Electronic Colloquium on Computational Complexity - Reports Series 1996 - available via:
ECCC
TR96-030
FTP: ftp.eccc.uni-trier.de:/pub/eccc/
WWW: http://www.eccc.uni-trier.de/eccc/
Email: ftpmail@ftp.eccc.uni-trier.de with subject 'help eccc'

APPROXIMATION FROM LINEAR SPACES AND APPLICATIONS TO COMPLEXITY

Meera Sitharam

Abstract. We develop an analytic framework based on linear approximation and point out how a number of complexity related questions – on circuit and communication complexity lower bounds, as well as pseudorandomness, learnability, and general combinatorics of Boolean functions – fit neatly into this framework. This isolates the analytic content of these problems from their combinatorial content and clarifies the close relationship between the analytic structure of questions.
(1) We give several, general results that characterize approximability from spaces of functions and hence also represent general analytic methods for showing non approximability.
(2) We point out that crucial portions of a significant number of the known complexity-related results can be given shorter and cleaner proofs using these general theorems: this clarifies their common analytic structure. We however provide only a few of the alternative proofs.
(3) We give several new complexity-related applications, including circuit complexity lower bounds, and results concerning pseudorandomness, learning, and combinatorics of Boolean functions.
(4) Finally, we suggest natural and promising directions for further investigation.

Key words. Circuit complexity; Communication complexity; Complexity Lower bounds; Fourier transforms; Linear approximation; Nonlinear approximation; Learning theory; Pseudorandomness.

Subject classifications. 68Q15, 68Q99.

1. Introduction

In the context of complexity lower bounds, the "approximation method" usually refers to the method originated by Razborov in [57] and [59] for proving monotone lower bounds. The approach was continued by [58] and [66] and several others including [6], [67], [9], [68], [71], [77], [39] for general lower bounds, and further used in monotone lower bounds such as [2] [78] and [10]. Other complexity lower bounds that can be generally classified as being "based on non approximability by low degree or sparse polynomials, or other basis functions" include many of the lower bounds on threshold circuit complexity and "voting polynomial representations" such as [11], [15], [42], [43] [26], [3], [28], etc. Some of these results have been collected in survey articles by [8] and [60]. Further results that could be seen to involve Boolean (non) approximability include results on general Boolean functions in [51], [52], [46], [37], and [47].

While these results are viewed as being based, broadly speaking, on the analytic notion of approximation of Boolean functions by sets of monomials, or other suitable basis functions, no closer relationship between the analytic structure of these problems has been established. In particular, the techniques that have been used for showing (non)approximability have not been unified, and are in fact considered to be quite different. For example, the paper [42] states: "Previous lower bound results on threshold representations are based on three different techniques, the discriminator method, a geometric method based on probabilistic communication complexity, and a spectral theoretic method for orthogonal bases." Similar statements can be found in several other papers throughout the literature.

In this paper, we develop an analytic framework based on linear approximation and point out how a number of complexity related questions – not only those considered in the papers above, but also other questions on circuit and communication complexity lower bounds and hardness, as well as pseudorandomness, learnability, and general combinatorics of Boolean functions – fit neatly into this framework. This isolates the analytic content of these problems from their combinatorial content and clarifies the close relationship between the analytic structure of questions. In addition, the framework facilitates a systematic study and application of analytic techniques, and, in particular, shows that many of the above proof techniques are minor variations of the same general technique.

It should be noted that the analytic methods studied here reduce complexity-related questions to their combinatorial essence, at which point possibly ad hoc combinatorial methods might be required to complete the solution. In other words, the analytic framework helps to break up complexity related questions into their constituent analytic and combinatorial subproblems, which are often of independent mathematical interest and require new and original techniques for solution.

The analytic methods used in the paper are various combinations of 2 basic ingredients
• modifications of the duality principle for linear approximation that hold when the function to be approximated is Boolean and/or the approximating space has a Boolean basis, and
• simple norm relationships specific to Boolean functions.

Organization. Section 2 gives basic preliminaries and conventions used in the paper. Section 3 introduces the duality principle for linear approximation and explains its relevance to complexity questions; the section presents a general analytic framework based on duality, by which one can view various complexity-theoretic questions as versions of the same analytic problem with different choices of parameters; and describes, with examples, the basic ingredients that constitute the analytic techniques for showing Boolean (non)approximability. Sections 4, 5 and 6 of the paper concern three (non)approximability questions that result from particular choices of parameters in the the analytic framework. These are: interpolation, one-sided approximation, and uniform approximation. The final section 7 covers algorithms for *finding* approximating functions. Sections 3 to 7 follow a fixed pattern of exposition.
(1) We give several, general results that characterize approximability from spaces of functions and hence also represent general analytic methods for showing non approximability.
(2) We point out that crucial portions of a significant number of the known complexity-related results can be given shorter and cleaner proofs using these general theorems: this clarifies their common analytic structure. We however provide only a few of the alternative proofs.
(3) We give several new complexity-related applications, including circuit complexity lower bounds, and results concerning pseudorandomness, learning, and combinatorics of Boolean functions.
(4) Finally, we suggest natural and promising directions for further investigation.

Scope. We note that our results and techniques are suitable primarily for spaces

of functions whose range is $\mathbb{R}$. Hence we do not deal with other valid representations of Boolean functions, for example, as functions whose ranges are finite fields, as in [66] [6], [9], [71], [77], etc. Furthermore, it should be noted that most of the general results and methods in this paper are inherently multivariate, and do not use rely on univariate approximation which have limited scope and on which, for example, the results of [51], [52], and some of [47] and [3] are based; we point out the key difference in Section 3. In addition, our discussion is directed towards non approximability results and resulting lower bounds. Hence even our approximability results are geared towards the eventual goal of proving non approximability. We pay scant attention to upper bounds that are obtainable from approximability results including the numerous threshold circuit complexity upper bounds in the literature (See [48] and [49]). Finally, our analytic framework is suitable primarily for questions that can be decomposed into *linear* approximation questions. Many of the lower bounds based on [57] and [59], such as [39] [2] [78] and [10], use distinctly non-linear approximation methods. While it is an open question whether these, too, can be treated using purely linear approximation methods, we discuss the current points of difference in Section 3.

General results. Below, we give an informal description of some of the general (non) approximability and interpolability characterizations in the order in which they appear in the paper. In addition, if easy to state, we mention known results that are generalized by these characterizations. We consider approximability of Boolean functions f from vector spaces X of functions from $\{-1,1\}^n$ to $\mathbb{R}$. We will use the inner product $< f,g >:= 1/2^n \sum_x f(x)g(x)$ and often refer to simple concepts from linear algebra, such as the orthogonal space $X^\perp$ consisting of functions that have 0 inner product with every function in X, and the projection $f|_X$ of a function f onto a space X.

The first result concerns a characterization in [15] which states that PT_1 functions (i.e, functions whose signs can be represented by linear combinations of at most polynomially many *Parity* functions), are defined uniquely by few of their Fourier coefficients. This gives a bound on the number of distinct PT_1 functions.

THEOREM 3.4 is stronger, and follows directly from a version of the duality principle that is specific to Boolean functions. It asserts the *equivalence* of two statements: for any Boolean function f, and any subspace X of functions from $\{-1,1\}^n$ to $\mathbb{R}$, there is a function in X with the same sign as f if and only if the projections $f|_X$ extends to a unique function bounded by 1.

This theorem gives a bound on the number of Boolean functions that are

approximable from any space X in terms of the dimension of X and some properties of bases for X.

THEOREM 4.1 converts a statement of approximability to an equivalent statement of interpolability. It states that for any subspace X of functions (from $\{-1,1\}^n$ to $\mathbb{R}$), and function f in the orthogonal space $X^\perp$, the following two statements are equivalent.
• No $g \in X$ has the the same sign as f on a set S on more than m points.
• For every set $\bar{S}$ of at most $2^n - m$ points, there is a function $h \in span((X \cup \{f\})^\perp)$ that interpolates f on $\bar{S}$, and is bounded above by f elsewhere.
Section 4 gives known results that provide context to this theorem.

THEOREM 4.4 converts a statement of non-interpolability to an equivalent statement of interpolability. let X be a subspace of the usual space of functions and let f be any function in $X^\perp$. The following statements are equivalent.
• No function in X of degree $\leq d$ interpolates f on a set S of $> m$ points
• For every set $\bar{S}$ of $< 2^n - m$ points, there is a function in $X^\perp \setminus f$, that interpolates f on $\bar{S}$.

This gives a different handle on an open problem posed by [66] namely to extend the results - on the non-interpolability of functions from a space of low degree polynomials over finite fields - to spaces of polynomials over the reals.

THEOREM 5.1 gives a systematic method of obtaining distributions that mimic the uniform distribution and "fool" any Boolean function f that is approximable from a given space X of functions. In general, the theorem shows that if every function in a complexity class C is approximable in the ∞-norm to within, say $1/n$, from a subspace X, then any (Boolean) function h in the orthogonal space $X^\perp$ is hard for C, and can be expressed as $h^* + c.One$, where One denotes the constant function; furthermore, h^* gives a distribution that fools every function in C. Such distributions that are, in addition, easy to generate, provide a large class of natural pseudorandom generators for all computations in C, and highlight the close relationship between hardness and pseudorandomness (see, for example, [50]).

THEOREM 5.6 gives an exact characterization of when a large class of distribution l fools a function f in terms of a notion of "one-sided" approximability of f. See Section 5 for known results that put this theorem into context.

THEOREM 6.1 converts non approximability in the ∞-norm into equivalent approximability results. Let f be a Boolean function and B a set of Boolean functions, and let $M \subseteq B$ consist of independent Boolean functions.

(1) The following are equivalent.
- There does not exist an approximation $g \in span(M)$ with $sign(f) = sign(g)$.
- There exists an approximation $l \in span(M)^{\perp}$ with $||l||_1 > 0$ and $sign(f(x)) = sign(l(x))$, whenever $l(x) \neq 0$.

(2) The following are equivalent.
- There does not exist an approximation $g = \sum_{h \in M} a_h h$ with $\sum_{h \in M} |a_h| \leq 1$, *and* $\epsilon \leq |g(x)| \leq 1$ everywhere, and $sign(f) = sign(g)$.
- There exists an approximation l *close* to $span(M)^{\perp}$ with $||l||_1 = 1$ and $sign(f) = sign(l)$, where ever $l \neq 0$. By "close to $span(M)^{\perp}$" we mean that $|\sum_x l(x)h(x)| \leq \epsilon$ for all $h \in M$.

Almost all the threshold circuit lower bounds known so far concerning non approximability in the ∞-norm (non expressibility of the sign) of a function f from the span of small number of LT_1 or other functions involve a *restriction* on the approximation: the the linear combinations that form the approximant have polynomially bounded coefficients. These include results in [33], [41], [44], [26]. In other words, these lower bounds apply only to circuits with an unweighted threshold gate at the top. These lower bounds *all* use the 'correlation/discriminator lemma,' proved in [33] and [26], which is nothing but 6.1(2), for which we give give a short and straightforward proof using the duality principle. To complete such a lower bound proof, one then needs to show that the scalar product $| < h, l > |$ is small for each $h \in M$. The methods for bounding this scalar product have been phrased in terms of communication complexity (for example [26]) and "variation rank" (for example [44]), but most of these also reduce to arguments based on duality and simple norm relationships for Boolean functions as will be shown in Section 6.

The only non approximability results *without* the above restriction are the following: the result of [23] on the non approximability of *Parity* by few functions computable by AC^0 circuits, i.e, $\{\wedge, \vee, \neg\}$- circuits of a fixed polynomial size and constant depth; related results of [43], for example, on the non approximability of an $AC^0[3]$ function by few *And* functions and the result of [42] on the non approximability of an $AC^0[3]$ function by few *Mod r* functions (which can be viewed as monomials over the reals, or $\mathbb{Z}_2^n$ characters; in fact, the result applies to $\mathbb{Z}_r^n$ characters, for any r). The main analytic technique in all of these papers reduces to 6.1 (1) (which is based on duality) although the respective papers do not state as such.

THEOREM 6.9 shows that when the set M of functions is an orthonormal set,

Theorem 6.1 implies a generalization of the "spectral method" of [15]: the number of *Parities* or monomials (or any other orthonormal set) needed to approximate a function exceeds the inverse of its maximum Fourier coefficient (or respectively, the maximum scalar product of the function with an element of the orthonormal set). I.e, $PT_1 \subseteq PL_\infty^{-1}$.

THEOREM 6.4 is a direct consequence of 6.1(1). Let f be Boolean and X any space. If $f \notin X$ and the maximum absolute value of the projection $||f|_X||_\infty \leq 1$ then there is no $g \in X$, with the same sign as f. This result is useful when the projection $f|_X$ can be easily found, for example, if a well-behaved orthonormal basis exists for X, and the scalar product of f with every basis function is small.

We use 6.4 directly together with the properties of the spectra of read-once $AC^0[3]$ functions to provide an alternative proof of the non approximability result that $AC^0[3] \not\subseteq PT_1$, [42].

THEOREM 6.6 considers a natural situation where a set M does not consist of orthonormal functions, and yet 6.4 can be used to show non approximability from $span(M)$. Here, all the functions in M, when viewed as vectors in $\mathbb{R}_{2^n}$, form vector bundles, such that all the vectors in any one bundle are close to each other (have large scalar product), but any two bundles are nearly orthogonal to each other. In particular, if all pairs of functions g_i, g_j in the class M of Boolean functions satisfy either $| < g_i, g_j > | \leq \delta$ or $| < g_i, g_j > | \geq 1 - \delta$, for δ being typically significantly less than $1/2$, and furthermore, for a given Boolean function f, the quantity $| < f, g_i > | \leq \epsilon$, for all $g_i \in M$, then if $|M| \leq min\{1/\delta^{1/3}, 1/\epsilon^{1/3}\}$, there is no $g \in span(M)$ with the same sign as f.

THEOREM 6.12 gives several results of the following form that illustrate the transitivity of approximability relationships: "f is approximable from the span of a set of m_1 functions g_i in some basis B, and g_i are all approximable from the span of a set of m_2 "simple" functions h then f is approximable from the span of $m_1 m_2$ "simple" functions." These results follow directly from 6.1.

We use results of this nature for building on previous non approximability results. For example, a result of the above form, together with the non approximability of f from the span of $m_1 m_2$ simple functions would imply that one of the approximability hypotheses is false.

Often, these transitive approximability results are used in conjunction with approximability results that are obtained from communication complexity upper bounds using the following facts.

FACT 6.10 and 6.11 state that lower bounds on deterministic and probabilistic communication complexity of a Boolean function f can be obtained by prov-

ing appropriate non approximability of f from the span of the characteristic functions of *cross-product* sets.

Results in 6.12 form the backbone of the lower bounds (non approximability results) of [33] and [26] concerning unweighted threshold circuits of depth 2 and 3, and their relationship to weighed and unweighed thresholds of *Parity* functions, namely the results stating that that $\hat{LT}_3 \not\subseteq \hat{LT}_2$, $LT_1 \not\subseteq \hat{PT}_1$, and $PT_1 \not\subseteq \hat{LT}_2$. The paper [26] uses communication complexity upper bounds to prove approximability of the functions in the relevant basis B by the span of a few cross-product functions and then, in effect, uses 6.1 (2) to show that f is not appropriately approximable from the span of few cross-product functions. The desired non approximability of f from the span of few functions in B then follows from 6.12. The papers, especially [26], however, employ the communication complexity paradigm throughout instead of treating the issue as one of approximability from cross-product functions. We point out that the word "communication complexity" can be usually removed from all (lower bound) proofs involving threshold functions, and many other lower bound proofs, without making the proofs any more difficult, or any less intuitive.

For example, the following result of [26], which uses communication complexity is a direct consequence of 6.12: "A circuit with an unweighted linear threshold gate on top, arbitrary linear threshold gates at the middle level, and gates from a class C in the lowest level can be simulated by a circuit with exactly the same gate on top, unweighted linear threshold gates in the middle level and exactly the same gates from C at the bottom." Furthermore, although, in theory, a communication complexity upper bound is stronger than an approximability result from the span of cross-product functions, usually, one can obtain such an upper bound by proving (possibly transitive) approximability as well. For example, 6.12 can be directly used to show an upper bound on the probabilistic communication complexity of LT_1 functions. As noted in the description of the scope of the paper, these are the only two examples of complexity upper bounds that we show to be obtainable from approximability results. However, it seems natural and promising to study the general use of approximability results for proving threshold and communication complexity upper bounds, especially since every *threshold* complexity upper bound is equivalent to an approximability result in the ∞ norm.

Similarly, although non approximability results from cross-product functions are stronger than communication complexity lower bounds, in fact, in practice, many lower bounds on communication complexity, including results in [33], [26], [32], [38], [56], [25], [29], [4], and [24], do yield stronger non approximability results, which could be proved independently using versions of 6.1(2)

in conjunction with 6.12. Thus proving non approximability from cross-product functions is a viable method for proving lower bounds in communication complexity.

OBSERVATION 6.14 contains a series of results, all of which show the non approximability of functions f in the ∞ norm (or non-expressibility of the sign) by decomposing the domain in a systematic way. These results are also useful in constructing hard f based on previously proven or easier non approximability results. Furthermore, these results give a general method for reducing a non approximability question into a combinatorial problem. This method has been used in several papers, although not stated as such, for example [44], [42], [26]. It is the crux of the proof in [44], although not stated as such, that the function DIP_2, which computes the inner product of 2 vectors in $\mathbb{F}_2$, is not closely approximable by the span of few symmetric functions. It is, in effect, also the crux of the proof of [42] showing that there are $AC_0[3]$ circuits that cannot be simulated by a threshold of quasi-polynomially many parity functions. I.e, $AC_0[3] \not\subseteq QT_1$.

THEOREM 6.22 provides several ways of showing that the scalar product between a Boolean function f and any function in a class of Boolean functions is small. Results of this nature are used in conjunction with the methods of 6.1 to show non approximability.

Finally, based on the analytic approximation framework, we are able to relate several approximation algorithms, some of which have appeared as learning algorithms, [46] [22], [34], [40], [64], and some of which are classical algorithms in the approximation theory literature.
THEOREM 7.1 states that to find an approximation with the same sign as a Boolean function f from a space X, there is a set of $dim(X)+1$ sample points on which it is sufficient to sample f. Furthermore, these points are the support of a function in $X^{\perp}$. (Recall 5.1 showing that such functions also provide distributions that fool f).

This result has fairly general consequences for obtaining deterministic approximation algorithms, which are however listed under the complexity-related applications below, since they also concern learning and pseudorandomness.

Specific applications. Some examples of specific, new complexity related applications are given below in the order in which they appear in the paper.

It is an elementary fact that no polynomial of degree bounded by $n-1$ has the same sign as *Parity*.

THEOREM 3.5 is a simple generalization of this result and has a two-line proof using an analytic technique developed in this paper. It states that scalar multiples of *Parity* are the *only* functions (Boolean or otherwise), over $\{1,-1\}^n$ whose sign does not coincide with that of any polynomial of degree bounded by $n-1$.

A conjecture of [47] states that any appropriately polylogwise independent distribution fools AC^0 functions.
OBSERVATION 5.5 states that a modified version of this conjecture - where the distributions are replaced by functions whose 1-norm is bounded by 1, but are allowed negative values - is false.

The proofs of [42] and [34] for the *approximability* result that $AC^0[2] \subseteq \hat{PT}_1$ involve a probabilistic existence argument for the approximant in the former case and an algorithm to find the approximant in the latter case. We give a transitive approximability result (see 6.12) whose proof is straightforward from 6.1(2)), which moreover gives a more general result a corollary.
In particular, COROLLARY 6.13 states that an unweighted linear threshold of polynomially many functions in PL_1, i.e, whose Fourier transforms have polynomially bounded 1-norms, can be simulated by an unweighted threshold of polynomially many *Parity* functions.

