f(x) - \frac{1}{2^n}\sum_{x\in\mathbb{F}_2^n} f(x)|$$

has a small upper bound. This shows that S fools f. The next two lemmas are required for the proof of the theorem, and extend Hastad's result (Theorem 2.2 (ii)) to bound general alternating sums involving functions computed by constant depth circuits.

LEMMA 3.1. *Let f over $\mathbb{F}_2^n$ be computed by a circuit of depth d and size M. Let $B = \{b^1, \ldots, b^e\}$ be a set of independent and mutually disjoint vectors. Then for any $y \in \mathbb{F}_2^n$,*

$$\frac{1}{2^e}|\sum_{x\in\mathcal{E}(B)} f(x+y) - \sum_{x\in\mathcal{O}(B)} f(x+y)| \leq M2^{-e^{1/d}}.$$

PROOF. Let g be the function of e variables obtained from f as follows.

$$g(x) =_{def} f(\sum_{x_i \neq 0} b^i) = f(\sum_{i=1}^{e} x_i b^i).$$

In other words, restrict f to only those variables that appear as non-zero bits in some vector of B, by setting all the other variables to 0. Further, identify any two variables of f to be equal if both are non-zero in the same vector from B. Observe that when $y = (0^n)$, the quantity on the left hand side of the lemma is simply

$$\frac{1}{2^e} |\sum_{\substack{x \in \mathbb{F}_2^e \\ |x| \ even}} g(x) - \sum_{\substack{x \in \mathbb{F}_2^e \\ |x| \ odd}} g(x)|,$$

and by Hastad's result, this quantity is bounded by $M2^{-e^{1/d}}$, since the vectors $b^1, \ldots, b^e$ are disjoint, and hence g, too, is computed by a circuit of depth d and size M. For arbitrary, fixed $y \in \mathbb{F}_2^n$, notice that the shifted function $f_y(x) =_{def} f(x + y)$ is also computed by a circuit of depth d and size M. Now the above argument applies when f is replaced by f_y, thus proving the lemma. □

Notice that in the above proof, we assume that g is computable by a sufficiently small constant depth circuit, since the vectors $b^1, \ldots, b^e$ are disjoint, and more specifically, because at most one of these vectors has a '1' at any given coordinate position. In addition, a consequence of the disjointness of the vectors $b^1, \ldots, b^e$ is that any vector x that is not orthogonal to all of these vectors (i.e, $\langle x, b^1 \rangle = 1, \ldots, \langle x, b^e \rangle = 1$) must have size at least e. As explained briefly in the introduction, this motivates the definition of polylog-wise decomposable distributions in the next section. The above lemma can be extended to the case when the vectors $b^1, \ldots, b^e$ are only *almost* disjoint.

DEFINITION 3.2. *A set of vectors $b^1, \ldots, b^k$ in $\mathbb{F}_2^n$ are* **almost disjoint** *if at most $\log n$ of them have a '1' at any coordinate position. More precisely, for any $j : 1 \leq j \leq n$, at most $\log n$ of $b_j^1, \ldots, b_j^k$ equal '1'.*

LEMMA 3.3. *Let f over $\mathbb{F}_2^n$ be computed by a circuit of depth d and size M. Let $B = \{b^1, \ldots, b^e\}$ be a set of independent and almost disjoint vectors. Then for any $y \in \mathbb{F}_2^n$,*

$$\frac{1}{2^e} |\sum_{x \in \mathcal{E}(B)} f(x + y) - \sum_{x \in \mathcal{O}(B)} f(x + y)|$$
$$\leq (M + n^2) 2^{-e^{1/d+2}}.$$

The proof is identical to that of Lemma 3.1, except that the function g of e variables, obtained from f as

$$g(x) =_{def} f(\sum_{x_i \neq 0} b^i),$$

can now be computed by a circuit of size $M + n^2$ and depth $d + 2$, since the sum $\sum_{x_i \neq 0} b^i$ can be computed by n circuits of depth 2 and size n.