The conjecture that $AC^0[3]$ is not contained in LT_2, i.e, the class of weighted thresholds of polynomially many threshold gates, has been posed by [42]. Settling this conjecture, would, in particular, settle the embarassing open question as to whether LT_2 is different from NP. In general, as mentioned earlier, very few lower bounds are known involving circuits with weighted threshold gates at the top.
THEOREM 6.19 is a partial result in this direction. A canonical $AC^0[3]$ function does not have an approximation with the same sign, from the span of a set $M \subseteq LT_1$, of polynomially bounded size provided M is closed under all the permutations of variables under which the canonical function is invariant.

THEOREM 6.20 gives a result that is weaker in some senses, but stronger in others than similar results concerning non approximability of an explicit function from the spans of *And* and AC^0 functions in [43] and [23], but using a different proof technique, where 6.1 (1) plays a crucial role.
THEOREM 6.21 extends the above result to *Flat* functions that are more general than *And* functions.

THEOREM 6.23 concerns correlations with LT_1 functions. Given a Boolean function f, and a function $g \in LT_1$ if for each subset $S \subseteq \{-1,1\}^n$ with $|S|$

containing more than an $(1+\epsilon)/2$ fraction of points, one of the several conditions holds on the geometric structure of f, then $< f, g > \leq \epsilon$. For example, one of these conditions is that $ConvexHull(f^{-}1(1) \cap S)$ and $ConvexHull(f^{-}1(-1) \cap S)$ intersect.

We give several general methods throughout Section 6 that use results of this nature for proving non approximability results.
THEOREM 6.24 states that, in particular, the correlation of a canonical read-once $AC^0[3]$ function with any function in LT_1 is comparable to the expected value of a canonical function.

REMARK 7.2 is based on Theorems 7.1 and 5.1 listed under the general results, and points out a direct connection between sets of pseudorandom elements for a general class of functions C and deterministic sample sets for approximating - in the ∞ and 2 norms - or learning functions in C. When easy to generate, these distributions can be used to derandomize randomized approximation and learning algorithms and randomized computations in C. Such a connection was previously established for the special case of AC^0 in [64].

Using these ideas, we derandomize the learning algorithm of [22] for AC^0 functions and portions of the learning algorithm of [34] for $\hat{PT}_1$ functions, and [40] for decision trees.

2. Background and conventions

Unless otherwise specified, all function domains consist of n-tuples in $\{-1, 1\}^n$ viewed as subsets of both $\mathbb{R}^n$ *and* the finite vector space $\mathbb{F}_2^n$, *with -1 mapping to* $1_{\mathbb{F}_2}$ *and 1 mapping to* $0_{\mathbb{F}_2}$. The number of arguments of a function is often omitted and is assumed to be n. Similarly, the words "polynomially many" and "polynomially bounded" usually refers to a polynomial in n. The range of all functions is $\mathbb{R}$, and for Boolean functions, the range is $\{1, -1\}$, viewed primarily as a subset of $\mathbb{R}$, (and occassionally as a subset of $\mathbb{F}_p$). Thus, for example, the functions $\wedge$, $\vee$ etc. map from $\{1, -1\}^n$ to $\{1, -1\}$ in the obvious way, *with -1 mapping to the usual* 1 *and 1 mapping to* 0.

The number of '1' entries in a vector $x \in \mathbb{F}_2^n$ is denoted $|x|$, and the n-tuple $(a, \ldots, a)$ is denoted (a^n). A vector $x \in \mathbb{F}_2^n$ is identified with the set of coordinates $1 \leq i \leq n$ where $x_i = 1$. Thus, given vectors x and y, we will refer to the vectors $x \cup y$, $x \cap y$, $x \setminus y$, $\bar{x}$ or $\neg x$ (for the bitwise complement of x), and expressions such as $i \in x$ (meaning $x_i = 1$). The inner product $< x, y >$ for $x, y \in \mathbb{F}_2^n$ is '1' if the parity of $|x \cap y|$ is odd.

The Fourier transform of a function f from $\mathbb{F}_2^n$ (or the group $\mathbb{Z}_2^n$) to $\mathbb{R}$ is denoted $\hat{f}$ and is given by

$$\hat{f}(x) = 1/2^n \sum_{u \in \mathbb{F}_2^n} f(u)(-1)^{<x,u>};$$

thereby $f(x)$ can be written as $\sum_{u \in \mathbb{F}_2^n} \hat{f}(u)(-1)^{<x,u>}$. The characters $\chi_u(x)$ are defined as $(-1)^{<x,u>}$, and are generally called parity functions. For $u = 1^n$, the function χ_u is called *Parity*, and for $u = 0^n$, the (constant) function is called *One*.

The following are basic properties of the Fourier spectra of Boolean functions.

FACT 2.1. *For functions f and g over $\mathbb{F}_2^n$ the following hold.*

(i) Parseval's identity:

$$||f||_2^2 = (1/2^n) \sum_{x \in \mathbb{F}_2^n} f^2(x) \; = \; \sum_{x \in \mathbb{F}_2^n} \hat{f}^2(x) = ||\hat{f}||_2^2.$$

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

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

Functions f over $\{-1,1\}^n$ are also representable uniquely as multilinear polynomials from $\mathbb{R}^n$ to $\mathbb{R}$. I.e, there is a unique multilinear polynomial over $\mathbb{R}^n$ that interpolates f at its domain points.

We will often use the following: when the range of functions is $\{1,-1\}$ then for $x \in \{1,-1\}^n$, The functions $\chi_u(x)$ are nothing but $\prod_{i \in u} x_i$. Thus, $f(x) = \sum_{y \in \mathbb{F}_2^n} \hat{f}(y) \prod_{i \in y} x_i$. In other words, the coefficient of $\prod_{i \in y} x_i$ in the multilinear polynomial over $\mathbb{R}^n$ that represents f on the domain $\{1,-1\}^n$ is nothing but the y^{th} Fourier coefficient of f. Notice that given a polynomial $\tilde{f}$ that represents f on the domain $\{0,1\}^n \subseteq \mathbb{R}^n$, the Fourier coefficients of f can be obtained by applying the change of variable $x_i \rightarrow \frac{1-x_i}{2}$, and $\bar{x}_i = (1-x_i) \rightarrow \frac{1+x_i}{2}$, to $\tilde{f}$; and finding the coefficients of the resulting polynomial in standard power form.

The finite vector space of functions from any subset S of $\mathbb{F}_2^n$ or of $\{1,-1\}^n$ embedded in $\mathbb{R}^n$ to $\mathbb{R}$ is denoted $\mathcal{F}_{2^n,S}$, and is equipped with the usual inner

product: for functions f and g, $< f, g >_S =_{def} 1/|S| \sum_x f(x)g(x)$. Sometimes the inner product is defined with respect to a distribution $\mathcal{R}$ over S; i.e, $\mathcal{R}(x) \geq 0$ for $x \in S$ and $\sum_x \mathcal{R}(x) = 1$. Then $< f, g >_{S,\mathcal{R}} =_{def} \sum_x \mathcal{R}(x)f(x)g(x)$. When S consists of all of the 2^n domain points, we simply omit the subscript S. The set of parity functions $\chi_u : u \in \mathbb{F}_2^n$ are mutually orthogonal in $\mathcal{F}_{2^n}$, but not neccessarily in $\mathcal{F}_{2^n,S}$, for arbitrary subsets S. However, these functions constitute a complete (possibly redundant) basis for $\mathcal{F}_{2^n,S}$, for any S. The norms are defined as usual: $||f||_{1,S} =_{def} \sum_{x \in S} |f(x)|$; and $||f||_{\infty,S} =_{def} max_{x \in S} |f(x)|$. However the $||.||_p$ norms for $1 < p < \infty$ are scaled by $1/|S|$, for convenience. For example, $||f||_{2,S} =_{def} \sqrt{< f, f >_S} = \sqrt{1/|S| \sum_{x \in \mathbb{F}_2^n} f(x)^2}$. The norms could also be defined with respect to a distribution $\mathcal{R}$ over S in the usual way. For example, $||f||_{2,S,\mathcal{R}} =_{def} \sqrt{< f, f >_{S,\mathcal{R}}}$ As usual, the norms $||.||_p$ and $||.||_q$ are said to be **dual** if $1/p + 1/q = 1$.

Clearly, $\mathcal{F}_{2^n,S}$ consists exactly of all multilinear polynomials over $S \subseteq \{1,-1\}^n \subseteq \mathbb{R}^n$. The subspace of this space formed by polynomials of degree bounded by d is called $\Pi^n_{d,S}$; the subscript S is often omitted. From the earlier discussion it is clear that this subspace is spanned by the basis parity functions $\{\chi_u : u \in \mathbb{F}_2^n, |u| \leq d\}$. For any subspace X of $\mathcal{F}_{2^n,S}$, the orthogonal space of functions $f \in \mathcal{F}_{2^n,S}$ satisfying $< f, g >_S = 0$, for all $g \in X$, is called $X^\perp_S$. *Notice that this is different from taking $X^\perp$ and restricting to S.* Thus $\Pi^{n,\perp}_d$ is nothing but the space spanned by the basis parity functions $\{\chi_u : u \in \mathbb{F}_2^n, |u| \geq d\}$, but this is not true over proper subsets S of $\mathbb{F}_2^n$. The space $X^\perp_{S,\mathcal{R}}$ can also be defined for a distribution $\mathcal{R}$ over S, by employing the inner product $< . >_{S,\mathcal{R}}$ in the definition of orthogonality. Given a subspace X and a function $f \in X$, $X \oplus f$ denotes the shifted set $\{g + f : g \in X\}$, $X \setminus f$ denotes the space $X \cap span(f)^\perp$ and $X \cup f$ denotes the space $X \cup (X \oplus f)$ Given a subspace X and a function f the function $f|_X$ is the projection of f on X. Thus $f = f|_X + f|_{X^\perp}$.

The Boolean functions in $\mathcal{F}_{2^n}$ with range $\{-1,1\}$ or $\{0,1\}$ form the vertices of the cubes, $\{1,-1\}^{2^n}$, or $\{0,1\}^{2^n}$ in $\mathbb{R}^{2^n}$. We will use the symmetries of these cubes to transfer results about certain Boolean functions to other Boolean functions. For example, the statement "*Parity* is not approximable by functions in Π_m" is equivalent to saying: "*One* is not approximable by functions in $\Pi^\perp_{n-m}$".

Complexity (resource) bounds on a function are always expressed in terms of the number of its variables. In the case of threshold complexity classes, the complexity of functions is given by the dimension of a good approximating space spanned by specific kinds of basis functions. Some common functions besides

Parity, *One* and the parity functions χ_s are the following: the functions $\wedge_{u,v}$ for disjoint $u, v \in \mathbb{F}_2^n$ are called the *And* functions, and are defined as

$$\wedge_{u,v}(x) := \bigwedge_{i \in u} x_i \bigwedge_{i \in v} \bar{x}_i;$$

i.e, when viewed as mapping from $\{1,-1\}^n$ to $\{1,-1\}$, $\wedge_{u,v}(x)$ takes the value -1 exactly when all the x_i's with $i \in u$ are -1's, and all the x_i's with $i \in v$ are 1's. The *Or* functions $\vee_{u,v}$ are defined analogously.

Some complexity classes that the paper deals with are the following. The class PT_1 (QT_1) consists of Boolean functions f that are approximable by a function g in the span of (quasi)polynomially many basis parity functions χ_s, with $||f-g||_\infty < 1$.
The class LT_1 consists of Boolean functions f that are approximable by a function g in the span of basis parity functions χ_s, with $|s| \leq 1$ and $||f-g||_\infty < 1$.
In general, LT_d is the class of Boolean functions f that are approximable by a function g in the span of polynomially many basis functions in LT_{d-1}, with with $||f-g||_\infty < 1$. The class $\hat{LT}_d$ is the class of Boolean functions f that are approximable by a function g which is in the span of polynomially many basis functions g_i in $\hat{LT}_{d-1}$, with the additional condition that when $g = \sum_i a_i g_i$, the coefficients a_i are normalized to $\sum_i |a_i| \leq 1$, and each a_i is a rational whose denominator is polynomially bounded. It should be noted that often in the literature, the normalization of $\sum_i |a_i|$ is removed, the condition $||f-g||_\infty < 1$ is simply written as $sign(f) = sign(g)$, and the a_i's are taken to be polynomially bounded integers. Finally, $AC^0[d]$ is the class of functions computable by (constant) depth d $\{\wedge, \vee, \neg\}$-circuits of polynomially bounded size.

3. Analytic framework using linear approximation

In this section, we develop an analytic framework based on linear approximation and point out how several questions concerning complexity lower bounds, pseudorandomness and learning algorithms fit naturally into this framework. This clarifies the relationship between the analytic structure of these questions. Therefore, the framework facilitates a systematic study of analytic methods which reduce these questions to their combinatorial essence, if there is any: in

some cases, the analytic methods alone are sufficient to answer the question at hand, as will be seen in this discussion.

This section is divided into four subsections. The first introduces the duality principle for linear approximation and explains its relevance to complexity questions. The second subsection presents a general analytic framework based on duality, by which one can view various complexity-theoretic questions as versions of the same analytic problem. The third describes the basic ingredients that constitute the analytic techniques for Boolean approximation, and gives examples that illustrate the pattern of exposition in the remaining sections.

3.1. The duality principle. The duality principle for any finite dimensional space of functions is the following. This can be found in any book on approximation theory. See [62] and [13], for a general treatment.

THEOREM 3.1. *Let U be a finite dimensional vector space of finite functions, with inner product $< f_1, f_2 >:= u \sum_x f_1(x)f_2(x)$, and 2-norm defined as $||f||_2^2 :=< f, f >$. Furthermore, let X be a linear subspace of U. For any function $f \in U$,*

$$\min_{g \in X} \|f - g\| = \max_{\substack{l \in X^\perp \\ \|l\|_* \leq 1}} |\sum_x l(x)f(x)|,$$

where $\|.\|_*$ is the dual norm of $\|.\|$.

REMARK 3.2. *The RHS of the above equality can be viewed as the maximum of $|L(f)|$ over all linear functionals L in the dual space that annihilate all the functions in X (i.e, $L(g) = 0$, for all $g \in X$), and have bounded norm. In fact, the RHS of the general duality principle for arbitrary normed linear spaces is such a supremum over linear functionals. See [13].*

PROOF. The dual norm $||.||_*$ is defined as:

$$||l||_* = \max_{f \in U} \frac{|\sum_x l(x)f(x)|}{||f||};$$

We first show that LHS $\geq$ RHS. For $g \in X$, and $l \in X^\perp$,

$$\sum_x l(x)(f - g)(x) = \sum_x l(x)f(x).$$

Thus, for all $f \in U$, $g \in X$, and $l \in X^\perp$,

$$|\sum_x l(x)f(x)| \leq ||l||_* ||f - g||,$$

by definition of the dual norm $||.||_*$, and

$$||l||_*||f-g|| \leq ||f-g||,$$

since $||l||_* \leq 1$.

To show the inequality in the other direction, for each $h \notin X$, we will exhibit a function l_h such that

$$|\sum_x l_h(x)h(x)| = \min_{g \in X} ||h-g||,$$

and additionally show that l_h satisfies the required conditions. Let $h^* \in X$ be such that $||h-h^*||$ is minimized. Define

$$l_h := \frac{||h-h^*||u(h-h|_X)}{||h-h|_X||_2^2}.$$

Writing h as $h-h|_X + h|_X$, and noticing that $< (h-h|_X), h|_X >= 0$, we get that

$$|\sum_x l_h(x)h(x)| = ||h-h^*||.$$

Furthermore, for a general $f \in span(X \cup \{h\})$,

$$\sum_x l_h(x)f(x) = \frac{< f, h-h|_X > ||h-h^*||}{||h-h|_X||_2^2}.$$

Now, for any $f \in X$, it holds that $< f, h-h|_X >= 0$, since $h-h|_X \in X^\perp$; therefore $l_h \in X^\perp$, thus satisfying the first required condition. Next, we establish that $||l_h||_*$ is bounded by 1 over $span(X \cup \{h\})$; l_h can clearly be modified and extended to the remainder of U without increasing the norm, (finite dimensional version of the Hahn-Banach theorem).

$$||l_h||_{*,span(X \cup \{h\})} \leq \max_{f \in span(X \cup \{h\})} \frac{|\sum_x l_h(x)f(x)|}{||f||}.$$

But

$$f = \frac{< f, h-h|_X > (h-h|_X)}{||h-h|_X||_2^2} + f|_X.$$

Substituting the expression for l_h as well, we obtain that

$$||l_h||_* \leq \frac{||h-h^*||}{||h-g||},$$

where $g = h|_X + cf|_X$, for some scalar c. But since h^* is chosen to be the best approximation for h from X, it follows that the above quantity is at most 1, thereby satisfying the second required condition on l_h. □

As a consequence the duality principle, the value of the correlation of any appropriately bounded function $l \in X^\perp$ with a function f gives a *lower* bound on the distance from f to its best approximation in X, and viceversa, the distance $\|f - g\|$ for any $g \in X$ gives an *upper* bound on the correlation for any bounded function $l \in X^\perp$. We now explain the relevance of the duality principle to various complexity questions, in a series of paragraphs enclosed by ◯'s.

◯ To show a complexity lower bound, i.e, to show that a function f is not in a complexity class C, one chooses an appropriate space X and a norm $||.||$, such that there is a good approximation from X to every function in C; then, one proceeds to show that there is a function $X^\perp$ of bounded norm that has a high correlation with f. Such a function is usually also a good approximation to f, and hence the lower bound comes down to showing the *existence of an approximation from $X^\perp$ to f.*

A number of lower bound results including many of those in [11], [15], [42], [43], [26], [33], [41], [44], [42], [23], [32], [38], [56], [25], [29], [4], and [24] can be phrased in the above form as will become clear during the course of this paper. See especially the next subsection and Sections 4 and 6. ◯

◯ Notice that if the functions in C have good approximations from X, then every function in $X^\perp$ has a small correlation with every function in C. This has several useful consequences.
The Boolean functions in $X^\perp$ are clearly natural hard functions for C.
Furthermore, assuming, for ease of exposition, that functions in C have a zero expected value, any positive function in $X^\perp$ could serve as a distribution that functions in C cannot distinguish from the uniform distribution, in the narrow sense of expected values. Elements drawn from these distributions serve as pseudorandom elements that "fool" randomized algorithms with complexity C. This makes a relationship between hardness and randomness (see [50]) transparent, and shows how to systematically characterize sets of pseudorandom strings and their generators for complexity classes: see Section 5 ◯

◯ Next, if the functions g in C have good approximations from X, then it is intuitively clear that $dim(X)+1$ independent pieces of information about such a function g should be adequate to find the approximation to g from X. If,

in addition, these independent pieces of information can be obtained by simply sampling g on a distribution supported at $dim(|X|+1)$ points, this would provide a small sampling distribution for approximating functions in the class C, which could potentially be converted to a fast approximation algorithm for functions in the class C, if the distribution, is, in addition, easy to generate. If, moreover, the approximation uniquely defines a (class of) function(s) in C (that coincide on a subdomain of large measure), then the algorithm is a learning algorithm, with "membership queries," for C. This is the idea behind the approximation-based learning algorithms of [46], [22] [40], [34], and [64] although not stated as such. Moreover, natural candidates for these sampling distributions are those that look uniform to functions in C. "Uniform" could be replaced by any other distribution as well, depending on the required approximation as will be seen in the next subsection and Sections 5 and 7, but for the moment, we restrict ourselves to uniform for the sake of exposition. We have seen in the previous paragraph that functions in $X^{\perp}$ can be modified to serve as such "close-to-uniform" distributions in the narrow sense of expected values. Already this was adequate in the case of the class C being AC^0, i.e, it was shown in [64] that in fact any set of pseudorandom strings that fool randomized circuits in AC^0 can also serve as a sampling set for approximating (and thereby learning) functions in AC^0. Since many of these sampling sets are easy to generate, they permit derandomization of AC^0 computations, as well as derandomization of learning algorithms for AC^0.

In fact, we will see in Section 7 that also for general classes C, functions in $X^{\perp}$ can be used to provide deterministic sampling sets for derandomizing computations in C as well as for approximating or learning functions in C. ◯

◯ Finally, a complexity upper bound of C can be shown by showing approximability from X. This is a promising direction to investigate, especially in the case of communication and threshold complexity upper bounds; however, as mentioned in the Section 1, with few exceptions in Section 6, our interest in approximability here is geared towards the eventual goal of showing non approximability. ◯

3.2. General analytic framework. In using the linear approximation framework of the last subsection to prove computational complexity lower bounds, etc., one could choose the approximating space X, and the norm $||.||_{\infty}$ appropriately to suit the problem at hand. In this section, we present a more general approximation framework that retains the main structure, but allows the choice of other parameters besides the approximating space and the norm. Thus, not only can a broader variety of questions be made to fit into this framework, but

also a finer differentiation between questions can be imposed.

We deal with the following kind of approximability of a Boolean function g, from a class B of Boolean basis functions, all of which are in the universal space U of functions.

$$\exists M \subseteq \mathbb{B}, with |M| \leq m, \exists(\forall) \text{ distributions } \mathcal{D} \text{ on } \mathbb{F}_2^n \text{ or } \{1,-1\}^n$$

$$\text{that induce a measure on subsets S of } \mathbb{F}_2^n,$$

$$\exists(\forall) \text{ subsets } S \subseteq \mathbb{F}_2^n \text{ with measure greater than } \sigma$$

$$\exists h \in span(M) \text{ with } ||g-h||_S \leq \gamma. \qquad \mathbf{A}$$

where s, $0 \leq \sigma \leq 1$ and $0 \leq \gamma \leq 1$ are some fixed constants, and m is typically a polynomial bound. (Notice that when $\gamma = 0$, we are dealing with a strict version of approximability, namely interpolability).

To show the corresponding non approximability of a hard function f, one shows:

$$\forall M \subseteq B, with |M| \leq m, \forall(\exists) \text{ distributions } \mathcal{D} \text{ on } \mathbb{F}_2^n \text{ or } \{1,-1\}^n$$

$$\text{that induce a measure on subsets S of } \mathbb{F}_2^n,$$

$$\forall(\exists) \text{ subsets } S \subseteq \mathbb{F}_2^n \text{ with measure greater than } \sigma$$

$$\exists l \in spanM_S^{\perp}, \text{ with } ||f||_S \leq 1 \text{ and } \sum_{x \in S} f(x)l(x) \geq \gamma. \qquad \bar{\mathbf{A}}$$

We enumerate all the parameters (besides m, σ and γ) and show how they can be appropriately chosen to fit several complexity-related problems.

- The choice of quantifiers in the definition of approximability: for example, the quantifiers on the the distributions $\mathcal{D}$, the subsets S, and the approximation $l \in span(M)^{\perp}$ can be chosen appropriately, or completely removed (i.e, the quantified parameter can be fixed to a constant). For example, for many complexity lower bound questions, the distribution $\mathcal{D}$ is fixed to be the uniform distribution and the subset S is sometimes fixed to be the entire domain, i.e, $\sigma = 1$. Sometimes, S is existentially quantified, with σ bounded away from $1/2$, which is often essential when the issue is interpolability, i.e, when $\gamma = 1$, as, for example, in the case of [58] and [66].

  When the approximation of interest is in the ∞-norm, as, for example,

for threshold circuit complexity lower bounds, the universal quantifier is implicit for the distribution $\mathcal{D}$: the approximation h in $\mathbf{A}$ must have the same sign as g everywhere. This fact has been extensively used in [26], [42], [44] and [43] and the usefulness of this universal quantifier is condensed in Theorem 6.14.

Notice also that a "PAC" learning algorithm, for functions g in a class C, is an algorithm for finding an approximation h, in $\mathbf{A}$, where the quantifier on the distribution is a universal quantifier and the quantifier on the subset S is an existential quantifier.

In some cases, for example, the monotone lower bound of [78], a *specific* function l in $\bar{A}$ is explicitly constructed to have a a high correlation with the given hard function f. In that particular case, the set from which l is chosen is not exactly a linear space of the form $span(M)^{\perp}$, but this assumption is reasonable, both intuitively, and for ease of exposition. Now, to show that f is not in the complexity class C, instead of showing that every function g in C has a poor correlation with *every* function in $span(M)^{\perp}$, or dually speaking, that g has an approximation from $span(M)$, it is sufficient to show that the correlation of g with the *specifically* constructed function l is small.

- The choice of the universal space U of all functions: this can be chosen, for example, to be functions from $\{-1,1\}^n$ to $\mathbb{R}$, as is usually the case, or from $\mathbb{F}_p^n$ to $\mathbb{F}_p$ (for example, as in [58], [66], [6], [67], [9], [68], [71], [77], as well as some of the results in [42], and [44]). In the latter case, γ is chosen to be 0 and $\sigma < 1$ in $\mathbf{A}$, i.e, one is interested only in the interpolability question over large subdomains, since finer notions of approximation do not make sense. Furthermore, the duality 3.1 and hence $\bar{\mathbf{A}}$ do not apply, since inner products and orthogonality are not well-defined for such spaces of functions.

  NOTE. In this paper, however, we restrict ourselves to functions from $\{-1,1\}^n$ to $\mathbb{R}$.

- The choice of the basis functions B: for example, these are the monomials from from $\mathbb{F}_p^n$ to $\mathbb{F}_p$ in the case of [66], and the set M is taken to be just the low degree monomials, whereby $span(M)$ is the space of all low degree polynomials. In some of the results on threshold circuit complexity in [42], [15], [26], and others in [46], [40], [34], [64], the basis

functions in $\mathbb{B}$ are, in effect, the *Parity* functions or $\mathbb{Z}_r^n$ characters; in the case of [43], the *And* functions; in [23], AC^0 functions; in [44], the basis functions are the symmetric functions; in the case of [33] and [26], the $\hat{LT}_1$ functions, and, in effect, characteristic functions of cross-product sets; and in general, in the papers [32], [38], [56], [25], [29], [4], [24], that deal with communication complexity, the basis functions are characteristic functions of cross-product sets. Many of these results, however, are not stated as nonapproximability results. Finally, several of the threshold circuit upper bounds (see [48] and [49]), are in fact approximability results from the spans of various sets of basis functions, although not phrased as such.

- The choice of norms and inner products: the most common norms are the 2-norm and the ∞-norm, and in several situations for Boolean functions, these are highly related, as we shall see in Section 6.

  Sometimes, the distribution $\mathcal{D}$ is included in the norm. For example, the 2-norm could be defined based on the inner product $<>_{\mathcal{D}}$.

- The linear approximating space $span(M)$ in $\mathbf{A}$ is often replaced by a convex polytope bounded by linear facets in $\mathcal{F}_{2^n}$, for example, when one restricts the coefficients of the linear combination that forms the approximation h to be polynomially bounded, as in the case of "unweighted" threshold circuit complexity, or positive, as is the case when g is a function of low communication complexity and the basis functions are cross-product functions; or if one requires the function l in $\bar{\mathbf{A}}$ to be a distribution, and hence to be positive, which happens when one is interested in a distribution that yields pseudorandom strings, as explained in the previous subsection on duality. These restrictions on the approximation nevertheless permit usual linear approximation methods: for example, modified versions of duality apply, as will be seen in Sections 5 and 6.

  Sometimes, however, the space $span(M)$ is replaced by a truly non-linear and non-convex structure, for example, as in the case of some (mainly monotone) lower bounds based on the approximation method of [57], and [59], such as [78], [39], [2], and [10]. However, even when the set of approximating functions has no linear structure, some of the general linear approximation methods can nevertheless be adapted. For example, to show that a particular function f is not in a class C, [78] chooses an approximating set M of functions as certain sums and products of

simple functions, and explicitly constructs a function l that is almost in $span(M)^{\perp}$ (i.e, it is almost orthogonal to every function in M), and $|\sum_x l(x)f(x)| = 1$ for the given function f. This establishes, via duality, that f is not closely approximable from the from the space $span(M)$. In fact, we will see in Section 6 that the existence of such a function l is equivalent - by one of the modified versions of duality mentioned in the last paragraph - to the statement that the sign of f is not obtainable by a any small-coefficient linear combination of functions in M. Thus, in particular, f is not approximable by any single function in M. Therefore, this part of the proof in [78] is based only on linear approximation ideas. To complete the proof that $f \notin C$, it is shown (as noted in the paragraph on the "choice of quantifiers") that the correlation of every function $g \in C$ with the *specifically* constructed function l is small. This is, however, done using a non-linear approximation technique of constructing a sequence of functions starting from a function in M, which we know to be almost orthogonal to l, and ending in g, in such a way that the functions in the entire sequence remain almost orthogonal to l. It is an open question whether the non-linear structure formed by the approximating functions in M can be replaced by a suitable linear space, by placing appropriate restrictions on the approximation.

Note that the general methods and results in this paper are meant to be used *after* a complexity question has been phrased as a Boolean approximation question as in **A**, i.e, all the choices described above have already been made, which is a nontrivial process.

The next three sections of the paper are organized based on the type of approximation chosen in **A**, which are broadly classified as:
• Interpolation (when the quantity γ in **A** is chosen to be 0),
• One-sided approximation (when the functions l in $\bar{\mathbf{A}}$ are forced to be positive and can be treated as distributions), and
• General uniform approximation (where the norm in **A** is usually chosen to be ∞-norm, but sometimes also to be the 2-norm).
The final section concerns
• Algorithms for finding the approximation h in **A**.

3.3. Basic ingredients of analytic techniques. We describe three basic ingredients that constitute most analytic methods specific to Boolean (non) approximability, including those in this paper.
• strong versions and modifications of the duality principle

• simple norm relationships specific to Boolean functions
• essentially univariate analytic techniques that apply in very specific multivariate settings. In providing some introductory examples of how these ingredients are used, we follow the pattern of exposition to be used in the remaining sections. We reiterate that the aim of the methods in this paper is to recognize and deal with the analytic content of the problem, thereby reducing it down to its combinatorial essence, at which point possibly adhoc combinatorial methods specific to the problem might be required.

Modifications of the duality principle

Several modifications of the duality principle are used in this paper, for example, Theorem 6.1, Theorem (interp), and Theorem (one-sided); these form the backbone for several previous complexity bounds as will be seen. Strong versions of the duality principle such as the one below depend on the Booleanness of the function to be approximated. Others such as 6.1(2) depend on the Booleanness of the approximating basis functions.

THEOREM 3.3. *Let U be a finite dimensional vector space of functions and let X be a linear subspace of U. For any Boolean function $f \in U$, the following are equivalent.*
• *There is no approximation $g \in X$ which is non-zero everywhere and has the same sign as f.*
• *There is a non-zero approximation l in $X^{\perp}$ with the same sign as f on $supp(l)$.*

PROOF. The first statement is equivalent to saying that for all $g \in X$, $||f - g||_\infty \geq 1$. Since the dual norm of the ∞-norm is the 1-norm, this is equivalent by 3.1 to saying that there is some $l \in X^{\perp}$, with $||l||_1 = 1$ such that $\sum_x l(x)f(x) \geq 1$. Now, since f is Boolean, this happens exactly when l has the same sign as f on $supp(l)$.

Notice that the theorem above converts a statement of non approximability into a statement of approximability, and depends on the Booleanness of f. In fact, due to the Booleanness of f, a geometric proof exists for the above result. This is given below, following the proof of 3.4. The first statement in the above theorem is also equivalent to a notion of orthogonality with respect to the $||.||_\infty$ norm, according to which a function f is orthogonal to a space X, if for every function $g \in X$, $||f + g||_\infty < ||f||_\infty$ (see [62] for an succinct treatment of $||.||_\infty$ norm, or uniform approximation).

Next, we give two introductory example applications of the general result of Theorem 3.3.

THEOREM 3.4. *The statement that for a Boolean function f, there is a function g in X with the same sign as f is equivalent to the following. For every f^* with $||f^*||_\infty \leq 1$, the projections $f^*|_X$ and $f|_X$ are identical implies $f^* = f$.*

PROOF. First notice that the functions f^* with the property that $f^*|_X = f|_X$ are exactly those that belong to $X^\perp \oplus f$, i.e, they are of the form $l + f$ for some l in $X^\perp$. Now the proof follows easily from duality for Boolean functions.

$$\exists g \in X \quad \text{and} \quad ||f - g||_\infty < 1 \quad \Longleftrightarrow$$

(by 3.1)

$$\forall l \in X^\perp \quad \text{with} \quad ||f||_1 \leq 1, \quad |\sum_x l(x)f(x)| < 1.$$

This is equivalent to saying that every non-zero function l in $X^\perp$ agrees in sign with f on atleast one point, and disagrees on atleast one point. This is, in turn, equivalent to saying that

$$\forall l \in X^\perp, \quad ||l + f||_\infty > 1 \quad \text{or} \quad l = 0$$

$$\forall f^* \quad \text{with} \quad ||f^*||_\infty \leq 1, f^*|_X = f|_X \Rightarrow f^* = f$$

since, f^* satisfying $f^*|_X = f|_X \Rightarrow f^* = f$, as observed in the beginning of the proof, must be of the form: $l + f$ for some l in $X^\perp$. □

The following is a *geometric* illustration of the the above proof, and in fact contains a geometric proof of 3.3 as well. The argument depends strongly on the fact that the function being approximated is Boolean.

View the Boolean functions in $\mathcal{F}_{2^n}$ as the vertices of the cube $\{-1, 1\}^{2^n}$. This cube is, in fact, the unit $||.||_\infty$ ball in $\mathbb{R}^{2^n}$, thus the points in the interior of this cube represent vectors (functions) with $||.||_\infty < 1$.

Spaces of functions such as X are plane/subspaces through the origin, and those such as $X \oplus f$ are shifted plane/subspaces through the vertex f of this cube. Now,

$$\exists g \in X \quad \text{and} \quad ||f - g||_\infty < 1 \quad \Longleftrightarrow$$

$$\exists g \in X \oplus f \quad \text{with} \quad ||g||_\infty < 1 \quad \Longleftrightarrow$$

The plane/subspace $X \oplus f$ cuts through the interior of the cube $\{-1, 1\}^{2^n}$ $\Longleftrightarrow$
(by the orthogonal geometry of cubes)
The orthogonal plane/subspace $X^\perp \oplus f$ touches the cube $\{-1, 1\}^{2^n}$ only at f
$\Longleftrightarrow$

$$\forall l^* \in X^\perp \oplus f, \quad ||l^*||_\infty > 1 \quad \Longleftrightarrow$$

$$\forall l \in X^{\perp}, \quad ||l+f||_{\infty} > 1 \quad \text{or} \quad f^* = 0 \quad \Longleftrightarrow$$
$$\forall f^* \quad \text{with} \quad ||f^*||_{\infty} \leq 1, f^*|_X = f|_X \Rightarrow f^* = f,$$

by the same argument as used in the above proof.

◯ The characterization result in [15] for PT_1 functions states the forward direction of 3.4 for the special case where X is given as the span of orthonormal parity functions χ_s. They use the result to show a bound of $2^{n.dim(X)}$ on the number of distinct PT_1 functions, since the number of distinct Fourier coefficient values of a Boolean function is at most 2^n. The above result gives a bound on the number of LT_2 functions as follows. Let M be any set of m independent LT_1 functions. Since there are at most $2^{O(n^2)}$ LT_1 functions, there are atmost $\binom{2^{O(n^2)}}{m}$ such sets. For any Boolean function f, $f|_{span(M)}$ can be uniquely written as $\sum_{h \in M} a_h h$. Let b be a bound on the total number of distinct values that the coefficients a_h take. Then there are at most $\binom{2^{O(n^2)}}{m} b^m$ LT_2 functions with m LT_1 gates at the bottom level. A straightforward upper bound of $m^{m/2}$ exists on b since LT_1 functions are Boolean, although it is not clear that the bound is tight. See [1], [72] and [31], for related tight bounds on the inverses of Boolean matrices. ◯

THEOREM 3.5. *Scalar multiples of Parity are the only functions (Boolean or otherwise), over* $\{1,-1\}^n$ *whose sign does not coincide with that of any polynomial of degree bounded by* $n-1$.

PROOF. A Boolean f has no approximation g from Π_{n-1} with the same sign, if and only if, by Theorem 3.3, there is a function $l \in \Pi_{n-1}^{\perp}$ such that l has the same sign as g wherever $l \neq 0$. But $\Pi_{n-1}^{\perp} = span(Parity)$, and hence f must be a scalar multiple of $span(Parity)$. □

◯ It is folk-lore that *Parity* cannot be approximated by any by polynomial of degree bounded by $n-1$. That requires only the first line of the above proof. ◯

Simple norm relationships specific to Boolean functions We give a few examples. Notice that if f is Boolean, then

$$||f||_1 = \sum_{x \in \mathbf{F}_2^n} |f(x)| = \sum_{x \in \mathbf{F}_2^n} f^2(x) = 2^n ||\hat{f}||_2^2.$$

Thus, for Boolean f, bounds on the 2-norm of $\hat{f}$ provide bounds on the 1-norm of f, and furthermore, the 1-norm of f ($\hat{f}$) provides an upper bound on the (2^n times) the 2-norm of f ($\hat{f}$). In addition, the 1-norm of $\hat{f}$ is a lower bound on the size of the support of $\hat{f}$, and an upper bound on the sparsity of an approximating polynomial having the same sign as f, (as will be seen in Theorem 6.9), which, in turn, yields a lower bound on the L_∞ norm of $\hat{f}$ (as will be seen in Theorem 6.3). One more property that is specific to Boolean functions and is constantly used in Section 6 is the following: when a function g has the same sign of as a Boolean function f, then the 1-norm of g is the magnitude of (2^n times) the scalar product or correlation $< f, g >$.

Univariate techniques

Any multivariate function f can be "univariatized"

$$f^*(x_1 + \ldots + x_m) := 1/m! \sum_{\pi} f((\pi(x_1) \ldots \pi(x_m))),$$

where the π are permutations acting on the set of arguments of f. It is not hard to see that f^* has degree bounded by the degree of f. Furthermore, univariate analytic techniques can be used to prove properties of f^*, which sometimes transfer partially to f. This is advantageous since much is known about univariate approximation and far less about multivariate approximation.

REMARK 3.6. *It should be noted that most of the general results and methods in the remaining sections of the paper are inherently multivariate, and do not rely on univariate techniques. Nevertheless, for completeness, we provide here a brief discussion of univariate analytic techniques and their applications.*

Univariate methods are particularly useful for symmetric functions. For a symmetric function f, $f^*(x_1 + \ldots + x_m) = f(x_1, \ldots, x_m)$. Thus symmetric functions can be treated essentially like univariate functions. The results of [52], for example, take full advantage of this fact. This univariate nature extends, in practice, also to functions f satisfying: $f(x_1, \ldots, x_m) = f'(\lambda^T x)$, where f' is univariate, $\lambda \in \mathbb{R}^n$, and, over the cube $\{-1, 1\}^n$, it is natural to require that the size of the set $\{\lambda^T x : x \in \{-1, 1\}^n\}$ is appropriately, say polynomially bounded. Such functions are called "ridge functions" in the approximation theory literature. Notice that $\hat{LT}_1$ functions have this property. It is not surprising, therefore that many complexity bounds involving symmetric functions transfer to $\hat{LT}_1$ functions as well.

◯ The results of [51], [52], and some of [47] and [3], for example, rely almost entirely on univariate approximation, and deal with local combinatorial properties of general Boolean functions. However, the related results of [37] do

use inherently multivariate techniques such as Beckner's inequalities [7]. Many of these results can be phrased as general combinatorial questions about the multidimensional unit cube, see for example, [27], and [47]. ◯

Two classical univariate approximation techniques that have been used for proving results about Boolean functions are the Bernstein-Markov, and the complementary Jackson inequalites for which the reader is referred to any introductory book on approximation theory, for example [61], or [55].

The following non approximability result of [51] is used later in Section 5. It relies on a Bernstein-Markov inequality. The complementary approximability results of [52] for the case of symmetric functions rely on the Jackson inequalities, but are not presented here.