LEMMA 3.4. *Let f over $\mathbb{F}_2^n$ be computed by a circuit of depth d and size M. Let $B = \{b^1, \ldots, b^k\}$ be a set of independent vectors, such that $\mathcal{O}(B)$ contains at least e independent, almost disjoint vectors. Then*

$$\frac{1}{2^k} |\sum_{x \in \mathcal{E}(B)} f(x) - \sum_{x \in \mathcal{O}(B)} f(x)| \leq (M + n^2) 2^{-e^{1/d+2}}.$$

PROOF. Since B can be replaced by any independent basis that preserves $\mathcal{E}(B)$ and $\mathcal{O}(B)$, without loss of generality we assume that $B' = \{b^1, \ldots, b^e\}$ is a set of almost disjoint vectors in B. It can be seen that

$$\begin{aligned}
&\sum_{x \in \mathcal{E}(B)} f(x) - \sum_{x \in \mathcal{O}(B)} f(x) = \\
&\sum_{y \in \mathcal{E}(B \setminus B')} \Big(\sum_{x \in \mathcal{E}(B')} f(x+y) - \sum_{x \in \mathcal{O}(B')} f(x+y) \Big) \; - \\
&\sum_{y \in \mathcal{O}(B \setminus B')} \Big(\sum_{x \in \mathcal{E}(B')} f(x+y) - \sum_{x \in \mathcal{O}(B')} f(x+y) \Big).
\end{aligned}$$

From Lemma 3.3, for any $y \in \mathbb{F}_2^n$,

$$\begin{aligned}
&\frac{1}{2^k} |\sum_{x \in \mathcal{E}(B')} f(x+y) - \sum_{x \in \mathcal{O}(B')} f(x+y)| \\
&\leq (M + n^2) 2^{-e^{1/d+2}} 2^{e-k},
\end{aligned}$$

and since $|\mathcal{E}(B \setminus B') \cup \mathcal{O}(B \setminus B')| = |span(B \setminus B')| = 2^{k-e}$, the lemma follows. □

Before we state the main theorem, we formally define and relate e-wise independent and e-wise decomposable distributions.

DEFINITION 3.5. *A subspace U of $\mathbb{F}_2^n$ has* **minimum distance** *e if it satisfies: for all $x, y \in U, |x - y| \geq e$. Subspaces S for which $\hat{S}$ has minimum distance*

at least $e+1$ are called e**-wise independent distributions.** *Recall from Fact 2.1(vi) that $\hat{S}$ is the support of the Fourier transform of the characteristic function of S, and is hence a subspace. In fact, $\hat{S}$ is the orthogonal subspace to S and is sometimes denoted $S^{\perp}$.*

It follows directly from the facts in Section 2 that if $\hat{S}$ has minimum distance e then the elements of S satisfy independent parities each of which has size at least e, i.e, S is the support of a product of parity functions as described in question $\star$ of Section 1. Another equivalent way of defining e-wise independent subspaces is as follows. Any subspace S of $\mathbb{F}_2^n$ of dimension k can be expressed as $S = \{y : y = xG, x \in \mathbb{F}_2^k\}$ where G is a $k \times n$ "generator" matrix of $0's$ and $1's$, whose rows form a basis of S. The subspace S is e-wise independent if every e columns of any such G are independent. In other words, every e coordinate positions of a random vector in S supply e independent random bits.

*A subspace S of dimension k is (m,e)-***decomposable** *if there are subspaces $S_1, \ldots, S_{m-k}$ with bases $B_1, \ldots, B_{m-k}$ such that*

(i) $S \subset S_1 \subset \ldots \subset S_{m-k}$, where $S = \mathcal{E}(B_1)$ and, in general, $S_i = \mathcal{E}(B_{i+1})$, for $1 \leq i < m-k$. In other words, $B_{i+1} = \{a\} \cup (B_i + a)$, for some $a \notin S_i$.

(ii) For $1 \leq i \leq m-k$, $\mathcal{O}(B_i)$ contains at least e independent, almost disjoint vectors.

*We refer to S_{m-k} as the (m,e)-***residue** *of S, and if S is (n,e)-decomposable, we refer to it as e-***wise decomposable**. *Notice that S_i has dimension $k+i$.*

REMARK 3.6. *The distribution given in [15] is a special case of a polylog-wise decomposable distribution. Furthermore, it is not hard to construct small, polylog-wise decomposable distributions that do not satisfy the requirements of [15].*

FACT 3.7. *Every e-wise decomposable subspace of $\mathbb{F}_2^n$ is at least $(e/\log n - 1)$-wise independent.*

PROOF. Let S be an e-wise decomposable subspace of $\mathbb{F}_2^n$ of dimension k, and let $S = \mathcal{E}(B_1) \subset S_1 = span(B_1) = \mathcal{E}(B_2) \subset \ldots \subset S_{n-k} = \mathbb{F}_2^n$ be its decomposition. Noticing that

$$\begin{aligned}\hat{S} = \hat{\mathcal{E}}(B_1) &= \{y : \ \forall x \in \mathcal{E}(B_1)\ \langle x, y\rangle = 0\} \\ &= \{y : \ \forall x \in \mathcal{O}(B_1)\ \langle x, y\rangle = 1\} \\ &\cup \{y : \ \forall x \in span(B_1)\ \langle x, y\rangle = 0\},\end{aligned}$$

and that

$$\{y: \ \forall x \in span(B_1)\ \langle x, y\rangle = 0\} = \hat{S}_1 = \hat{\mathcal{E}}(B_2),$$

we obtain

$$\hat{S} = \hat{\mathcal{E}}(B_1) = \{y: \ \forall x \in \mathcal{O}(B_1)\ \langle x, y\rangle = 1\} \cup \hat{\mathcal{E}}(B_2).$$

Successively expanding $\hat{\mathcal{E}}(B_2), \hat{\mathcal{E}}(B_3), \ldots, \hat{\mathcal{E}}(B_{n-k})$ we obtain

$$\hat{S} = \bigcup_{i=1}^{n-k} \{y : \forall x \in \mathcal{O}(B_i)\ \langle x, y\rangle = 1\} \cup \hat{S}_{n-k}.$$

Since for each i, $\mathcal{O}(B_i)$ contains at least e independent and almost disjoint vectors, it follows that any y that satisfies: $\forall x \in \mathcal{O}(B_i)\ \langle x, y\rangle = 1$ must have an odd (non-empty) intersection with each of these e vectors, and hence $|y|$ must be at least $e/\log n$, since each non-zero bit of y "covers" at most $\log n$ of the e almost disjoint vectors. Further, since $S_{n-k} = \mathbb{F}_2^n$, it follows that $\hat{S}_{n-k} = \{(0^n)\}$, and hence for every non-zero $y \in \hat{S}$, $|y| \geq e/\log n$. Thus, from Definition 3.5, S is at least $(e/\log n - 1)$-wise independent. □

THEOREM 3.8. *Let f over $\mathbb{F}_2^n$ be computed by a circuit of size M, and depth d, and let S be an (m, e)-decomposable subspace of dimension k and residue S_{m-k}. Then*

$$\begin{aligned} &|\frac{1}{2^k}\sum_{x\in S} f(x) - \frac{1}{2^m}\sum_{x\in S_{m-k}} f(x)| \\ &\leq (M + n^2)(m-k)2^{-e^{1/d+2}}. \end{aligned}$$

In particular, if S is e-wise decomposable, then $S_{n-k} = \mathbb{F}_2^n$, and hence

$$|\frac{1}{2^k}\sum_{x\in S} f(x) - \frac{1}{2^n}\sum_{x\in \mathbb{F}_2^n} f(x)| \leq (M + n^2)(n-k)2^{-e^{1/d+2}}.$$

PROOF. Let $S = \mathcal{E}(B_1) \subset S_1 = span(B_1) = \mathcal{E}(B_2) \subset \ldots \subset S_{m-k}$ be the decomposition of S. Starting with

$$\frac{1}{2^k}\Big[\sum_{i=1}^{m-k}\frac{1}{2^i}\Big(\sum_{x\in\mathcal{E}(B_i)} f(x) \ - \sum_{x\in\mathcal{O}(B_i)} f(x)\Big)\Big]$$

and using the fact that $\mathcal{E}(B_i) = \mathcal{E}(B_{i-1}) \cup \mathcal{O}(B_{i-1})$ to write each $\mathcal{E}(B_i)$ in terms of $\mathcal{E}(B_1)$ and $\mathcal{O}(B_j)$'s for $1 \leq j \leq i$, we obtain the equivalent quantity

$$\frac{1}{2^k}\Big[\sum_{i=1}^{m-k}\Big(\frac{1}{2^i}\sum_{x\in\mathcal{E}(B_1)} f(x) - \frac{1}{2^{m-k}}\sum_{x\in\mathcal{O}(B_i)} f(x)\Big)\Big]$$
$$= \frac{1}{2^k}\Big[(1-\frac{1}{2^{m-k}})\sum_{x\in\mathcal{E}(B_1)} f(x) - \frac{1}{2^{m-k}}\sum_{i=1}^{m-k}\Big(\sum_{x\in\mathcal{O}(B_i)} f(x)\Big)\Big]$$