THEOREM 3.7. *The degree of a polynomial g that approximates a Boolean function f with $||f-g||_\infty \leq 1/3$ (any constant strictly less than 1/2) must be at least $\sqrt{sensitivity(f)/6}$, where the sensitivity of a Boolean function f is the quantity*

$$\max_x |\{y : x \oplus y = 1, \text{ and } f(x) \neq f(y)\}|.$$

PROOF. Consider a point x where f attains its maximum sensitivity, denoted $S(f)$, and let I be the set of $S(f)$ coordinates

$$\{i : \exists y \neq x \text{ s.t. } y_j = x_j \ \ \forall j \neq i, \text{ and } f(x) \neq f(y)\}.$$

We now restrict our domain to the $S(f)$-dimensional cube defined by the coordinates in I, with the rest of the coordinate values fixed identical to x. Notice that on this smaller cube, the value that f takes at the vertex x is different from its value at all the $S(f)$ neighboring vertices (i.e, points that differ on exactly one coordinate from x).

Therefore, to complete the proof, it is sufficient to show that when a function f over the vertices of $\{1,-1\}^m$ takes the value $+1$ at one vertex and the value -1 at all the neighboring vertices, then any g approximating f with $||g-f||_\infty \leq 1/3$ must have degree at least $\sqrt{m}/6$.

To this end, we take the well-known "univariatization" of f by defining:

$$f^*(x_1 + \ldots + x_m) := 1/m! \sum_\pi f((\pi(x_1 \ldots x_m))).$$

It is not hard to see that f^* has degree bounded by the degree of f and furthermore, for any p such that $||f-p||_\infty \leq 1/3$, the corresponding univariatization p^* also satisfies $||f^*-p^*||_\infty \leq 1/3$, and the degree of p^* is bounded by the

degree of p. It therefore suffices to show that any p^* that approximates f^* with $||f^* - p^*||_\infty \leq 1/3$ must have degree at least $\Omega(\sqrt{m})$.

To achieve this, first shrink the domain of f^* and p^* from $[-m, m]$ to $[-1, 1]$. Then, since by assumption $f^*(-1) = -1$, $f^*(-1+2/m) = +1$, $p^*(-1) \leq -2/3$, and $p^*(-1 + 2/m) \geq +2/3$, it follows by the mean value theorem that there must exist a $z : -1 \leq z \leq -1 + 2/m$ such that the derivative $|p^{*'}(z)| \geq 4n/6$. Furthermore, notice that $max\,|p^*(x)| \leq 4/3$ over all the points x in the set $J =_{def} \{-1, -1 + 2/m, \ldots, 0, 1 - 2/m, 1\}$. Now we apply the most straightforward version of the univariate Bernstein-Markov bound relating the $||p^{*'}||_\infty$, and $||p^*||_\infty$ over the interval $[-1, 1]$:

$$||p^{*'}||_\infty \leq \deg^2(p^*)||p^*||_\infty.$$

Now, by the mean value theorem, $||p^*||_\infty$ over $[-1, 1]$ can be bounded above by

$$\max_{x \in J} |p^*(x)| + ||p^{*'}||_\infty/n.$$

Thus

$$\deg^2(p^*) \geq n[\frac{||p^{*'}||_\infty/n}{(\max_{x \in J} |p^*(x)| + ||p^{*'}||_\infty/n)}].$$

Now, since $||p^{*'}||_\infty/n \geq 4/6$, and $\max_{x \in J} |p^*(x)| \leq 4/3$, we obtain that $\deg(p^*) \geq \sqrt{n}/6$, which completes the proof. □

Certain multivariate extensions of Bernstein-Markov, and Jackson theorems exist in the literature, but their usefulness for Boolean approximation is yet to be investigated, see, for example, [54]. In general, while there is a well-grounded and classical univariate approximation theory, multivariate approximation techniques are still very much under development, see for example [16], and [17].

4. Interpolation

We prove two general results, both of which are modified versions of the duality principle. The first equates a question of approximability in the ∞-norm to a question of interpolability, and the second equates non-interpolability to interpolability. Applications are given for both results.

THEOREM 4.1. *Let X be any subspace of functions (from $\{-1, 1\}^n$ to $\mathbb{R}$), and let $f \in X^\perp$. The following statements are equivalent.*

• *No $g \in X$ has the the same sign as f on a set S of more than m points.*
• *For every set $\bar{S}$ of at most $2^n - m$ points, there is a function $h \in X^\perp \setminus f$ (which is also $(X \cup f)^\perp$) that interpolates f on $\bar{S}$, and is bounded above by f elsewhere.*

REMARK 4.2. *Notice that the last sentence states that $h \leq f$; it may not hold that $||h||_\infty \leq ||f||_\infty$.*

PROOF. The first statement is equivalent, by duality, to the following. For any set S of greater than m points, there is a function h in $X_S^\perp$ such that $|f(x) - h(x)| \leq 1$ for $x \in S$, and $h(x)$ is not identically zero on S. Denoting the complement of (points outside) S to be $\bar{S}$, notice that $X_S^\perp = \{g \in X^\perp : g = 0 \text{ on } \bar{S}\}$. Therefore, the first statement can be restated as follows: for any set $\bar{S}$ of $< 2^n - m$ points, there is a function $h \in X^\perp$ such that $h = 0$ on $\bar{S}$, h is not identically 0 on S and $sign(f) = sign(h)$ whenever $h \neq 0$.

Now let $h = h_1 + h_2$, where $h_1 \in X^\perp \setminus f$ and h_2 is some scalar multiple of f. Note that h_1 and h_2 are orthogonal. Also, h_2 cannot be a negative multiple of f, because then h_1 would have the same sign as f everywhere that it is non-zero (at least at one point), which would imply that $< h_1, f > \neq 0$. But h_1 was chosen to be orthogonal to f, which creates a contradiction. Thus h_2 is a positive multiple of f, say $c * f$. Now the statement in the previous paragraph is equivalent to the following: for any set $\bar{S}$ of $< \sum_{k=0}^{l/2} \binom{n}{k}$ points, there is a function $-h_1/c \in X^\perp \setminus f$ that is f on $\bar{S}$, and is at most f elsewhere, which is exactly the statement of the theorem. □

◯ As a straightforward application of the above result, we give a statement of non-interpolability that is equivalent to a non approximability result of [3], which was proven using a standard univariate approximation technique (see discussion in Subsection 3.3 on univariate techniques).

THEOREM 4.3. *The following statements are equivalent.*
• *For any polynomial g over $\{1, -1\}^n$ of degree at most d, (i.e, $g \in \Pi_d$), $sign(g) = sign(Parity)$ on a set S of at most $\sum_{k=0}^{(n+d+1)/2} \binom{n}{k}$ points.*
• *Given any set $\bar{S}$ of at most $\sum_{k=0}^{l/2} \binom{n}{k}$ points, there is a function $h \in \Pi_l \setminus \{One\}$ that is equal to 1 (interpolates the constant function One) on all points in $\bar{S}$ and is at most One elsewhere.*

PROOF. After a direct application of the above theorem 4.1, with f taken as *Parity*, and $X = \Pi_d$, use the symmetry of the Boolean functions in $\mathcal{F}_{2^n}$ to replace *Parity* by *One* and $\Pi_d^\perp$ by Π_{n-d-1}. □

○

Next, we prove a general result that transforms a statement of non-interpolability to a statement of interpolability.

THEOREM 4.4. *Let X be a subspace of the usual space of functions and let f be any function in $X^\perp$. The following statements are equivalent.*
• *No function in X of degree $\leq d$ interpolates f on a set S of more than m points*
• *For every set $\bar{S}$ of $< 2^n - m$ points, there is a function in $X^\perp \setminus f$, that interpolates f on $\bar{S}$.*

PROOF. Take an orthonormal basis $B = B_1 \cup B_2 \cup f$, where B_1 is an orthonormal basis for X, and B_2 is an orthonormal basis for $X^\perp \setminus f$. Let H be the matrix whose rows b correspond to functions in B and whose columns x correspond to points $x \in \{-1, 1\}^n$. Thus, pairs of rows are orthogonal pairs in $\mathbb{R}^{2^n}$, which makes H an orthogonal matrix with $H^T H = I$, a fact that will be constantly used. The entry $H_{b,x}$ is simply $b(x)$. We will denote by B_1 the set of rows $b \in B_1$, and by B_2 the set of rows $b \in B_2$. We will refer to the remaining row as f. For any set S of columns and a set M of rows, let $H_{M,S}$ be the submatrix of H formed by those rows and columns. Furthermore, denote by f_S the f row vector restricted to the set S of columns. It is not hard to see that the statement we want to prove is the following.

$$\forall S : |S| > m,\ f_S \notin span(\text{rows of } H_{B_1,S}) \iff$$

$$\forall S' : |S'| < 2^n - m,\ f_{S'} \in span(\text{rows of } H_{B_2,S'}) \qquad A$$

We will show that assuming either the LHS of A is true and the RHS is false, or that the LHS is false and the RHS is true lead to contradictions.

First, assume the LHS is true and the RHS is false. I.e, there is some set S of columns such that $f_S \notin span(\text{rows of } H_{B_1,S})$, and denoting by S' the *remaining rows*, $f_{S'} \notin span(\text{rows of } H_{B_2,S'})$. These imply that

$$\begin{bmatrix} f_S \\ H_{B_1,S} \end{bmatrix} * \begin{bmatrix} g_S \end{bmatrix} = \begin{bmatrix} 1 \\ 0^{|B_1|} \end{bmatrix}$$

has a non-zero solution for g_S and and that

$$\begin{bmatrix} f_S \\ H_{B_2,S'} \end{bmatrix} * \begin{bmatrix} g_{S'} \end{bmatrix} = \begin{bmatrix} 1 \\ 0^{|B_2|} \end{bmatrix}$$

has a non-zero solution for $g_{S'}$. Here, 0^n denotes a $n \times 1$ column vector of 0's.

This, in turn, implies that

$$\begin{bmatrix} f_{S'} & f_S \\ H_{B_2,S'} & H_{B_2,S} \\ H_{B_1,S'} & H_{B_1,S} \end{bmatrix} * \begin{bmatrix} 0^{|S'|} \\ g_S \end{bmatrix} = H * \begin{bmatrix} 0^{|S'|} \\ g_S \end{bmatrix} = \begin{bmatrix} 1 \\ H_{B_2,S} g_S \\ 0^{|B_1|} \end{bmatrix}$$

has a non-zero solution for g_S and that

$$H * \begin{bmatrix} g_{S'} \\ 0^{|S|} \end{bmatrix} = \begin{bmatrix} 1 \\ 0^{|B_2|} \\ H_{B_1,S'} g_{S'} \end{bmatrix}$$

has a non-zero solution for $g_{S'}$. Multiplying both sides of the the two equations above, we get

$$[\, g_S \quad 0^{|S'|} \,] * H^T H * \begin{bmatrix} g_{S'} \\ 0^{|S|} \end{bmatrix} = 0.$$

But since H is an orthogonal matrix, $H^T H = I$, and thus we obtain a contradiction.

Next, we assume that the LHS of A is false and that the RHS is true. I.e, there is some set S of columns such that $f_S \in span(\text{rows of } H_{B_1,S})$, and denoting by S' the *remaining rows*, $f_{S'} \in span(\text{rows of } H_{B_2,S'})$. Let H_i be the i^{th} row of H; we can say that there are weights a_i such that

$$f + \sum_{i \in B_1} a_i H_i = \begin{bmatrix} 0^{|S'|} \\ *^{|S|} \end{bmatrix}$$

and

$$f + \sum_{i \in B_2} a_i H_i = \begin{bmatrix} *^{|S'|} \\ 0^{|S|} \end{bmatrix}.$$

Here the $*$'s represent indeterminates. Now, multiplying both sides of the above two equations, all terms on the left except for $f^T f = 1$ vanish, because of the orthogonality of H, and the right side is 0, causing a contradiction. □

◯ To apply this result, we turn to a well-known theorem of [66]. For the moment, we view the domain ($\{1, -1\}^n$) as subsets of $\mathbb{F}_3^n$ (or a vector space

over any finite field $\mathbb{F}$), with *with -1 mapping to* $1_{\mathbb{F}_3}$ *and 1 mapping to* $0_{\mathbb{F}_3}$, and similarly the range $\{1,-1\}$ of Boolean functions is also viewed as $\{0_{\mathbb{F}_3}, 1_{\mathbb{F}_3}\}$. Thus the function $Parity(x)$ evaluates to $1_{\mathbb{F}_3}$ if $|x|$ is odd, and to $0_{\mathbb{F}_3}$ otherwise.

We denote by Π^3_d the space of polynomials of degree d, over $\{0_{\mathbb{F}_3}, 1_{\mathbb{F}_3}\}^n$. The well-known result of [66] is the following.

THEOREM 4.5. *The Parity function is interpolable from* $\Pi^3_{\sqrt{n}}$ *on at most* $2^{n-1} + o(2^n)$ *points.*

In the terminology used by [66], this is equivalent to saying that *Parity* is "$U^n_{\mathbb{F}_3}$-complete". The complementary result in [66] concerns the "nearly-$\mathbb{F}_3$-easiness" of the functions circuits computed by constant depth, subexponential size circuits of $\vee$, $\wedge$ and $\neg$ gates; this notion captures the interpolability of such functions on large domains by low degree polynomials. The two results together provide a lower bound of $2^{n^{1/2d}}$ on the size of $AC^0[d]$ circuits computing *Parity*. To obtain finer approximation bounds such as those obtained by using the switching lemma [30], [66] suggests the open problem of using $\mathbb{R}$ instead of finite fields such as $\mathbb{F}_3$, formulating analogous notions such as "$U^n_{\mathbb{R}}$-completeness," and "nearly-$\mathbb{R}$-easiness." and obtaining analogous results based on these notions.

REMARK 4.6. *We now switch back to our prevailing custom of viewing* $\{1,-1\}$ *as a subset of* $\mathbb{R}$ *and of* $\mathbb{F}_2$*, the Parity function as taking values in* $\{1,-1\}$*, and polynomials as being over* $\mathbb{R}^n$.

Intuitively, "nearly-$\mathbb{R}$-easiness" would imply a form of approximability over large domains by low degree polynomials over $\mathbb{R}^n$ with respect to a chosen norm. Similarly, "$U^n_{\mathbb{R}}$-completeness" would mean a form of non approximability over large domains by low degree polynomials.

In this sense, a version of the $U^n_{\mathbb{R}}$-completeness of *Parity* has been proven by the result of [3] in 4.3. However, no complementary result has been established for the "nearly-$\mathbb{R}$-easiness" of constant depth subexponential size circuits of $\vee$, $\wedge$ and $\neg$ gates. Therefore a stronger version of $U^n_{\mathbb{R}}$-completeness for *Parity* is desirable, such as the the following conjecture by [45], since this would make it sufficient to prove a weaker "nearly-$\mathbb{R}$-easiness" result.

CONJECTURE 4.7. *Any polynomial that interpolates Parity on more than* 2^{n-1} *points must have degree* $\Omega(\sqrt{n})$.

Now Theorem 4.4 shows that proving the above conjecture is equivalent to proving one of two statements of *interpolability*.

THEOREM 4.8. *The above conjecture is equivalent to the following statements.*

• *For every subdomain of* $< 2^{n-1}$ *points in* $\{1,-1\}^n$, *there is a polynomial in* $\Pi^{\perp}_{o(\sqrt{n})} \setminus$ *Parity that interpolates Parity on the subdomain.*
• *For every subdomain of* $< 2^{n-1}$ *points, there is a polynomial in* $\Pi^{\perp}_{n-o(\sqrt{n}} \setminus$ *One that is constant on the subdomain.*

PROOF. The former sentence is a corollary of Theorem 4.4 and the latter sentence is equivalent to the former because of the symmetries of the the cube of Boolean functions in $\mathcal{F}_{2^n}$. □

◯

5. One-sided approximation

As the main general result of this section, Theorem 5.6 we prove a version of the duality principle involving the notion of one-sided approximation [18]. As one application, we obtain a systematic method of characterizing distributions that look uniform to (and hence "fool") any function that is approximable from a given space of functions.

The first general result follows directly from Theorem 3.1, and leads us to consider a slightly strengthened version of a conjecture of [47]. The main result, Theorem 5.6, is motivated by showing that this modified conjecture is false.

THEOREM 5.1. *If a Boolean function* f *has an approximation* g *in a subspace* X, *such that* $||f-g||_\infty \leq \epsilon$, *then for any positive* $l \in (X \setminus One)^{\perp}$ *with* $||l||_1 = 1$, $|\sum_x l(x)f(x) - 1/2^n \sum_x f(x)| \leq 2\epsilon$.

PROOF. Notice that $l^* = l - ||l||_1/2^n \in X^{\perp}$, with $||l^*||_1 \leq 2$. Then, by 3.1, $|\sum_x l^*(x)f(x)| \leq 2\epsilon$, which completes the proof. □

◯ In general, the above theorem shows that if every function in a complexity class C is approximable in the ∞-norm to within, say $1/n$, from a subspace X, then any positive $h \in (X \setminus One)^{\perp}$ forms a distribution that looks uniform to (and hence fools) every function in C. Furthermore, any function $h \in X^{\perp}$ can be expressed as $h^* + c.One$, where h^* is positive and in $(X \setminus One)^{\perp}$. Clearly h is hard for C by 3.1, and h^* gives a distribution that fools every function in C. Such distributions that are, in addition, easy to generate, provide a large class of natural pseudorandom generators for all computations in C, and

highlights the close relationship between hardness and pseudorandomness (see, for example, [50]). We will see in Section 7 that these distributions also serve as sample sets for derandomizing learning algorithms for the class C. ◯

◯ Next, we consider the special case where the class C above is AC^0. The conjecture of [47] states that all polylog-wise independent distributions fool AC^0 functions. This conjecture has not been settled, although partial positive results appear in [64]. It is not hard to see (the reader is referred to [64]) that an e-wise independent distribution l is nothing but a positive function l with $||l||_1 = 1$ and $l \in (\Pi_e \setminus One)^\perp$. Thus, the conjecture can be restated as follows.

CONJECTURE 5.2. *[47] For all functions f computed by an $AC^0[d]$ circuit of size $s(n)$, there is an e that depends only on s and d and is polylogarithmic in n (if s is polynomial in n and d is a constant), such that for every positive function $l \in \Pi_e^\perp \cup \{One\}$ with $||l||_1 = 1$,*

$$|\sum_x l(x)f(x) - 1/2^n \sum_x f(x)| \leq 1/n. \qquad I$$

To settle this conjecture, using Theorem 5.1, it would be *sufficient* to show that all functions f in AC^0 have an approximation $g \in \Pi_e$, where e is an appropriate polylogarithm in n, such that $||f - g||_\infty \leq 1/n$. Unfortunately, the latter statement is false, as is shown below.

FACT 5.3. *Let $RO[2](x) := \bigvee_{i=1}^{\sqrt{n}} \bigwedge_{j=1}^{\sqrt{n}} x_{ij}$. Then an function g such that $||f - g||_\infty \leq 1/3$ must have degree at least $\Omega(n^{1/4})$.*

PROOF. The proof of Fact 5.3 follows from 3.7 and the fact that the function $RO[2]$ has sensitivity $\geq \sqrt{n}$, for example, at the minterms and maxterms. □

Since the converse of Theorem 5.1 does not hold, the above non approximability result for AC^0 functions does not falsify the conjecture 5.2. However, using a more direct application of 3.1 we observe below that even a slight strengthening of the conjecture 5.2 is false.

REMARK 5.4. *To fully understand the relationship between 5.2 and the observation below, notice that the function l^* defined as $l^*(x) := l(x) - 1/2^n$ - where l is the function in 5.2 - is in $\Pi_e^\perp$, since $\Pi_e^\perp$ and span(One) are orthogonal spaces. Thus the quantity in 5.2 (I) becomes $|\sum_x l^*(x)f(x)|$, with $||l^*||_1 \leq 2$, and $l^* \in \Pi_e^\perp$, but l^* need not be positive. Note also that some scalar multiple of each function in $\Pi_e^\perp$ is obtainable in this way, however, its norm might be arbitrarily small.*

OBSERVATION 5.5. *The following stronger version of 5.2 is false. "For all functions f computed by an $AC^0[d]$ circuit of size $s(n)$, there is an e as in 5.2 such that for every function $l^* \in \Pi_e^\perp$ with $||l^*||_1 \leq 2$, $|\sum_x l^*(x)f(x)| \leq 1/n$."*

PROOF. Let f be a Boolean function. If for every function $l^* \in \Pi_e^\perp$ with $||l||_1 \leq 1$, $|\sum_x l^*(x)f(x)| \leq 1/2n$, if and only if by 3.1, there is a function $g \in \Pi_e$ such that $||f-g||_\infty \leq 1/2n$. But since e is at most *polylog*(n), Fact 5.3 provides the required contradiction.