which, since $\bigcup_{i=1}^{m-k} \mathcal{O}(B_i) \cup \mathcal{E}(B_1) = S_{m-k}$,

$$= \frac{1}{2^k}\sum_{x\in\mathcal{E}(B_1)} f(x) \;-\; \frac{1}{2^m}\sum_{x\in\mathcal{E}(B_1)} f(x) \;-\; \frac{1}{2^m}\sum_{x\in S_{m-k}\setminus\mathcal{E}(B_1)} f(x)$$

$$= \frac{1}{2^k}\sum_{x\in\mathcal{E}(B_1)} f(x) \;-\frac{1}{2^m}\sum_{x\in S_{m-k}} f(x) \tag{3.1}$$

Since $\mathcal{O}(B_i)$ has at least e independent, almost disjoint vectors it follows by Lemma 3.4 that for each $1 \leq i \leq m-k$,

$$\frac{1}{2^{k+i}}\Big|\sum_{x\in\mathcal{E}(B_i)} f(x) - \sum_{x\in\mathcal{O}(B_i)} f(x)\Big| \;\leq\; (M+n^2)2^{-e^{1/d+2}},$$

and hence the LHS of Equation (3.1) is bounded by $(M+n^2)(m-k)2^{-e^{1/d+2}}$. The RHS is similarly bounded, thereby proving the theorem. □

REMARK 3.9. *If the definition of "almost disjoint" vectors is generalized to allow* $\log^l n$ *vectors to have* 1*'s in the same coordinate position, then the RHS of Lemma 3.3 becomes* $O((M+n^{2l})2^{-e^{1/(d+2l)}})$. *This is because the parity of* $\log^l n$ *bits can be computed by an* AC^0 *circuit of size* $O(n^{2l})$ *and depth* $2l$*; and the RHS of the Theorem 3.8 then becomes* $O((M+n^{2l})n2^{-e^{1/(d+2l)}})$.

4. Pseudorandom generators and learning algorithms.

The following theorem follows directly from Theorem 3.8 and shows that polylog-wise decomposable subspaces provide pseudorandom generators for AC^0 functions.

THEOREM 4.1. *Every $(3\log n + \log M)^d$-wise decomposable subspace S of $\mathbb{F}_2^n$ has the following property. For every f over $\mathbb{F}_2^n$ computed by a circuit of size M, depth d,*

$$|\frac{1}{|S|}\sum_{x\in S} f(x) - \frac{1}{2^n}\sum_{x\in \mathbb{F}_2^n} f(x)| \leq \frac{1}{n}.$$

In particular, any $((3+l)\log n)^d$-wise decomposable subspace S of $\mathbb{F}_2^n$ serves as a source of good pseudorandom strings that are indistinguishable from random strings from $\mathbb{F}_2^n$ by any $f \in AC^0[d]$ that is computed by a circuit of size $O(n^l)$. I.e, the fraction of S on which f evaluates to a '1' is close (within $1/n$) to the fraction of $\mathbb{F}_2^n$ on which f evaluates to a '1.' Thus the subspace S yields a deterministic $O(|S|)$-time simulation of any bounded error probabilistic circuit of depth d and size $O(n^l)$.

Next we show that polylog-wise decomposable subspaces provide not only good pseudorandom generators, but also deterministic $O(2^{\text{polylog } n})$-time learning algorithms for AC^0 functions, of size and depth depending on the degree of the polylog. Below is a formal definition of a "good" learning algorithm with a uniform training set for a class of functions.

DEFINITION 4.2. *A* **good, deterministic learning algorithm** *A with a* **uniform training set** *S for a class C of functions over $\mathbb{F}_2^n$ is defined as follows. In the learning phase, A queries some function, $f \in C$, $|S|$ times, to obtain the set $\{f(x) : x \in S\}$. In the second phase, on input $y \in \mathbb{F}_2^n$, the Boolean output $A(y)$ of the algorithm satisfies:*

$$\frac{1}{2^n}\sum_{y\in \mathbb{F}_2^n} |A(y) - f(y)| \leq \frac{1}{n}.$$

The next theorem follows from the results of [12], (Section 4), some of the basic properties of Fourier transforms given in Fact 2.1, and Shannon's sampling theorem [18].