This suggests that in order to settle the [47] conjecture, one should take advantage of the special properties of the *positive* bounded functions $l \in \Pi_e^\perp \cup$ *One*. In other words, we would like to prove a stronger version of 5.1, with a weaker approximability hypothesis, but with the same consequence. ◯

The above discussion motivates our next general result on one-sided approximation. This result is also a modification of the duality principle. See [18].

THEOREM 5.6.

$$\sup_{\substack{l \in \Pi_e^\perp \cup One \\ \|l\|_1 = 1, l+ve}} |\sum_x l(x)f(x) - 1/2^n \sum_y f(y)| \qquad (Q1)$$

$$= \inf_{g \in \Pi_e \setminus One} |\sup_x (f(x) - 1/2^n \sum_y f(y) - g(x))| \qquad (Q2)$$

$$= \inf_{\substack{g \in \Pi_e \setminus One \\ g \geq f - \hat{f}(0^n)}} |1/2^n \sum_x g(x))|. \qquad (Q3)$$

REMARK 5.7. *The positive distribution l above, which can be viewed as a positive linear functional that annihilates all but the constant function in Π_e. Finding such functionals is called the "moment problem," and is studied extensively in [36], yielding methods for efficiently generating distributions l satisfying the above conditions, which is useful in obtaining pseudorandom generators for randomized computations based on f. The paper [18] takes the dual approach and gives methods based on quadrature formulas for finding the one-sided approximation g.*

PROOF. The first equivalence follows directly from the same argument as in the proof of the duality result 3.1, and does not use the orthogonality of $\Pi_e^\perp$ and *One*. We will show that $Q3 \geq Q1$ and $Q2 \geq Q3$.

To show the former, denote by f_1 the function $f - 1/2^n \sum_y f(y)$, and notice that for all $g_1 \geq f_1$, if l is positive, then

$$\sum_x l(x)g_1(x) \geq \sum_x l(x)f_1(x).$$

Furthermore, if $||l||_1 = 1$, l positive, $l \in \Pi_e^{\perp} \cup One$, and $g_1 \in \Pi_e$, then it is not hard to see

$$\sum_x |l(x)g_1(x)| == |1/2^n \sum_x One(x)g_1(x))| = |1/2^n \sum_x g_1(x))|.$$

It therefore follows that for all $g_1 \in \Pi_e$ with $g_1 \geq f_1$, and all $l \in \Pi_e^{\perp} \cup One$ with $||l||_1 = 1$, l positive,

$$\sum_x l(x)f_1(x) \leq |1/2^n \sum_x g_1(x))|$$

thus showing that $Q3 \geq Q1$.

To show that the $Q2 \geq Q3$, again denote $f - \hat{f}(0^n)$ by f_1 and assume to the contrary that there is a $g^* \in \Pi_e \setminus One$ such that for all $g_1 \in \Pi_e$ with $g_1 \geq f_1$,

$$|1/2^n \sum_x g_1(x)| > |\sup_x(f_1(x) - g^*(x))|.$$

We derive a contradiction to this assumption as follows: the function $g_2 \in \Pi_e$ defined as

$$g_2 := g^* + \sup_x(f_1(x) - g^*(x))$$

satisfies $g_2 \geq f_1$ and

$$1/2^n \sum_x g_2(x) = 1/2^n \sum_x g^*(x) + \sup_x(f_1(x) - g^*(x));$$

however, $1/2^n \sum_x g^*(x) = 1/2^n \sum_x One(x)g^*(x)$ equals 0 since g^* is orthogonal to One, thereby resulting in a contradiction. □

◯ It follows from 5.6 that the conjecture in [47] is equivalent to the following statement: for every function f computed by $AC^0[d]$ circuits of a fixed polynomial size, and for e being chosen as an appropriate polylogarithm, either there is a function g of degree at most e, with $g \geq f - \hat{f}(0^n)$ and $|\hat{g}(0^n)| \leq 1/n$, or that there is a function g of degree at most e with $\hat{g}(0^n) = 0$, and $|\sup_x(f(x) - \hat{f}(0^n) - g(x))| \leq 1/n$.

On the other hand, to show that the conjecture in [47] is false, it is sufficient to extend 5.5 in the following manner. For the case of the function $RO[2]$ in 5.3, Fact 3.1 establishes the existence of a function l, with $||l||_1 \leq 1$ and $l \in \Pi^{\perp}_{n^{1/4}}$ such that $|\sum_x l(x)r(x)| > 1/2n$. The only hitch is that this proof is not constructive. If l can be constructed, then to disprove the [47] conjecture, it is sufficient to find a positive function in $\Pi^{\perp}_{n^{1/4}} \cup One$ with a small 1-norm, which approximates l well. In particular, such functions can easily be constructed if l satisfies the condition that $|2^n \inf_x l(x)|$ is bounded above by $o(n)$: simply take l^* to be $l + \inf_x l(x)$ and normalize to get the required positive function $l^*/||l^*||_1$ in $\Pi^{\perp}_{n^{1/4}} \cup One$, for which $|\sum_x l^*(x)f(x)/||l^*||_1 - 1/2^n \sum f(x)|$ is larger than $1/n$. ◯

6. Uniform approximation

This section contains general results, methods and complexity-theoretic applications that involve the approximability of Boolean functions in the $||.||_\infty$ norm from various spaces. We also consider approximability in the $||.||_2$ norm in those cases where it is useful in studying $||.||_\infty$ norm approximability. We repeat that the treatment is restricted to inherently multivariate methods rather than univariate approximation theoretic techniques such as those employed in [51] [52], and to some extent, [3], and [47], some of which were discussed in Subsection 3.3.

This section is divided into 4 subsections. The first shows that non approximability can be reduced to approximability using duality relationships; the second discusses methods for establishing approximability; the third deals with the transitive nature of approximability relationships, which can be used to establish new (non)approximability results from already established ones; the fourth discusses how showing non approximability in the ∞-norm naturally lends itself to decomposition and divide-and-conquer paradigms which allows non approximability problems to be reduced, using analytic techiques, into their combinatrial constituents; and finally, the fifth shows that any of the non approximability methods in the previous four subsections can be used to establish small correlation between functions, which, in turn, is an essential ingredient for showing certain non approximability results.

Each subsection is, similar to the earlier sections, organized as follows. Following each theorem containing a result of a general nature, complexity-theoretic applications of the result are given as short remarks enclosed by ◯'s. These include earlier results from the literature that were proved using different

methods, as well as new observations. Finally, at the end of the discussion in each subsection, applications requiring longer exposition are presented. These employ the general results discussed so far, and include new results, as well as a few alternate proofs of old results.

6.1. Nonapproximability to approximability. The first theorem enumerates a number of systematic and general results showing that a Boolean function f is not approximable in the $||.||_\infty$ norm from any space spanned by a small number of Boolean functions from a class B, when it is known that f differs from all the functions in B with respect to some natural characteristic. In addition, some of the results assume various other realistic properties of f and the functions in B. The theorem applies duality relationships to convert a statement of non approximability of a Boolean function in the $||.||_\infty$ from the span of a small number of basis functions from a given class into a statement of approximability. The latter two statements below treat non approximability by functions with additional properties.

NOTE. As noted in Section 2, all approximating functions g from the span of a set $\{g_i\}$ are assumed to be of the form $g = \sum_i a_i g_i$, where the $a_i \in \mathbb{Q}$ and $\sum_i |a_i| \leq 1$.

THEOREM 6.1. *Let f be a Boolean function and B a set of Boolean functions, and let $M \subseteq B$ consist of independent Boolean functions.*

(1) The following are equivalent.

- *There does not exist an approximation $g \in span(M)$ with $sign(f) = sign(g)$.*
- *There exists an approximation $l \in span(M)^\perp$ with $||l||_1 > 0$ and $sign(f(x)) = sign(l(x))$, whenever $l(x) \neq 0$.*
- *There exists a distribution $\mathcal{R}$ such that $\sum_x f(x)h(x)R(x) = 0$ for all $h \in span(M)$.*

(2) The following are equivalent.

- *There does not exist an approximation $g \in span(M)$ with $sign(f) = sign(g)$, and $\epsilon \leq |g(x)| \leq 1$ everywhere. This happens when the coefficients in the approximation satisfy not only $\sum_i |a_i| \leq 1$, but also the a_i are rationals with denominators bounded by $1/\epsilon$; therefore the statement implies that f is realizable as an unweighted threshold of functions in M, if ϵ is chosen to be, say, $1/|M|$. We will call such a function g as a* **close** *approximation to f, or an approximation within $1 - \epsilon$ to f.*
- *There exists an approximation l close to $span(M)^\perp$ with $||l||_1 > 0$ and*

$||f-l||_\infty \leq 1$. *By "close to* $span(M)^\perp$*" we mean that* $|\sum_x l(x)h(x)|/||l||_1 \leq \epsilon$, *for all* $h \in M$.

- *There exists a distribution* $\mathcal{R}$ *such that* $| < h, f > |_\mathcal{R} \leq \epsilon$ *for all* $h \in M$.

(3) *There does not exist an approximation* $g \in span(M)$ *with* $|\sum_x f(x)g(x)| \geq \epsilon$ *(if, in addition,* $sign(f) = sign(g)$, *then* $||g||_2 \geq \epsilon$ *would be sufficient, since then* $|\sum_x f(x)g(x)| \geq ||g||_1 \geq ||g||_2$; *we will call such a function* g *as a* **high energy** *approximation to* f; *notice that a close approximation is a high energy approximation)*
if
for every subset $S \subseteq \{-1, 1\}^n$ *with* $|S| \geq 2^n\epsilon$, *there exists a distribution* $\mathcal{R}$ *over* S *such that* $| < f, h > |_\mathcal{R} \leq \epsilon$ *for all* $h \in M$,
or if
for all $h \in M$, $| < f, h > | \leq \epsilon$.

PROOF. The result (1) follows directly from 3.3. The equivalence of the latter two statements is simply due to the Booleanness of f.

For (2), the equivalence of the latter 2 statements is straightforward. For the first of the two statements, The '$\Leftarrow$' direction is shown as follows: if g is a close approximation to f, then for all distributions $\mathcal{R}$, it holds that $|\sum_x \mathcal{R}(x)g(x)f(x)| \geq \epsilon$. Since the coefficients of the linear combination $\sum_{h\in M} a_h h = g$ satisfy $\sum_x |a_h| \leq 1$, there must exist a Boolean function $h \in M$ such that $| < f, h > |_\mathcal{R} \geq \epsilon$. The '$\Rightarrow$' direction is shown as follows: the hypothesis is equivalent to the non-existence of a function g in $span(M)$ with $||f - g||_\infty \leq 1 - \epsilon$. This is equivalent, by 3.1, to the existence of $l_1 \in span(M)^\perp$ such that $||l_1||_1 \leq 1$ and $|\sum_x l_1(x)f(x)| > 1 - \epsilon$. Let S be the subdomain where $sign(l_1) \neq sign(f)$, and $l_1 \neq 0$. Since f is Boolean, $||l_1||_{1,S} < \epsilon/2$. Construct the function l by setting l_1 to 0 on S, and normalizing to get $||l||_1 = 1$. Now, l is no longer in $span(M)^\perp$, i.e, $\sum_x h(x)l(x)$ is no longer 0 for all $h \in M$, but since the functions $h \in M$ are Boolean, clearly, $|\sum_x h(x)l(x)| \leq \epsilon/(2(1-\epsilon/2)) \leq \epsilon$, since $\epsilon \leq 1$, and since $|\sum_x h(x)l_1(x)| = 0$.

For (3), the last of the 2 hypotheses clearly implies the non-existence of a high-energy approximation, since the linear combination $\sum_{h\in M} a_h h = g$ satisfies $\sum_{h\in M} |a_h| = 1$, and if g is a high-energy approximation, then $|\sum_x g(x)f(x)| \geq \epsilon$. The first of the 2 hypotheses implies non approximability because every high-energy approximation g to f $|\sum_x g(x)f(x)| \geq \epsilon$ is a close approximation to f

over at least an ϵ fraction of the domain, hence (2) can be applied over this domain.

REMARK 6.2. *Notice that (1) does not use the Boolean-ness of the basis set M, although (2) and (3) do. In fact, (1) depends only on the space $span(M)$, whereas the concepts of close and high energy approximation are specific to a particular basis M. This generates an open question as to whether a stronger result can be proven instead of (1), using the Boolean-ness of the basis set M.*

The statement (1) is equivalent to a notion of orthogonality with respect to the $||.||_\infty$ norm, according to which a function f is "orthogonal" to a space $span(M)$, if for every function $g \in span(M)$, $||f+g||_\infty < ||f||_\infty$. This reduces to the usual orthogonality in the case of the $||.||_2$ norm (see [62] for an succinct treatment of $||.||_\infty$ norm, or uniform approximation). In addition, we will see in Section 7 that the $|M|+1$ points forming the support of the function h in $span(M)^\perp$, together with the sign of f at these points form an "extremal signature" that characterizes f, i.e, they constitute points where $|(f-g)(x)|$ attains its maximum (i.e, $||f-g||_\infty$), when g is a *best* approximation to f.

◯ Almost all the threshold complexity lower bounds, for example, [33], [41], [44], [26], known so far concerning non approximability of a function f by from the span of small number of LT_1 or other functions involve a *restriction* on the approximation: the the linear combinations that form the approximant have polynomially bounded coefficients. In other words, these lower bounds apply only to circuits with an unweighted threshold gate at the top. These lower bounds *all* use the 'correlation/discriminator lemma,' proved in [33] and [26], which is nothing but 6.1(2). To complete such a lower bound proof, one then needs to show that the scalar product $| < f, h > |$ is small for each $h \in M$, over an appropriate distribution $\mathcal{R}$. The methods for bounding this scalar product have been phrased in terms of communication complexity (for example [26]) and "variation rank" (for example [44]), but most of these also reduce to arguments based on duality and simple norm relationships for Boolean functions as we shall see in Theorems 6.12 and 6.22. ◯

◯ The only two non approximability results *without* the above restriction are the following: the result of [23] on the non approximability of *Parity* by few functions computable by $\{\wedge, \vee, \neg\}$- circuits of a fixed polynomial size and constant depth; and related results of [43], for example, on the non approximability of an $AC^0[3]$ function by few *And* functions and the result of [42] on the non

approximability of an $AC^0[3]$ function by few *mod r* functions (which can be viewed as monomials over the reals, or $\mathbb{Z}_2^n$ characters; in fact, the result applies to $\mathbb{Z}_r^n$ characters, for any r). Both papers use straightforward duality from 6.1 (1) as the main technique - although it is not stated as such. The former paper uses 6.1 (1) in conjunction with the switching lemma of [30]. The later paper "divides" the problem into "pieces", as will be discussed in Observation 6.14, and uses 6.1 (1) to "conquer" the pieces.

A quote from the latter paper [42] states: "Previous lower bound results on threshold representations are based on the discriminator method, a geometric method, a method based on probabilistic communication complexity, and a spectral theoretic method for orthogonal bases."

In fact, Theorem 6.1, Theorem 6.12, Observation 6.14, and Theorem 6.22 show that the main analytic content of *all three methods* is straightforward duality, and simple norm relationships for Boolean functions. ◯

Theorem 6.1 (3) seems, on the surface, to be cumbersome to use for proving non approximability results, since it would involve proving a sweeping universal statement, and furthermore the applicability seems questionable since the conditions on the approximation seem too strong. However, the following theorem motivates by giving a natural situation under which the statement of Theorem 6.1 (3) is useful.

THEOREM 6.3. *Let f be a Boolean function and B a set of Boolean functions, and let $M \subseteq B$ consist of orthonormal Boolean functions. Any approximation $g \in span(M)$ with $sign(f) = sign(g)$, $g = \sum_{h \in M} a_h h$, and $\sum_{h \in M} |a_h| \leq 1$, is in fact a high energy approximation satisfying $||g||_1 = \sum f(x)g(x) \geq 1/|M|$.*

PROOF. We simply show that any function g, which is a linear combination as in the theorem, satisfies $||g||_1 \geq 1/|M|$, using the fact that the functions in M are orthonormal. If, in addition, $sign(f) = sign(g)$, then it follows that $|\sum_x f(x)g(x)| \geq 1/|M|$, since f is Boolean.

Since M forms an orthonormal set,

$$\sum_{h \in M} a_h^2 = \sum_x g^2(x) = 2^n ||g||_2^2.$$

Assuming without loss that $\sum_{h \in M} |a_h| = 1$, it follows that

$$\sum_{h \in M} a_h^2 \geq 1/|M|,$$

and hence $2^n||g||_2^2 \geq 1/|M|$. Moreover, the functions h are Boolean, and therefore $||g||_\infty \leq 1$, since $\sum_{h \in M} |a_h| = 1$; thus

$$||g||_1 \geq 2^n||g||_2^2 \geq 1/|M|.$$

◯ Theorem 6.3, together with Theorem 6.1 (3) directly implies the "spectral method" of [15] that the number of *Parities* or monomials, or any other orthonormal set needed to approximate a function exceeds the inverse of its maximum Fourier coefficient, or respectively, the maximum scalar product of the function with an element of the orthonormal set. I.e, $PT_1 \subseteq PL_\infty^{-1}$. ◯

The next theorem provides natural conditions which imply non approximability through Theorem 6.1 (1). Intuitively the theorem states that if $f|_{X^\perp}$ behaves like f, then then f is not approximable by the span of functions in M.

THEOREM 6.4. *Let f be Boolean and M be a set of Boolean functions. If $f \notin X$ and $||f|_X||_\infty \leq 1$ then f is not approximable in the ∞ norm from X, i.e, there is no $g \in X$, with the same sign as f.*

PROOF. First split f into two parts as $f = f|_X + f|_{X^\perp}$. Since $f \notin X$, it follows that $f|_{X^\perp}$ is not identically 0. Furthermore, since f is Boolean, $||f|_X||_\infty \leq 1$ if and only if $f|_{X^\perp}$ has the same sign as f where ever non-zero. The theorem then follows from 6.1(1). □

Is the converse of Theorem 6.4 true? The answer is no. In other words, it could be that f is not approximable from X and yet $f|_{X^\perp}$ does not behave like f. However, a version of the converse does hold, thereby giving another equivalent condition to non approximability, via Theorem 6.1 (1).

FACT 6.5. *We use $f|_{S,X}$ to denote the projection of f_S on X_S. Notice that this is different from taking the projection $f|_X$ and then restricting it to S. In other words, as mentioned in the background section, the space $X_S^\perp$ is not the same as taking $X^\perp$ and restricting to S. The following are equivalent.*

- *There is a nonempty subdomain S with $||f|_{S,X}||_{\infty,S} \leq 1$*
- *There is no $g \in X$, with the same sign as f.*

PROOF. For the forward direction, define $l \in X^\perp$ to be 0 outside S, and $f|_{X_S^\perp} = f_S - f|_{S,X}$ on S. Since $f|_{S,X}(x) \leq 1$ for all $x \in S$, it follows that on its support, l has the same sign as f. Now apply 6.1(1).

For the reverse direction, we use a geometric argument to find the set S^* such that $f|_{X|^{\perp}_{S^*}} = f_S - f|_{S^*,X}$ has the same sign as f where ever non-zero.