THEOREM 4.3. *Let S be a subspace of $\mathbb{F}_2^n$ with the property that for any function f in some class C, the following hold:*

$$\sum_{|y|\leq e'} (\sum_{x\in \hat{S}} \hat{f}(x+y) - \hat{f}(y))^2 \leq u,$$

and

$$\sum_{|x|\geq e'} \hat{f}^2(x) \leq v$$

for some fixed functions e', u and v of n that are independent of f, then there is a deterministic learning algorithm A that uses S as a uniform training set, runs in time max $(|S|, O(n^{e'}))$ and satisfies

$$\frac{1}{2^n} \sum_{y \in \mathbb{F}_2^n} |A(y) - f(y)| \leq u + v$$

for all $f \in C$.

PROOF. Let s be the characteristic function of S. In the learning phase, the algorithm A computes $\widehat{fs}(x)$ for $|x| \leq e'$, and defines the function f_1 as:

$$\hat{f}_1(x) =_{def} (2^n/|S|)\ \widehat{fs}(x) \quad \text{for } |x| \leq e',$$

and 0 otherwise. Note that f_1 need not be Boolean. In the second phase, on input $y \in \mathbb{F}_2^n$, the algorithm "Booleanizes" f_1 and outputs $A(y) =_{def} 0$ if $sign(f_1(y) - 1/2)$ is negative, and 1 if positive. Clearly, this algorithm runs deterministically in time max $(|S|, O(n^{e'}))$. We prove that that

$$1/2^n \sum_{y \in \mathbb{F}_2^n} \left(f_1(y) - f(y)\right)^2 = \sum_{y \in \mathbb{F}_2^n} \left(\hat{f}_1(y) - \hat{f}(y)\right)^2$$

$$\leq u + v.$$

It would then follow directly from the above that $1/2^n \sum_{y \in \mathbb{F}_2^n} |A(y) - f(y)| \leq u + v$, thus completing the proof of the theorem.

First, observe from the definition of f_1 that $\sum_{y \in \mathbb{F}_2^n} \left(\hat{f}_1(y) - \hat{f}(y)\right)^2$ can be split into two parts:

$$\sum_{\substack{y \in \mathbb{F}_2^n \\ |y| \leq e'}} \left(\hat{f}_1(y) - \hat{f}(y)\right)^2 \text{ and } \sum_{\substack{y \in \mathbb{F}_2^n \\ |y| > e'}} \hat{f}^2(y),$$

where the latter part is clearly bounded by v by the assumption of the theorem. It remains to show that the former part is bounded by u. By the definition of f_1, the former part is simply

$$\sum_{\substack{y \in \mathbb{F}_2^n \\ |y| \leq e'}} \left((2^n/|S|)\widehat{fs}(y) - \hat{f}(y)\right)^2.$$

By applying the convolution identity, and the fact that S is a subspace, and therefore by Fact 2.1, $\hat{s}(x) = |S|/2^n$ for all $x \in \hat{S}$, and $\hat{s}(x) = 0$ otherwise, we obtain

$$\widehat{fs}(y) = \left(\hat{f} \circ \hat{s}\right)(y) = \sum_{x \in \hat{S}} \hat{f}(y+x)\hat{s}(x) =$$
$$(|S|/2^n) \sum_{x \in \hat{S}} \hat{f}(y+x).$$

Thus

$$\sum_{\substack{y \in \mathbb{F}_2^n \\ |y| \leq e'}} \left(\hat{f}_1(y) - \hat{f}(y)\right)^2 = \sum_{\substack{y \in \mathbb{F}_2^n \\ |y| \leq e'}} \left(\sum_{x \in \hat{S}} \hat{f}(x+y) - \hat{f}(y)\right)^2,$$

and the latter quantity is bounded above by u by the assumptions of the theorem. □

The next theorem shows that any polylog-wise decomposable subspace provides a uniform training set for learning AC^0 functions.