Let C_f be a cone or orthant in $\mathcal{F}_{2^n}$ - viewed as $\mathbb{R}^{2^n}$ - and given by $\{g : sign(g) = sign(f)\}$. Notice that each facet of this cone is also a cone that contains exactly those functions that are 0 outside some subdomain S, and are either 0 or have the same sign as f on S. The entire cone C_f corresponds to $\bar{S}$ being empty. So we will denote the facet corresponding to subdomain S as $C_{f,S}$. Now $X^{\perp}$ is a subspace that satisfies at least one of the following properties.
(a) it completely contains some proper facet (of at least one lower dimension) $C_{f,S}$ of C_f,
(b) it cuts through a proper facet $C_{f,S}$ of C_f, or
(c) it is completely contained in the subspace formed by extending some proper facet $C_{f,S}$ of C_f to all orthants, i.e, the subspace containing exactly all functions that are 0 outside S.

In case (a), we simply choose $S^* = S$. Clearly, $f|_{S^*,X} = 0$ and we are done. In cases (b) and (c), we continue this process on a smaller cone $C_{f,S}$, starting with the function f_S instead of f, and the subspace $X^{\perp}_S$ instead of X. For the base case, when $|S| = 2$, in cases (b) and (c), it is easy to see that $f_S - f|_{S,X}$ does in fact have the same sign as f_S. □

One situation when 6.4 is applicable is if X has a Boolean orthonormal basis M and $| < f, g > | \leq 1/|M|$ for $g \in M$. Because, by orthonormality, $f|_X = \sum_{g \in M} < f, g > g$, and since the g are Boolean and the quantities $| < f, g > |$ are small we obtain that $||f|_X||_\infty \leq 1$. Unfortunately, under this situation, 6.4 is not particularly useful in the sense that with these strong conditions, the non approximability conclusion of the theorem can be directly obtained using Theorem 6.3 and Theorem 6.1 (3). However, 6.4 is applicable in some situations where Theorem 6.3 and Theorem 6.1 (3) cannot be used. The next two theorems consider a natural situation where M does not consist of orthonormal functions, and yet 6.4 (6.5) can be applied to show non approximability. Here, all the functions in B, when viewed as vectors in $\mathcal{F}_{2^n}$, form vector bundles, such that all the vectors in any one bundle are close to each other (have large scalar product), but any two bundles are nearly orthogonal to each other.

THEOREM 6.6. *If all pairs of functions g_i, g_j in the class B of Boolean functions satisfy either $| < g_i, g_j > | \leq \delta$ or $| < g_i, g_j > | \geq 1 - \delta$, for δ being typically significantly less than $1/2$, and furthermore, for a given Boolean function f,*

$| < f, g_i > | \leq \epsilon$, *for all* $g_i \in B$, *then for* $M \subseteq B$ *with* $|M| \leq min\{1/\delta^{1/3}, 1/\epsilon^{1/3}\}$, *there is no* $g \in span(M)$ *with the same sign as* f.

PROOF. We first construct a subdomain S and show that $f|_{span(M)_S^{\frac{1}{2}}} = f_S - f|_{S,span(M)}$ has the same sign as f_S. Theorem 6.4 or 6.1(1) then completes the proof. The subdomain is constructed by first dividing M into bundles such that for pairs g_i, g_j in each bundle, $| < g_i, g_j > | \geq 1 - \delta$ and for g_i and g_j in different bundles, $| < g_i, g_j > | \leq \delta$. For each bundle M_k, we find a representative function $g_k \in M_k$ and remove from the S all points where $g_k \neq g_i$, for some $g_i \in M_k$. The subdomain S thus constructed is no less than $1 - |M|\delta$ of the entire domain; therefore, the values $| < f, g_i >_S |$ are still no larger than $(\epsilon + |M|\delta)/(1 - |M|\delta)$, and the values $| < g_i, g_j >_S |$ are either 1, i.e, $g_{i,S} = g_{j,S}$, or is at most $(\delta + |M|\delta)/(1 - |M|\delta)$, for $g_i, g_j \in M$. I.e, $g_{i,S}$ and $g_{j,S}$ are either identical or almost orthogonal. Furthermore, $||g_i||_{2,S}$ is still 1, since the g_i are Boolean. Intuitively, the function $\sum_i < f, g_i >_S g_{i,S}$ is a reasonable approximation to the true projection $f|_{S,span(M)}$, since the g_i are almost orthonormal over S. Furthermore,

$$||\sum_i < f, g_i >_S g_{i,S}||_{\infty,S} \leq \sum_i | < f, g_i >_S | \leq \frac{(\epsilon + |M|\delta)}{(1 - |M|\delta)}|M|,$$

which is at most 1 provided $|M|$ is sufficiently small as in the statement of the theorem.

To find the true projection, we find orthonormal basis functions g_i^* for $span(M)|_S$, from the functions $g_i|_S$, using, for example, Gram-Schmidt orthonormalization. We omit the exact calculations. Basically, since $g_{i,S}$ already forms a close-to-orthonormal basis, the orthonormalization does not blow-up either the ∞-norm of the functions g_i^*, or the values $| < f, g_i^* >_S |$, and thus the projection $f|_{S,span(M)} = \sum_i < f, g_i^* >_S g_i^*$ still continues to have a small ∞-norm provided $|M|$ is at most the bound given in the theorem. □

REMARK 6.7. *In general, by 6.1(3), and 6.4, it follows that if a distribution* $\mathcal{D}$ *can be found such that every pair of distinct Boolean functions* g_i *and* g_j *in a set* M *are almost orthonormal with respect to* $<, >_{\mathcal{D}}$, *and if* $< f, g_i >_{\mathcal{D}}$ *is small for every* $g_i \in M$, *then there is no approximation from* $span(M)$ *to* f *with the same sign.*

◯ We now turn to a specific application. In the next theorem, we use 6.4 to provide an alternative proof of the non approximability result that $AC^0[3] \not\subseteq$

PT_1, [42]. In this case, while M does consist of orthonormal functions, the scalar product of f with some functions in M is fairly large, and hence Theorem 6.1 (3) does not apply. The proof of Theorem 6.8 also uses spectral methods for orthogonal bases which were eschewed in [42] as not being applicable for this purpose. Furthermore, duality is the primary ingredient, since Theorem 6.4 uses only Theorem 6.1 (1).

THEOREM 6.8. *If the $AC^0[3]$ function $f(x) := \bigvee_{i=1}^{l_1} \bigwedge_{j=1}^{l_2} \bigvee_{k=1}^{l_3} x_{ijk}$, where $n = l_1 l_2 l_3$ and $l_1 := l_2 := l_3 :=$ has an approximant $g \in span(M)$ where M consists of parity functions χ_s, with $||f - g||_\infty < 1$, then $|M| \geq \Omega(2^{n^c})$, for any constant $c < 1$.*

PROOF. By 6.4, it is sufficient to show that for all sets M consisting of parity functions χ_s, $||f|_{span(M)}||_\infty \leq 1$, unless $|M| \geq \Omega(2^{n^c})$, for constant $c < 1$. Since

$$||f|_{span(M)}||\infty = || \sum_{\chi_s \in M} \hat{f}(s)\chi_s||_\infty \leq \sum_{\chi_s \in M} |\hat{f}(s)|,$$

we will simply bound this last quantity.

Set the quantities $q_1 := 1/2^{l_1}$, $q_2 := (1 - q_1)^{l_2}$ and $q_3 := (1 - q_2)^{l_3}$. It is not hard to see (see [65]) that $|\hat{f}(x)|$ is largest for $|\hat{f}(1^n)| = |1 - 2q_3|$ and for x such that $x_{ijk} = 1$ except for a single value of i and j. In the latter case, $|\hat{f}(x)| \leq 2q_1q_2q_3/((1 - q_1)(1 - q_2))$. Since

$$\sum_{\chi_s \in M} |\hat{f}(s)| \leq 2q_3 - 1 + |M| * 2q_1q_2/(1 - q_1)q_3/(1 - q_2). \qquad I$$

Now the conditions on $|M|$ corresponding to $\sum_{\chi_s \in M} |\hat{f}(s)| \leq 1$ depend on whether

(i) $q_3 \geq 1/2$ or

(ii) $q_3 \leq 1/2$.

Now, in case (i), $\sum_{\chi_s \in M} |\hat{f}(s)|$ is bounded by 1 as long as

$$|M| \leq (1/q_1 - 1)(1/q_2 - 1)(1/q_3 - 1);$$

and in case (ii), $\sum_{\chi_s \in M} |\hat{f}(s)|$ is bounded by 1 as long as

$$|M| \leq (1/q_1 - 1)(1/q_2 - 1);$$

We choose case (i), and ignore the latter. Now for any $c < 1$, l_1, l_2, l_3 can be chosen such that both $q_3 \geq 1/2$, and $(1/q_1 - 1)(1/q_2 - 1)(1/q_3 - 1) \geq \Omega(2^{n^c})$. which completes the proof. □

○

6.2. Approximability. Approximability results are useful for showing non approximability, via 6.1. They also permit us to build on previous non approximability results by using the transitive nature of approximation, as will be seen in the next section. They are needed to show complexity lower bounds using the framework **A**: in order to conclude from the non approximability of f from a space of functions, that a function f is not in a class C, one needs to show that every function in the class C is approximable from the space. They are of independent interest in proving complexity upper bounds and are equivalent to upper bounds for threshold circuit complexity. In fact, as will be seen in the next subsection, approximability results also result in communication complexity upper bounds. (However, as mentioned in Section 1, we do not concentrate on obtaining threshold and communication complexity upper bounds as approximability results.) Finally, constructive approximation results provide learning algorithms, as will be seen in Section 7

In this section, we discuss the two main techniques that have been used so far for showing uniform approximability ($||.||_2$-norm approximability from a space of functions is straightforward, since the best approximation is simply the projection of the given function onto the space: this is easy to deal with, if the space has a nice orthonormal basis, and needs adhoc methods otherwise; furthermore, a close or high energy uniform approximation is automatically a good 2-norm approximation).

The first technique follows from an investigation of the conditions under which a converse version of Theorem 6.3 holds, i.e, where the existence of an approximation with large norm (from an orthonormal basis) implies the existence of an approximation from a space spanned by only a few of the basis functions.

THEOREM 6.9. *If f is closely approximable by $g \in span(B)$, with $|g(x)| \geq 1/m$ for all x, where B is an orthonormal basis, then f is closely approximable by $g \in span(M)$ with $|g(x)| \geq 1/m^2$ everywhere, where $M \subseteq B$, and $|M| \leq m^2$.*

◯ The proof of the theorem above is along the same lines of the proof of [15]. They showed a special case of the above result, namely that $PL_1 \subseteq PT_1$. ◯

The second method uses the fact that upper bounds on communication complexity provide a method of showing approximability from a specific set of basis functions, namely the functions over $\{-1,1\}^n \times \{-1,1\}^n$ that are characteristic functions of cross-products $A \times B$, $A, B \subseteq \{-1,1\}^n$. We call these *cross-product* functions. The following facts follow directly from the definition of deterministic and probabilistic communication complexity.

FACT 6.10. *If the deterministic communication complexity of a Boolean function f is at most $\log m$, then f is exactly interpolated by $\sum_{i \leq m} r_i + (m-1)$, where the r_i are cross-product functions not including the constant function.*

FACT 6.11. *If the $1 - \delta$-error probabilistic communication complexity of a Boolean function f is at most $\log m$, then there is a very close approximation g with the same sign as f, of the form $g = \sum_{i \leq m} a_i r_i$, where, as usual, $\sum_{i \leq m} |a_i| \leq 1$, the the r_i are cross-product functions, and $|g(x)| \geq \delta$ everywhere.*

◯ The above facts seem to indicate that non approximability results are stronger than lower bounds on communication complexity, and hence the lower bounds can be obtained from non approximability results, but not viceversa. However, in practice, several known communication complexity lower bounds such as [33], [26], [32], [38], [56], [25], [29], [4], [24], do yield the corresponding non approximability results that can be directly proven using the duality-based methods being discussed in this paper although, theoretically, it might have been easier to obtain such lower bounds by other means than prove non approximability.

Similarly, not only do upper bounds on communication complexity give approximability results, but usually, the converse also holds. See, for example, [26]. In the next subsection on transitive approximability, we discuss this relationship in more detail. ◯

6.3. Transitive approximability. Here, we give results of the following form: "f is approximable from the span of a set of m_1 functions g_i in some basis B, and g_i are all approximable from the span of a set of m_2 "simple" functions h, then f is approximable from the span of $m_1 m_2$ "simple" functions." *These results will follow directly from 6.1 and 6.3.*

Results of this nature are useful for building on previous non approximability results. For example, a result of the above form, together with the non approximability of f from the span of $m_1 m_2$ simple functions would imply that one of the approximability hypotheses is false.

Moreover, statements of this type remain meaningful when the word "not approximable" is replaced by "small scalar product." This will be used in Subsection 6.5.

THEOREM 6.12. *Let f be a Boolean function, B a set of Boolean functions, and Q a set of "simple" Boolean functions.*

(1) If

• *f has a high energy approximation g, with $|\sum_x f(x)g(x)| \geq \delta$, from the span of functions in B*
and
• *every function in B is the linear combination of Boolean functions in Q, with the sum of the absolute values of the coefficients bounded by m,*
then
$|< f, h >| \geq \delta/m$, for some function $h \in Q$.

(2) If

• *f has a high energy approximation g from the span of functions in B with $|\sum_x g(x)f(x)| \geq \epsilon$,*
and
• *for every $g \in B$ there is a close approximation h, with the same sign as g, and in the span of m Boolean functions in Q, satisfying $|h(x)| \geq \delta$ for all x*
then

(i) there is a set of m of functions $h^* \in Q$, and a subset S consisting of at least $(1+\epsilon)/2$ fraction of the domain, such that for every distribution $\mathcal{R}$ over S, $|< f, h^* >|_{\mathcal{R}} \geq \delta$, for at least one of the m functions h^*; and

(ii) $|< f, h^* >| \geq \epsilon + \delta - 1$, for some $h^* \in Q$.

(3) If

• *f has a close approximation g from the span of a set of m_1 functions g_i, (i.e, g has the same sign as f with $|g(x)| \geq \delta$ for all x,)*
and
• *each g_i has a close approximation h_i from the span of a set of m_2 functions h_{ij}, with $|h_i(x)| \geq \epsilon$ for all x, and $1 - \epsilon < \delta/m_1$,*
then
there is a close approximation h^* to f from the span of the $m_1 m_2$ resulting functions $h_{ij} \in Q$ with $|h^*(x)| \geq \epsilon\delta$ for all x.

(4) If

• *f has a close approximation g from the span of m_1 functions $g_i \in B$ with with $|g(x)| \geq \delta$ for all x,*
and
• *each g_i can be expressed as a linear combination of functions in a set Q with the sum of the absolute values of the coefficients bounded by m_2,*

then
f has a close approximation h^*, with $|h^*(x)| \geq \delta/m_2$ everywhere. It follows from 6.1(2) that for every distribution $\mathcal{R}$, there is a funciton $h \in Q$ such that $< f, h >_{\mathcal{R}} \geq \delta/m_2$. Moreover, if Q happens to consist of orthonormal functions, then, by 6.9, f has a close approximation h^* from the span of at most $4m_1^2m_2^2$ functions in Q.

(5) If
- *f has a high energy approximation g, with the same sign as f, from the span of m_1 functions in B with $|\sum_x g(x)f(x)| \geq \epsilon$,*

and
- *each function in B has an approximation with the same sign from the span of m_2 simple functions $h \in Q$,*

then
f is approximable on some subset S consisting of at least $(1+\epsilon)/2$ fraction of the domain from the span of m_1m_2 functions $h \in Q$.

PROOF. For (1) (2) and (5), since f has a high energy approximation $g \in span(B)$, by 6.1 (3), there is a function $g^* \in B$ such that $|\sum_x f(x)g^*(x)| \geq \delta$. (1) now follows immediately.

For (2) and (5), notice that since both f and g^* are Boolean, f and g^* must coincide in sign on at least a $(1+\epsilon)/2$ fraction of the domain. (5) now follows immediately.

For (2), since g^* has a close approximation h from the span of m functions $h^* \in Q$, (i) follows by 6.1(2). The consequence (ii) follows from the fact that h is a high energy approximation to g^* with $|\sum_x g^*(x)h(x)| \geq \delta$. Thus we obtain the existence of 3 Boolean functions f, g^* and h^* such that

$$< f, g^* > \geq \epsilon, \text{ and } < g^*, h^* > \geq \delta.$$

Therefore $< f, h^* > \geq \epsilon + \delta - 1$.

For (3), we use the fact that the functions g_i have a very close approximation h_i from the span of m_2 functions h_{ij}, since $1 - \epsilon \leq \delta/m_1$. Now, to form the required close approximation to f from the span of the m_1m_2 functions h_{ij} in Q, modify g as follows: simply replace each of the g_i's that form g, with the corresponding approximation h_i.

For (4), choose the close approximation h^* to f as g/m_2. □

◯ 6.12 (1) and (2) form the backbone of the lower bounds (non approximability results) of [33] and [26] that $\hat{LT}_3 \not\subseteq \hat{LT}_2$, $LT_1 \not\subseteq \hat{PT}_1$, and $PT_1 \not\subseteq \hat{LT}_2$. [26]

uses communication complexity upper bounds to prove approximability of the functions in the relevant class B by the span of a few cross-product functions (in Q), and then, in effect, uses of 6.1 (2) to show that f is not appropriately approximable from the span of few cross-product functions. The desired non approximability of f from the span of few functions in B then follows from 6.12. The papers, especially [26] employ the communication complexity paradigm throughout instead of treating the issue as approximability from cross-product functions.

It should be noted that while it is often easier to show that *specific* functions are appropriately approximable from the span of cross-product functions by giving a *direct* upper bound on the communication complexity, the word "communication complexity" can otherwise be removed from all (lower bound) proofs involving threshold functions, or linear (non) approximability results, without making the proofs any more difficult, or any less intuitive. In fact, translating "low communication complexity" as "appropriate approximability by few cross-product functions" allows one to take advantage of transitive approximability. This, in turn, allows a natural extension of approximability notions, such as 6.12 that are already being employed and and often makes the proofs shorter and more transparent. ◯

◯ An example application of 6.12(3) is the following result proved in [26] using communication complexity. This can be obtained directly from 6.12(3) using the fact that every LT_1 function has a very close approximation from the span of few $\hat{LT}_1$ functions.
"A circuit with an unweighted linear threshold gate on top, arbitrary linear threshold gates at the middle level, and gates from a class C in the lowest level can be simulated by a circuit with exactly the same gate on top, unweighted linear threshold gates in the middle level and exactly the same gates from C at the bottom."

Another example application of 6.12(3) is the following: LT_1 functions have a very close approximation from the span of few $\hat{LT}_1$ functions by a result of [26]. Moreover, $\hat{LT}_1$ functions have a very close approximation from the span of few cross-product functions by a simple probabilistic communication complexity upper bound, in fact they are even interpolable from the span of a few more cross-product functions by the straightforward deterministic communication complexity upper bound. Therefore, LT_1 functions have a very close approximation from the span of few cross-product functions, and this approximation does yield a probabilistic communication complexity upperbound for LT_1 functions, although this is not generally a consequence of 6.11 alone. It is

natural and promising to investigate if approximability results are, in general, a viable method for proving communication complexity upper bounds. ◯

◯ Non-approximability results from cross-product functions, on the other hand, are theoretically stronger than lower bounds on communication complexity, but nevertheless provide a viable method of proving such lower bounds, since in practice, several of the known lower bounds on communication complexity such as the results of [33], [26], [32], [38], [56], [25], [29], [4], [24] actually yield the stronger non approximability results from cross product functions, which can be obtained independently using the methods of this Section. For example, using 6.12(1) and (2), a lower bound of $\log m$ on the communication complexity of a function g can be obtained by showing g is not a linear combination -with small coefficients - of m cross-product functions or by showing that there is a function f such that $| < f, s > | \leq \epsilon$ for cross-product functions s, and yet $| < f, g > | \geq m\epsilon$. Similarly, a lower bound of $\log m$ can be obtained on the $(1-\delta)/2$-error probabilistic communication complexity of g as follows: show the negation of 6.12 (2) (i) that for each set of m cross-product functions, for some ϵ and each subset S with $|S| \geq (1+\epsilon)/2$ fraction of the domain, there is a distribution $\mathcal{R}$ over S with $< f, s >_{\mathcal{R}} \leq \delta$, but $< f, g > \geq \epsilon$. ◯

◯ The hypothesis of 6.12 (2) also holds when the functions g are functions in $\hat{LT}_1$, with $h = \sum_i a_i x_i + a_0$ and, as usual, $\sum_i |a_i| \leq 1$ and the a_i are rationals with denominators bounded above by $1/\delta$. Here the simple functions in Q are the linear monomials and the constant function. In this case, showing the the negation of 6.12 (2) (ii) for $1/\delta$ being polynomially bounded would simply mean that $g \notin \hat{LT}_1$. ◯

◯ A special version of 6.12(4) is used in the proofs of [42] and [34] i.e, the *approximability* result that $AC^0[2] \subseteq \hat{PT}_1$. [42] gives a separate probabilistic argument. [34] gives an algorithm to find the approximating function which we will discuss in Section 7. Our proof of 6.12(4) above is straightforward from 6.1(2), and 6.9, and clarifies exactly where the orthonormality of the functions in Q is needed, and moreover gives the general result below as a corollary. ◯

COROLLARY 6.13. *$\hat{LT}_1$ of $PL_1 \subseteq \hat{PT}_1$. I.e, an unweighted threshold of polynomially many functions in PL_1 can be simulated by an unweighted threshold of polynomially many Parity functions.*

6.4. Non-approximability via divide and conquer. Next we consider non approximability results for functions f that involve decomposing the domain. These results are useful in constructing hard f based on previously

proven or easier non approximability results, and, in addition, give methods for reducing a non approximability question to its combinatorial core.