THEOREM 4.4. *If S is $(3\log n + \log M)^{(d+6)^2}$-wise decomposable subspace of $\mathbb{F}_2^n$, then S is a uniform training set for a good deterministic learning algorithm that runs for $O(|S|)$ steps and learns any function f computed by a circuit of size M and depth d. Here, a function evaluation is assumed to take one step. In particular, any $((3+l)\log n)^{(d+6)^2}$-wise decomposable subspace is a uniform training set for learning any $AC^0[d]$ function computed by a circuit of size $O(n^l)$.*

PROOF. We show that if S is e-wise decomposable, has dimension k, and $e' \leq e/\log n$ then

$$\sum_{|y| \leq e'} \left(\sum_{x \in \hat{S}} \hat{f}(x+y) - \hat{f}(y)\right)^2$$

$$\leq (2(M+n^2)(n-k)2^{-(e-e')^{1/d}})^2 n^{e'} =_{def} u, \tag{4.1}$$

Further, by [12] (Theorem 2.2 (i))

$$\sum_{|x| \geq e'} \hat{f}^2(x) \leq M 2^{e^{1/d}+3} =_{def} v.$$

The proof is then complete on applying Theorem 4.3, and checking that when e is set to the values given in the statement of the theorem, and e' is correspondingly set to $(3\log n + \log M)^{(d+6)}$, (or $((3+l)\log n)^{d+6}$, if $M = n^l$,) then $u + v \leq 1/n$. To show (4.1) it is sufficient to establish that for each $y : |y| \leq e'$,

$$\sum_{x \in \hat{S}} \hat{f}(x+y) - \hat{f}(y) \quad \leq \quad 2(M+n^2)(n-k)2^{-(e-|y|)^{1/d+2}}. \tag{4.2}$$

For the special case $|y| = 0$, it follows directly from Fact 2.1 that:

$$\sum_{x \in \hat{S}} \hat{f}(x) - \hat{f}(0^n) \quad = \quad \frac{1}{|S|} \sum_{x \in S} f(x) - \frac{1}{2^n} \sum_{x \in \mathbb{F}_2^n} f(x),$$

and since S is e-wise decomposable, by Theorem 3.7, the above quantity is at most

$$(M+n^2)(n-k)2^{-e^{1/d+2}},$$

thereby showing (4.2) for the special case $y = (0^n)$.

When $y \neq (0^n)$, there are two possible ways of tackling the quantity $\sum_{x \in \hat{S}} \hat{f}(x+y) - \hat{f}(y)$. The first is to analyze the shifted function $\hat{f}(. + y)$ and the second is to analyze the shifted subspace (coset) $\hat{S} + y$.

The first method is my interpretation of a comment by an anonymous referee, and depends on additional properties of AC^0 functions. Defining $\hat{f}_y(x)$ as the shifted function $\hat{f}(x+y)$, the quantity $\sum_{x \in \hat{S}} \hat{f}(x+y) - \hat{f}(y)$ becomes $\sum_{x \in \hat{S}} \hat{f}_y(x) - \hat{f}_y(0^n)$. As mentioned in the previous paragraph, this quantity is bounded by

$$(M'+n^2)(n-k)2^{-e^{1/d'+2}},$$

where M' and d' are the size and depth of the circuit computing f_y. Now $\hat{f}_y(x)$ is directly seen to be $\hat{f}(x)(-1)^{\langle x,y \rangle}$, and $(-1)^{\langle x,y \rangle}$ is just (a simple linear transformation of) the *parity* function involving the non-zero bits in y. Thus a careful choice of $|y|$ i.e, e' and additional properties of AC^0 circuits (see Remark 3.9) together ensure that M' and d' are small enough to establish the required bound. In particular, M' is seen to be $O(M + 2^{e'}$ and d' is $max\{d, 2^{e'}\} + 1$.

The second method, namely analyzing the coset $\hat{S} + y$, depends only on the e-wise decomposability of S and not on additional properties of AC^0 circuits. We describe this method below. We show that when $|y| \leq e' \leq e/\log n$,

$$S_y =_{def} (\hat{S} \cup (y + \hat{S}))^{\vee} = S \cap \{x : \langle x, y \rangle = 0\}$$

is $(n-1, e-|y|)$-decomposable, with residue exactly $\{(0^n), y\}^{\vee}$.

Let

$$S = \mathcal{E}(B_1) \subset S_1 = \mathcal{E}(B_2) \subset \ldots \subset S_{n-k} = \mathbb{F}_2^n$$

be the decomposition of S. This yields an equivalent decomposition:

$$\hat{S}_{n-k} \subset \hat{S}_{n-k-1} \subset \ldots \subset \hat{S}_1 \subset \hat{S}.$$

Since $|y| < e' \leq e/\log n$, and since S is e-wise decomposable and by Fact 3.7, $e/\log n$-wise independent, it follows that $\hat{S}$ does not contain y, and clearly none of the $\hat{S}_i$'s contain y. Denoting $S_{i,y} =_{def} (\hat{S} \cup (y + \hat{S}_i))^\vee$, we construct the $(n-1, e-|y|)$-decomposition of S_y as (recall that S_y has dimension $k-1$, and hence such a decomposition has $n-k$ terms besides S_y):

$$S_y \subset S_{1,y} \subset \ldots \subset S_{n-k,y} = \{(0^n), y\}^\vee.$$

It remains only to show that for the bases $B_{i,y}$ of the $S_{i,y}$ that satisfy $\mathcal{E}(B_{i+1,y}) = span(B_{i,y})$, it holds that $\mathcal{O}(B_{i,y})$ contains at least $e-|y|$ almost disjoint vectors. Without loss of generality assume that each of the B_i's in the decomposition of S is a standard basis, i.e, if the vectors in B_i are the rows of a $(k+i) \times n$ matrix G, then G is of the form $[I_{k+i,k+i} A]$, where $I_{k+i,k+i}$ is the identity matrix with $k+i$ rows and columns, and A is an arbitrary $(k+i) \times n-(k+i)$ matrix of 0's and 1's. Furthermore, it can also be assumed that the e almost disjoint vectors of $\mathcal{O}(B_i)$ in fact belong in B_i. Clearly, the B_i's can be identically modified and reduced by one vector to give corresponding bases $B_{i,y}$ of $S_{i,y}$ that satisfy $\mathcal{E}(B_{i+1,y}) = span(B_{i,y})$. Furthermore, since y is the only additional vector that is orthogonal to $B_{i,y}$ but not to B_i, it follows that $B_{i,y}$ contains all but $|y|$ vectors from B_i (the independent even combinations of these $|y|$ vectors of B_i form the remaining $|y|-1$ vectors of $B_{i,y}$). Now since B_i contains at least e almost disjoint vectors, it follows that $B_{i,y}$ contains at least $e-|y|$ almost disjoint vectors, thus showing that S_y is $(n-1, e-|y|)$-decomposable with residue $\{(0^n), y\}^\vee$. Therefore, by Theorem 3.8,

$$\begin{aligned} \frac{1}{|S_y|} \sum_{x \in S_y} f(x) - \frac{1}{2^{n-1}} \sum_{x \in \{(0^n), y\}^\vee} f(x) & \\ \leq \quad M(n-k) 2^{-(e-|y|)^{1/d+2}}. & \qquad (4.3) \end{aligned}$$

Now,

$$\sum_{x \in \hat{S}} \hat{f}(x+y) - \hat{f}(y) \quad = \quad \sum_{x \in \hat{S}_y} \hat{f}(x)$$

$$
\begin{aligned}
&- \Big(\sum_{x\in\hat{S}} \hat{f}(x) - \hat{f}(0^n)\Big) \quad - \quad \Big(\hat{f}(y) + \hat{f}(0^n)\Big) \\
&= \Big(\frac{1}{|S_y|}\sum_{x\in S_y} f(x) - \frac{1}{2^{n-1}} \sum_{x\in\{(0^n),y\}^\vee} f(x)\Big) \\
&- \Big(\frac{1}{|S|}\sum_{x\in S} \hat{f}(x) - \frac{1}{2^n}\sum_{x\in\mathbb{F}_2^n} f(x)\Big).
\end{aligned}
$$

Now (4.3) and the ϵ-wise decomposability of S bound the above two terms, thus proving (4.2), and thereby the theorem. □

OPEN QUESTION 4.5. *There is no known example of a polylog-wise independent distribution that is not polylog-wise decomposable. Furthermore, using Remark 3.9 and the definition of decomposable distributions, it is sufficient to prove (or disprove) the following in order to find such an example (or completely settle Question $\star$).*

There is a linear, $\log^{3l} n$-wise independent distribution $S \subseteq \mathbb{F}_2^n$, for some $l \geq n$ such that for all vectors $a \notin S$ and all sets of $log^{3l}n$ independent vectors in $S + a$, there is some coordinate position at which strictly greater than $\log^l n$ of these vectors have a '1.'