Before giving a broad description of the general method, we reiterate that non approximability of f over the whole domain is equivalent to saying that from each candidate approximating space there is no approximation of f over *some* distribution or subdomain (even though no restriction of f may be hard to approximate over all). This is already implicit in both Theorem 6.1 (1) and (2) and is false for the case of high energy approximations, which is why 6.1 (3) has no natural converse, but even so, Theorem 6.1 (3) can and will be used in what follows.

The results below have the following form. The domain of f is decomposed into the (not neccessarily disjoint) union $\bigcup_i P_i$ of subdomains P_i that typically look identical; for example, their characteristic functions could be shifts of the same function, as in the case of the subdomains covered by any one row or one column of a communication matrix. The P_i are so chosen so that some structure is visible concerning the behavior of f as well as the behavior of the approximating space $span(M)$, over the pieces. Now, to show that f is not approximable from $span(M)$, one uses the above structure to show the *existence of at least one piece* P_i over which f is not approximable from $span(M)$. Not surprisingly, this allows more freedom in the proof process, although the argument used over the individual pieces is still based on 6.1 and the other techniques discussed in the earlier theorems of this section. This general method is useful when it is known that f and the functions in the approximating space $span(M)$ differ in some local characteristic, so that global measures such as the scalar product do not capture the difference, but on the other hand, the exact subdomain where they differ cannot be determined using any local characteristic of f. This seems to be the situation while trying to find a suitable function that is not in LT_2. We explain some examples of decomposition in more detail in what follows.

OBSERVATION 6.14. *Let $\bigcup_i P_i \subseteq \{-1,1\}^n$, and for all i, let $P_i = P_0 \oplus s_i$, where '$\oplus$' stands for addition when $\{-1,1\}^n$ is viewed as $\mathbb{F}_2^n$, and s_i is the shift vector for P_i. In other words, the P_i's are shifts of each other. Given a function f on $\{-1,1\}^n$, we denote by f_{P_i} its restriction to P_i; furthermore, we shall view all of the functions f_{P_i} as being over P_0, by defining $f_{P_i}(x) := f(x \oplus s_i)$. Let f be Boolean, B be a set of Boolean functions, and M be a (typically small) subset of B.*

(i) If the functions f_{P_i} form an orthonormal basis for the space of functions from P_0 to the reals, and the functions $\{g_{P_i} : g \in M, i \in \mathbb{N}\}$ span a

subspace X_M of dimension m, then there is no close approximation h to f from span(M), with $|h(x)| \geq \sqrt{m/|P_0|}$ for all $x \in P_0$.

(ii) If for some i, the set $\{g_{P_i} : g \in M\}$ forms an orthonormal set and $< f, g >_{P_i} < 1/|M|$ for all $g \in M$ then there is no approximation to f from span(M). This can be extended to the case where the set $\{g_{P_i} : g \in M\}$ forms orthonormal bundles as in Theorem 6.6.

(iii) If for every P_i, $f_{P_i} = f_{P_0}(\pi_i)$, with the map π_i being, for example, a permutation of the variables, and f_{P_0} is not approximable from a space X that is typically closed under the maps π_i. Note that P_i is simply a shift of P_0, and is not $P_0(\pi_i)$. Furthermore, for some i, $M_i = \{g_{P_i} : g \in M\}$ is a subset of X. Then f is not approximable from the span of functions in M.

For the following, assume that the pieces P_i of the domain are not shifts of each other, but rather, $P_i = \pi_i(P_0)$, for some uniform set of maps π_i, for example certain permutations of the variables. In this case, for any function, we define f_{P_i} over the domain P_0 to be $f_{P_0}(\pi_i)$, i.e, $f_{P_i}(x) = f(\pi_i(x))$.

(iv) Let the set of functions M be such that if $g \in M$, then either $g(\nu_i) \in M$ for all the maps ν_i, or the expectation $\mathbf{E}_i g_{P_i}(x) \leq \delta/m$ for all x. Furthermore, let f be invariant under the maps ν_i, so the f_{P_i}'s are all identical to f_{P_0}. Finally, assume that f_{P_0} is not approximable over P_0 from the span of any set M of at most m functions from B, which is closed under the maps ν_i. Then f is not closely approximable by $g \in$ span(M) with $|g(x)| \geq \delta$ for all x.

REMARK 6.15. *All of these results can be seen to employ just 6.1 together with the earlier theorems of this section on the pieces. In addition, analogous results can be shown when the subdomains P_i are replaced by arbitrary distributions, much in the same way that remark 6.7 is analogous to 6.6.*

PROOF. For (i), we show that there is a P_i such that for the distribution $\mathcal{R}$ that is 1 on P_i and 0 elsewhere, $< f, g >_R < \sqrt{m/|P_0|}$, for all $g \in M$. Then the proof follows by 6.1(2). We in fact show something stronger. Since the functions $f_{P_i} : i \in \mathbb{N}$ form an orthonormal basis for a space of dimension $|P_0|$, and the set $\{g_{P_i} : g \in M, i \in \mathbb{N}\}$ only spans a subspace X_M of dimension m, it can be shown using simple linear algebra and geometry, that there must be at least one P_i such that the projection $f_{P_i}|_{X_M}$ has a 2-norm at most $\sqrt{m/|P_0|}$. In

other words, it is easy to show that if for all the P_i the projections $f_{P_i}|_{X_M}$ had 2-norms exceeding δ, then the space spanned by the projections $f|_{P_i}|_{X_M}$, and therefore the space X_M would have dimension at least $|P_0|\delta^2$. Thus for some P_i, $< f_{P_i}, h^* >_{P_i} < \sqrt{m/|P_0|}$, for all $h^* \in X_M$, with 2-norm bounded by 1; and taking h^* to be any of the functions in $\{g_{P_i} : g \in M, i \in \mathbb{N}\}$, we have what we require. Notice that the above proof depends only on the dimension of the space X_M and goes through independent of the exact basis M.

For (ii), since the g_{P_i} form an orthonormal set for some P_i, by 6.3, any approximation g from their span to f is a high energy approximation with $< f, g >_{P_i} \geq 1/|M|$. The result follows.

The proof of (iii) is straightforward.

For (iv), notice that f is closely approximable by $g^* = \sum_{g \in M} a_g g$, with $|g^*(x)| \geq \delta$ for all x, if and only if f is similarly closely approximable by

$$\mathbf{E}_i g^*(\nu_i) = \sum_{g \in M} a_g \mathbf{E}_i g(\nu_i).$$

Now construct a new set M_1 of only those functions g in M for which $g(\nu_i)$ is also in M. Since f is invariant under the maps ν_i, if f is closely approximable from the span of functions in M, then it is closely approximable from $M(\nu_i)$ for *all* i. In particular, it is closely approximable by $\sum_{g \in M} a_g \mathbf{E}_i g(\nu_i)$ as well. Furthermore, since for every function $g \in M$ that is not in M_1, $\mathbf{E}_i g_{P_i}(x) \leq \delta/m$ for all x, it follows that f is also approximable by $\sum_{g \in M_1} a_g \mathbf{E}_i g(\nu_i)$, and thus by $\sum_{g \in M_1} a_g g$. But since $|M| \leq m$, this contradicts the fact that f is not approximable by any set of fewer than m functions that is, in addition, closed under the permutations ν_i.

◯ This general method has been used in several papers, although not stated as such, for example [44], [42], [26]: in particular, (i) above is the crux of the proof in [44] that DIP_2 is not closely approximable by the span of few symmetric functions. The proof of [42] to show that $AC_0[3] \not\subseteq QT_1$, is a combination of of (iii) and the main idea in the proof of (iv) for a special case of maps π_i, and ν_i. ◯

The following is straightforward from 6.1(1) and reduces non approximability into a purely geometric problem that is amenable to decomposition.

OBSERVATION 6.16. *Any set M of Boolean provides a map T_M from $\{-1,1\}^n$ to $\{-1,1\}^{|M|}$ by mapping $x \to (h_1(x), \ldots, h_{|M|}(x))$. Thus any Boolean function*

f over $\{-1,1\}^n$ is transformed into a function f_M over the image of T_M. I.e, $f_M(x) =_{def} f_M(T_M(x))$. Now, the convex hulls of $f_M^{-1}(1)$ and $f_M^{-1}(-1)$ intersect if and only if f does not have an approximation with the same sign from span(M). Two convex sets intersect if any of their subsets intersect, and thus this observation allows a non approximability question to be decomposed.

◯ Next, we turn to a few specific non approximability results, that apply some of the general results discussed so far. First we show that a universal function $RO[3]$ in $AC^0[3]$ is not approximable from the span of a set M of polynomially many LT_1 functions which is, in addition, closed under a class of permutations Π, of variables. The permutations are chosen so that the universal function is invariant under them. It seems to be intuitively the case that if f is approximable from the span of any set of polynomially many LT_1 functions then it would also be approximable from the span of a set that is closed under these permutations. A proof of this, together with the theorem below would imply that $AC^0[3] \not\subseteq LT_2$.

First, we define the universal function $RO[3]$ and the class of permutations under which it is invariant.

FACT 6.17. *The read-once function $RO[3](x) = \bigwedge_{i=1}^{m_1} \bigvee_{j=1}^{m_2} \bigwedge_{k=1}^{m_3} x_{ijk}$ is invariant under the class Π consisting of*

(i) one permutation for each fixed i, j and each k_1, k_2 that maps x_{ijk_1} to x_{ijk_2} (and fixes the other variables),

(ii) one permutation for each i and each j_1, j_2 that maps x_{ij_1k} to x_{ij_2k} for all k,

(iii) one permutation for each i_1, i_2 that maps x_{i_1jk} to x_{i_2jk} for all j, k.

Next, we notice a property of small sets of LT_1 functions that are closed under the above permutations.

FACT 6.18. *Any set M of LT_1 functions g that satisfy $|\{g(\pi) : \pi \in \Pi\}| \leq n^t$ must satisfy the following. Let $g(x) = sign \sum_{ijk}(a_{ijk}x_{ijk} + a_0)$. Then*

(i) there is a set s_3 of values for the subscript k with $|s_3| \geq m - 3t$ such that $a_{ijk_1} = a_{ijk_2}$ for $k_1, k_2 \in s_3$, and for all $g \in M$.

(ii) *there is a set s_2 of values for the subscript j with $|s_2| \geq m_2 - 3t$ and with $a_{ij_1k_1} = a_{ij_2k_2}$ for all $j_1, j_2 \in s_2$, for all $k_1, k_2 \in s_2$, and for all $g \in M$.*

(iii) *there is a set s_1 of values for the subscript i with $a_{i_1j_1k_1} = a_{i_2j_2k_2}$, with $|s_1| \geq m_1 - 3t$, for $i_1, i_2 \in s_1$, for all j_1, j_2, k_1, k_2, and for all $g \in M$.*

The theorem below uses Theorem 6.1 (1) and Theorem 6.14 (iii) and (iv).

THEOREM 6.19. *The function $RO[3]$ does not have an approximation with the same sign, from the span of a set $M \subseteq LT_1$, where M is closed under the permutations in Π, i.e, $\{g(\pi) : g \in M \text{ and } \pi \in \Pi\}$, and $|M|$ is polynomially bounded.*

PROOF. We will show that there exists a function $l \in span(M)^{\perp}$, such that $sign(l) = sign(RO[3])$. The function l will be chosen such that l is 0 except on 2 sets of points, one being $\{\pi(a) : \pi \in \Pi\}$, where $RO[3](a) = 1$, and the other of the form $\{\pi(b) : \pi \in \Pi\}$, where $RO[3](b) = -1$. For x in the first set, denoted L^+, $l(x)$ will equal $1/2|L^+|$, and for x in the second set L^-, $l(x)$ will equal $-1/2|L^-|$.

It is sufficient to show that for any M satisfying the conditions of the theorem, there is a solution l, as described above, to the system of equations E : $\sum l(x)h(x) = 0,\ h \in M$. Since l is invariant under Π, M is closed under Π, and the functions in M satisfy the conditions of Fact 6.18, this system E becomes analyzable.

First, we rewrite the system E as follows. Let $M^* \subseteq M$ be the set of representative functions in M, i.e, if $g_1, g_2 \in M^*$, then $g_1 \neq g_2(\pi)$, for any $\pi \in \Pi$. In other words, M^* is the set of equivalence classes, each of which contains functions in M that are equivalent up to the permutations (in Π) of variables.

For each $g \in M^*$, let $N_{g,a}$ denote the number $|\{\pi \in \Pi : g(\pi(a)) = 1\}|$. The system E then reduces to a system E' of $|M^*|$ equations of the form $N_{g,a} - N_{g,b} = 0, g \in M^*$. Notice that these equations depend on the parameters m_1, m_2 and m_3. To massage the system E' further, we place the following additional restrictions on the vectors a and b that generate the support of l: the vectors a and b satisfy the following.

$$\forall i, \quad m_0 =_{def} |\{(j,k) : a_{ijk} = 1\}| = |\{(j,k) : b_{ijk} = 1\}|, \qquad I$$

and

$$\forall i, j, \quad t^a_{ij} =_{def} |\{k : a_{ijk} = 1\}| \text{ and } t^b_{ij} = |\{k : b_{ijk} = 1\}|.$$

Moreover, the matrix with the t^a_{ij} as entries is denoted T^a, the matrix with the t^b_{ij} as entries is denoted T^b, each entry in these matrices is equal to m_3, $m_3/2$, or 0. *II*

We also define the variables

$$t^a_{pqr} =_{def} |\{i : |\{j : t^a_{ij} = m_3\}| = p, |\{j : t^a_{ij} = m_3/2\}| = q,$$

and

$$|\{j : t^a_{ij} = 0\}| = r\}|,$$

and t^b_{pqr} is similarly defined.

Clearly, due to the conditions I and II, t^b_{pqr} and t^a_{pqr} are 0 if $p+q+r \neq m_2$, and if $p + 2q \neq m_0$. Furthermore, $\sum_{pqr} t^b_{pqr} = \sum t^a_{pqr} = m_1$. The only difference between the equations constraining the variables t^b_{pqr} and those for t^a_{pqr} are the following, which is a consequence of the fact that $RO[3](a) = 1$, and $RO[3](b) = -1$:

$$t^a_{0qr} = 0, \text{ and } t^b_{0qr} \neq 0,$$

for the only relevant values of q and r, namely $q = t/2, r = m_2 - t/2$. *III*

For any matrix T^c with entries t^c_{ij} being $m_3, m_3/2$ or 0, denote by $N_{c,a}$, the number of distinct permuted matrices $T^{a'}$ formed by permuting the rows and columns of T^a, such that T^c is a top left minor of $T^{a'}$. and similarly define $N_{c,b}$. In addition, the quantities $N_{c,a}$ and $N_{c,b}$ can be expressed in terms of the entries c_{ij} of the matrix T^c, and the variables t^a_{pqr} t^b_{pqr} respectively. Clearly, the expressions for $N_{c,a}$ and $N_{c,b}$ are identical for all matrices T^c that are identical up to row and column permutations. In fact, for determining the expressions for $N_{c,a}$ and $N_{c,b}$ all the required information in the matrix T^c can be captured by a set of values t^c_{pqr}, analogous to t^a_{pqr} or t^b_{pqr}.

To see that the system E' is satisfiable, notice that for some choice of the parameters $m_2, m_1, m_0 \in \mathbb{N}$ as functions of m_3, the following system E_{m_0,m_1,m_2,m_3} of diophantine equations in the variables t^a_{pqr} or t^b_{pqr} is satisfiable over $\mathbb{N}$, and therefore E' is satisfiable.

(i) For each class of matrices T^c that are identical up to row and column permutations, and the corresponding set of constants t^c_{pqr}, there is one equation of the form $N_{c,a} - N_{c,b} = 0$, in the variable sets t^a_{pqr} and t^b_{pqr}. Clearly, these equations depend on the values of $m_0, \ldots, m_3$, since the quantities $N_{c,a}$ and $N_{c,b}$ and even the number of variables t^a_{pqr} and t^b_{pqr} depend on them.

(ii) There are equations that enforce the conditions III. These clearly depend on the values of $m_0, \ldots, m_3$, as well.

○

○ Below, we give an alternative result (obviously weaker in one sense, but stronger in another) than [23], but using a different proof technique.

THEOREM 6.20. *For any given k and m with $k \leq m \leq n$, and for any set M of And functions $\wedge_{u,v}$ where $M_{big} =_{def} \{\wedge_{u,v} \in M : |u| + |v| \geq n - m\}$, with*

$$|M_{big}| \leq \frac{2^k}{\prod_{i=0}^{k-1} 1 + m/(n-i)},$$

there does not exist a function $g \in span(M)$ i.e, with the same sign as Parity.

PROOF. By 6.1 (1), for any set M as in the statement of the theorem, it is sufficient to construct a function $l \in span(M)^{\perp}$ such that l is not identically 0, and having the same sign as *Parity* wherever non-zero. Our l will be constructed as follows.
• the support of l is of size 2^{n-k}, and l will equal *Parity* on its support;
• the support of l is a subcube of $\{-1, 1\}^n$ where an *And* function $\wedge_{u^*,v^*}$ equals -1, with $|u^*| + |v^*| = k$;
• all the functions in M_{big} will be constant on the support of l, so that l is in $span(M_{big})^{\perp}$ already by construction.
The functions in $M \setminus M_{big}$ are of the form $\wedge_{u,v}$ with $|u| + |v| < n - m$. Since $k \leq m$, and $|u^* + v^*| = k$, denoting by n^* the set of $n - k$ "free" coordinates of $\wedge_{u^*,v^*}$, i.e, the coordinates outside $u^* \cup v^*$, we notice that the set $n^* \setminus u \cup v$ (the set of free coordinates of $\wedge_{u,v}$, among the n^*) must be non-empty, and therefore, the set

$$\{x : \wedge_{u,v}(x) = -1\} \cap \text{ support}(l) = \{x : \wedge_{u^*,v^*}(x) = -1\}$$

splits into two equal halves, one where *Parity* = 1, and the other where *Parity* = -1. Now since l is defined to be *Parity* on its support, it follows that $l \in span(M \setminus M_{big})^{\perp}$ as well, and therefore, $l \in span(M)^{\perp}$.