Acknowledgements

I thank Eric Bach for his extensive comments on an earlier version of the manuscript; Anne Condon, Johan Hastad, Deborah Joseph, Gary Lewandowski, Nati Linial, Randy Pruim, and two anonymous referees for their patience and valuable suggestions; and Shmuel Safra for putting me in touch with some of these people.

References

[1] M. BELLARE, A technique for upper bounding the spectral norm with applications to learning. In *Proc. 5^{th} Ann. IEEE Symp. Computational Learning Theory (COLT)*, 1992.

[2] J. BRUCK, Harmonic analysis of polynomial threshold functions. *SIAM J. Discrete Mathematics*, **3** (2), 1990, 168-177.

[3] R. BOPPANA, M. SIPSER, *The complexity of finite functions*. Technical Report, Massachussetts Institute of Technology, Laboratory for Computer Sciences, MIT/LCS/TM-405, 1989.

[4] J. Bruck, R. Smolensky, Polynomial Threshold Functions, AC^0 Functions and Spectral Norms. *SIAM J. Computing,* **21** (1), 1992, 33-42.

[5] H. Dym, H.P. McKean, *Fourier series and integrals.* Probability and Mathematical Statistics series, Academic Press, 1972.

[6] M. Dowd, M. Sitharam, Shannon bounds for functions over $\mathbb{Z}_n^2$. Technical Report, Kent State University, Department of Mathematics and Computer Science, 1990.

[7] M. Furst, J. Jackson, S. Smith, Improved learning of AC^0 functions. In *Proc.* 5^{th} *Ann. IEEE Symp. on Computational Learning Theory (COLT),* 1992.

[8] M. Furst, J. Saxe, M. Sipser, Parity, circuits and the polynomial time hierarchy. *Mathematical Systems Theory,* **17**, 1984, 17-27.

[9] J. Hastad, *Computational limitations of small depth circuits.* Ph. D thesis, Massachussetts Institute of Technology press, 1986.

[10] J. Kahn, J. Kalai, N. Linial, The influence of variables on Boolean functions. *Proc.* 29^{th} *Ann. IEEE Symp. Foundations of Computing (FOCS),* 1988, 68-80.

[11] E. Kushilevitz, Y. Mansour, Learning decision trees using the Fourier transform. *SIAM Journal on Computing,* **22** (6), 1993, 1331-1348.

[12] N. Linial, Y. Mansour, N. Nisan, Constant depth circuits, Fourier transforms, and learnability. To appear in *JACM*; In *Proc.* 30^{th} *Ann. IEEE Symp. on Foundations of Computing (FOCS)*, 1989, 574-579.

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

[14] F.J. MacWiliams, N.J.A. Sloane, *The theory of error-correcting codes.* North Holland, 1977.

[15] N. Nisan, A.W. Widgerson, Hardness vs. randomness. To appear in *JCSS*; In *Proc.* 29^{th} *Ann. IEEE Symp. on Foundations of Computing*, 1988.

[16] N. Nisan, M. Szegedy, On the degree of Boolean functions as real polynomials. In *Proc.* 24^{th} *Ann. ACM Symp. Theory of Computing,* 1992, pp. 462-467.

[17] A.A. Razborov, Lower bounds on the monotone complexity of some Boolean functions. *Soviet Mathematics Doklady,* **31**, 1985, 354-357.

[18] C.E. Shannon, Communication in the presence of noise. In *Proceedings of the IRE,* **37**, 1949, 10-21.

[19] K.I. Siu, J. Bruck, On the power of threshold circuits with small weights. *SIAM Journal of Discrete Mathematics,* 4 (3), 1991, 423-435.

[20] A.C. Yao, Lower bounds by probabilistic arguments. In *Proc.* 24^{th} *Ann. IEEE Symp. Foundations of Computing,* 1983, 420-428.

[21] A.C. Yao, Separating the polynomial time hierarchy by oracles. In *Proc.* 26^{th} *Ann. Symp. on Foundations of Computing* , 1985, 1-10.

Manuscript received 10 January 1994

Meera Sitharam
Department of Mathematics and Computer Sciences
Kent State University
Kent, OH 44240, USA
sitharam@mcs.kent.edu