It remains to describe the support of l, i,e, to describe u^* and v^*; and to ensure that the functions in M_{big} are constant on the support of l. Clearly, by the pigeon hole principle, there is some coordinate position i where $|M_{big}|(n - m)/2n$ of the *and* functions in M_{big} "coincide", i.e, for all of these functions g, $g(x) = -1$ only if $x_i = -1$, or for all of these functions g, $g(x) = -1$ only if $x_i = 1$. In the former case, put i into v^*, and in the latter case into u^*. We can continue this process on the remaining $|M_{big}|(n + m)/2n$ functions in M_{big},

increasing the size of v^*, and u^* until all the functions in M_{big} are exhausted. When the process terminates, the size $|v^* \cup u^*| \leq k$ provided

$$|M_{big}| \prod_{i=0}^{k-1} \frac{(n-i)+m}{2(n-i)} \leq 1,$$

i.e, when

$$|M_{big}| \leq \frac{2^k}{\prod_{i=0}^{k-1} 1 + m/((n-i))}.$$

□

◯

◯ As a corollary to the above theorem, we obtain that $AC^0[4] \not\subseteq LT_1 - Ands$, since $AC^0[4]$ functions embed the *Parity* function of $\log^2 n$ bits. ◯

◯ The main idea of the proof of the above theorem extends to the case of spaces spanned by other functions that behave similar to *And* functions, for example the *Flat* functions $\phi_{u,v}$ defined for sets u and v of *pairs* of variables in $\{1,\ldots,n\} \times \{1,\ldots,n\}$ as follows:

$$\phi_{u,v}(x) =_{def} \bigwedge_{(i,j)\in u} \neg(x_i \oplus x_j) \bigwedge_{(i,j)\in v} x_i \oplus x_j.$$

In other words, $\phi_{u,v}(x) = -1$ if and only if the bit-pairs of x are equal when the corresponding coordinate pairs are in u, and are unequal when the corresponding coordinate pairs are in v. Notice that u and v might contain redundant pairs of bits that can be removed without affecting the definition of the function. Hence we assume that $|u|$ and $|v|$ are minimal. The set of points where $\phi_{u,v} = -1$ defines a "flat" of dimension $n - |u| + |v|$, i.e, these points form both a subspace of that dimension in $\mathbb{F}_2^n$, as well as the intersection of a subspace of that dimension in R^n with $\{-1,1\}^n$. Flats are a generalization of the subcubes where the functions $\wedge_{u,v} = -1$.

Repeating an analogous construction as in the proof of 6.20, we obtain the following.

THEOREM 6.21. *The function* $\bigoplus_{i,j}(x_i \oplus x_j)$ *does not have an approximation with the same sign from the span of subexponentially many Flat functions.*

◯

6.5. Small correlation and non approximability. The next two theorems provide general methods for showing that the scalar product $< f, g >$ is small, for some fixed f, and for all functions g in some class B that is modelled after LT_1. As mentioned earlier, this step is needed for employing Theorem 6.1 (2) and (3) to prove a non approximability result for f. However, many of the methods given below themselves use duality.

It is generally assumed that every $g \in B$ is either in the span of simple functions or is approximable (to some degree of closeness) in the $||.||_\infty$ norm from the span of simple functions.

Showing that the scalar product $< f, g >$ is small is a combination of two tasks. First, a transitive approximability relationship is shown roughly of the form: if $< f, g >$ is large, and g is closely approximable from the span of simple functions, then f must be also be closely approximable in some sense from the span of few simple functions. Second, a strong non approximability result is proven that f cannot be thus approximated. The second part is stronger than what is required, which is only that f not be approximable by linear combinations of those simple functions that are used to approximate the functions $g \in B$. But such sets of simple functions are hard to isolate, so one is forced to consider all sets. The second part could be based on any of the methods of the previous four subsections. The next theorem presents natural combinations of these two parts. The proofs are straightforward and use Theorem 6.12 and Theorem 6.1.

THEOREM 6.22. *Let f be a Boolean function, B a set of Boolean functions, and Q as set of "simple" Boolean functions.*

(1) If every $g \in B$ is the linear combination of functions h in Q with the coefficients summing, in absolute value, to at most m, and $< f, h > \leq \epsilon$ for each $h \in Q$. Then $< f, g > \leq m\epsilon$. (Application of 6.12(1)).

(2) If for every $g \in B$ there is a function h, with the same sign as g, and in the span of m functions $h \in Q$, satisfying $|h(x)| \geq \delta$ for all x; and furthermore, if f is a function such that $< f, h > \leq \epsilon$, for every $h \in Q$, then $< f, g > \leq \epsilon + 1 - \delta$. (Application of 6.12(2)).

(3) If g is in the span of m functions $h \in Q$, and if for every subset S of the domain that contains more than an $(1 + \epsilon)/2$ fraction of the points and any set M of m functions in Q, f is not approximable from $span(M)$ on S, then $< f, g > \leq \epsilon$. Any appropriate method of this section could be used to show that f is not approximable on S from $span(M)$. For example,

(i) *if there is some function* $l \in span(M)^{\perp}$ *that is* 0 *outside* S *and has the same sign as* f *on* S *(application of 6.1 (1));*

(ii) *if there is some distribution* $\mathcal{R}$ *on* S*, over which the functions* $h \in M$ *are orthonormal, and* $< f, h >_R \leq 1/|M|$*, for all* $h \in M$*, (application of 6.14 (ii) and 6.1 (3)).*

◯ The combination of 6.22(iii) and 6.14 yield the following for the special case where the functions in B are LT_1 functions, and Q is the set of the linear monomials.

THEOREM 6.23. *Given a Boolean function* f*, and a function* $g \in LT_1$*. if for each subset* $S \subseteq \{-1,1\}^n$ *with* $|S|$ *containing more than an* $(1+\epsilon)/2$ *fraction of points, one of the following holds, then* $< f, g > \leq \epsilon$*.*

(i) $ConvexHull(f^-1(1) \cap S) \cap ConvexHull(f^-1(-1) \cap S)$ *is non-empty.*

(ii) The linear monomials and the constant function One continue to remain almost orthonormal with respect to S*, with* $| < x_i, x_j >_S | \leq \delta$*, and* $| < x_i, One >_S | \leq \delta$*, but* $| < f, x_i >_S | \leq (1-\delta)^2/(n+1)$*, for all* i *and* $| < f, One >_S | \leq (1-\delta)^2/(n+1)$*.*

(iii) There is a set Π *of permutations of the variables such that* f *is invariant under the permutations in* Π*, and for all linear functions* g*,* $\sum_{\pi \in \Pi} g(\pi)(x)$ *has the form* $a \sum_i x_i + a_0$*, for some* a *and* a_0*. Finally,* f *is not approximable by any function of the form* $a \sum_i x_i + a_0$*, over* $\bigcap_{\pi \in \Pi} S(\pi)$*.*

(iv) There is a set Π *of permutations of the variables such that* f *is invariant under the permutations in* Π*, and for some point* a *in* $f^-1(1) \cap S$ *and* b *in* $f^-1(-1) \cap S$*,* $\sum_{\pi \in \Pi, \pi(a) \in S} \pi(a)_i = 0$*, and* $\sum_{\pi \in \Pi, \pi(b) \in S} \pi(b)_i = 0$*, for all* i*.*

As a direct application of 6.23(i), we obtain the following.

FACT 6.24. $< RO[3], g > \leq 2\hat{RO}[3](0^n)$*, for all* $g \in LT_1$*.*

PROOF. We illustrate the convex-hull intersection argument (neccessary for employing 6.23(i)) by proving the straightforward result that $RO^0[2] \notin LT_1$. The proof of the fact is carried out along the same lines, by showing $RO^0[3]|_S \notin LT_{1,S}$ for any large subdomain S. We will consider the canonical $AC^0[2]$ function $RO[2](x) := \bigvee_{j=1}^{k} \bigwedge_{i=1}^{k} x_{ij}$. If an LT_1 function g equals $RO[2]$,

then, in particular, $g(x) = -1$ when $x_{i1} = -1$, for $1 \leq i \leq k$ and $x_{ij} = 1$, for all i and all $j \neq 1$;
and $g(y) = -1$ when $y_{i2} = -1$, for $1 \leq i \leq k$ and $y_{ij} = 1$, for all i and all $j \neq 2$.
Moreover, for every z, w such that $x + y = z + w$ (where the + stands for addition in $\mathbb{R}^n$), either $g(z) = -1$ or $g(w) = -1$, since g, being in LT_1, is the characteristic function of a halfspace of $\mathbb{R}^n$. However, for our chosen function $RO[2]$, we can find z and w with $z + w = x + y$, with both $f(z) = f(w) = 1$. For example, choose z with $z_{i1} = -1$ and $z_{i2} = -1$ for $1 \leq i \leq k/2$ and $z_{ij} = 1$, for all i and all $j \neq 1, 2$, and choose w such that $x+y = z+w$. This contradicts the assumption that $g = f$. □

○

7. Algorithms for approximation

In this section, we are interested in efficient constructive solutions to the general approximation problem **A** in Section 3, when the Boolean functions g in the class C are known to be appropriately approximable. The universal space U is the space of real functions over $\{-1, 1\}^n$, so we are do not deal with approximation/interpolation algorithms over different domains (such as, for example, [5]) or ranges (such as, for example, [12]). Notice that since we are interested only in Boolean g, and are dealing with real-valued functions, the construction of an interpolant, i.e, with $\epsilon = 0$ is algorithmically equivalent to the construction of an ∞-norm approximant with $\epsilon \leq 1$. In addition, to ∞-norm approximations with error $\epsilon \leq 1$, we also consider 2-norm approximations, since they are often easier to find, and moreover provide ∞-norm approximations over large domains S in **A**. Finally, we also consider close ∞-norm approximations (i.e, small ϵ), and high-energy approximations, as in 6.1 (2) and (3), since both are highly related to 2-norm approximations (see Section 6).

We assume that the approximation algorithms are able to evaluate the function g pointwise at any set of sample points, which we shall refer to as σ (each evaluation costs one unit of time). We also consider randomized algorithms, that may choose to (but are not required to) sample randomly on the pertinent distribution $\mathcal{D}$ in the framework **A**. These randomized algorithms produce the required approximation with suitably high probability depending on the running time. There are two independent issues to be considered in designing such algorithms.

• First, how many sample points (pieces of information about g) are required to *determine* the set of valid approximations $h \in X =_{def} span(M)$.

• Secondly, how to find such a set Σ of sample points, and using the values of g over Σ, how to construct an approximation h efficiently.

The main general result of this section, Theorem 7.1, shows that a function $l \in X^{\perp}$ of support $dim(X)+1$ provides a set $\Sigma = supp(l)$ of sample points, so that the values of g over Σ are sufficient to determine an ∞-norm approximation to g from X. This result depends mainly on the duality principle. In an ensuing observation, Remark 7.2, connections are drawn between sample sets Σ and sets of pseudorandom elements given by Theorem 5.1.

This is followed by a list of relevant approximation algorithms in the approximation theory and computational learning theory literature. We classify these algorithms based on their choices of parameters in the general approximation framework **A**; describe their connection to Theorem 7.1, and Remark 7.2; derandomize some of these algorithms, and point out promising natural extensions that that have not been investigated. The main application of such approximation algorithms is for learning classes of Boolean functions. As usual, we sandwich all discussions concerning this application by ◯'s.

◯ If the approximation algorithm for the problem **A** produces an approximation h that uniquely defines a Boolean function g in C, when restricted to the relevant σ-fraction of the domain, i.e, a class of functions in C that coincide on the subdomain S of approximation, then the algorithm is a learning algorithm for C with **accuracy** σ. This is the case if the norm of approximation is the ∞-norm, the error of approximation $\epsilon \leq 1$ and the functions in C are Boolean. On the other hand, if the approximation h only uniquely defines a class of functions in C that coincide on a large σ' fraction of S, then the algorithm is a learning algorithm for C with accuracy $\sigma\sigma'$. This is the case if the norm of approximation is the 2-norm, the error of approximation ϵ is $1-\sigma'$, and the functions in C are Boolean.

These learning algorithms are said to use **membership** oracles, if the sample set Σ is arbitrary, and are said to use **example** oracles, if the sample set Σ is drawn at random from the relevant distribution $\mathcal{D}$ in the framework **A**. Thus our approximation algorithms use both kinds of oracles.

Moreover, when the quantifier on the distribution $\mathcal{D}$ is universal, and on the subdomain S is existential, in the framework **A**, such an approximation algorithm is a PAC learning algorithm provided it works for all $\sigma < 1$, with an appropriate increase in running time, and furthermore, if it works for $\sigma = 1$, then it is called an exact learning algorithm. ◯

We now give the main general result concerning sample sets Σ. This can be found in [62] and provides a restriction on the size of the support of the

function $l \in X^{\perp}$.

THEOREM 7.1. *Let X be a space of functions and let g be any function. Let Σ be the support of a function $l^* \in X^{\perp}$ that maximizes $|\sum_x l(x)g(x)|$ over all $l \in X^{\perp}$ with $||l||_1 \leq 1$. Then a best uniform approximation to g from X, i.e, a function g^* such that $||g - g^*||_\infty$ is minimum, is determined by the values of g on Σ. In fact, a best approximation to g remains a best approximation to g_Σ. Furthermore, there is is a set Σ satisfying the above conditions, of size at most $dim(X) + 1$.*

PROOF. It is clear that a best approximation to g remains a best approximation to g on the support Σ of the l^*, since $l^* \in X_\Sigma^{\perp}$, and maximizes $|\sum_{x\in\Sigma} l(x)g(x)|$. Let g^* be a best approximation to g, and let $||g - g^*|| = \epsilon$. Let Σ be the support of a function l^* as in the statement of the theorem. By 3.3, $|\sum_{x\in\Sigma} l^*(x)g(x)| = \epsilon$. However, $\sum_{x\in\Sigma} l^*(x)g(x) = \sum_{x\in\Sigma} l^*(x)(g - g^*)(x)$. Thus, $|(g - g^*)(x)| = \epsilon$, for each $x \in \Sigma$, and in fact, the sign of $(g - g^*)(x)$ is equal to the sign of l^*. It only remains to show that there is such a function l^*, whose support Σ satisfies $|\Sigma| \leq dim(X)+1$. We construct this function l^* as $h(g-g^*)$ (normalized), where h is a positive function that has at most $dim(X)+1$ points of support.

First, notice that $g - g^*$ is "orthogonal" to X in the sense of the ∞-norm. Recall that by this notion of orthogonality, $f \perp X$, if $||f||_\infty \leq ||f + p||_\infty$ for all $p \in X$. We use a property of functions that are orthogonal in this sense.

Claim. A function f is orthogonal to a space X in the sense of the ∞-norm if and only if there is a positive function h with at most $dim(X) + 1$ points of support contained in the extremal set of f, i.e, $E(f) =_{def} \{x : f(x) = ||f||_\infty\}$, such that fh is in $X^{\perp}$ (the orthogonal space in the usual inner product sense).

Now, let the function h be as in the claim. The function $l^* =_{def} h(g-g^*)/||h(g-g^*)||_1$ clearly has only $dim(X)+1$ points of support; and is in $X^{\perp}$ by the claim. It only remains to show that l^* is extremal, i.e, $|\sum_x l^*(x)g(x)|$ is maximum over all $l \in X^{\perp}$ with $||l||_1 \leq 1$. For this, by 3.1, it is sufficient to show that $|\sum_x l^*(x)g(x)| = ||g - g^*||_\infty$. Since $l^* \in X^{\perp}$,

$$|\sum_x l^*(x)g(x)| = |\sum_x l^*(x)(g - g^*)(x)| = \frac{|\sum_x h(x)(g - g^*)(x)(g - g^*)(x)|}{||h(g - g^*)||_1}.$$

By the claim, h is positive and its support is restricted to lie within $E(g-g^*)$, therefore, it holds that $||h(g-g^*)||_1 = |\sum_x h(x)|\ ||g-g^*||_\infty$ and

$$\frac{|\sum_x h(x)(g-g^*)(x)(g-g^*)(x)|}{||h(g-g^*)||_1} = \frac{||g-g^*||_\infty^2 |\sum_x h(x)|}{||h(g-g^*)||_1} = ||g-g^*||_\infty.$$

□

Proof of claim. For one direction, suppose $fh \in X^\perp$, and h is positive, with $||h||_1 = 1$, and its support is contained in the extremal set of f as stated in the claim (the size of the support is not neccessary here). Then for every $e \in X$,

$$||f||_\infty^2 = \sum_x h(x)f(x)f(x) = \sum_x h(x)f(x)(f(x)-e(x))$$

$$\leq ||f||_\infty \sum_x h(x) \max_{x \in supp(h)} |f(x)-e(x)| \leq ||f||_\infty ||f-e||_\infty,$$

thereby showing that f is orthogonal to X in the sense of the ∞-norm.

For the reverse direction, suppose f is orthogonal to X in the sense of the ∞-norm. Let $dim(X) = m$, and let $b_1 \ldots, b_m$ be a basis for X and consider the set of points S in $\mathbb{R}^m$ given by

$$S := \{(f(x)b_1(x), \ldots, f(x)b_m(x) : x \in E(f)\}.$$

Notice that the origin must be contained in the convex hull of S. Otherwise, there is a hyperplane separating the points in S from the origin, i.e, there are constants $a_1, \ldots, a_m \in \mathbb{R}$ such that

$$\sum_i a_i f(x) b_i(x) > 0,$$

for all $x \in E(f)$. This would mean that $f(x)\sum_i a_i b_i(x) > 0$, for all $x \in E(f)$, implying that there is a function $e = \sum_i a_i b_i \in X$ with the same sign as f on $E(f)$. This would imply that for some $e \in X$, $||f+e||_\infty \geq ||f||_\infty$, thereby contradicting the assumption that f is orthogonal to X in the sense of the ∞-norm.

Thus the origin must be contained in the convex hull of S. It follows that there is a set $m+1$ points in S such that the origin is a convex combination of these points, which means that there is a positive function h with support consisting of $m = dim(X)+1$ points in $E(f)$ such that

$$\sum_x h(x)f(x)b_i(x) = 0, \quad \forall b_i,$$

and hence, $hf \in X^\perp$. This proves the claim. ■

REMARK 7.2. *The fact that a function in $X^\perp$ provides a good sample set Σ for approximation of a function g from X is not surprising. Intuitively, a good sample set Σ is one over which the behavior of the function g is similar to its behavior over the relevant distribution $\mathcal{D}$ in the approximation framework* **A**. *Theorem 5.1 and the following remark show how such sample sets - which behave as though uniformly distributed in the sense of expected values - are obtained naturally from functions in $X^\perp$. This can be extended to give sample sets that behave like other distributions as well. These sets were used in Section 5 to serve as distributions of pseudorandom elements that fool the functions that are approximable from X. In fact, we will show that similar distributions serve as sample sets for approximating functions in the $||.||_2$ norm. Moreover, the above theorem shows that the supports of certain functions in $X^\perp$ also serve as small sample sets σ for approximating g from X in the ∞ norm. When easy to generate, such sampling distributions, as noted in Section 5, therefore help not only to derandomize computations based on approximable functions like g, but are used in some of the algorithms listed below, especially in the context of finding deterministic sample sets Σ and therefore derandomizing randomized approximation or learning algorithms.*

Based on the following parameters in the framework **A**, we list and classify known approximation algorithms and derandomize some of them. The parameters are the following.

• The norm, which could be $||.||_\infty$, or $||.||_2$; we also consider high-energy approximations as in 6.1(3).

• The space $X = span(M)$ which could be fixed or allowed to vary, with the only constratint being on $dim(X) = |M|$.

• The basis functions in the set B which could chosen orthonormal or not.

• The distribution $\mathcal{D}$, which is arbitrary, i.e, universally quantified, or fixed to be relatively close to the uniform distribution.

• The size of the domain of approximation S, quantified by σ, which could be the entire domain, i.e, $\sigma = 1$, or a subdomain of suitably large measure, with respect to the distribution $\mathcal{D}$.

We are also concerned with:

• The sample set Σ, which could be deterministically chosen, or randomly chosen, and its connection to supports of functions in $X^\perp$ i.e, possible uses of Theorem 7.1, Remark 7.2 and Theorem 5.1.

CASE 1 We first consider the case of ∞ norm approximations over the whole domain, i.e, $\sigma = 1$, and the space $X = span(M)$ is fixed. Here we can assume

that the distribution $\mathcal{D}$ is universally quantified, by the nature of ∞-norm approximation.

In this case, adaptations of the univariate primal-dual method of Remez (see [55]) can be used, which work also for general, non-Boolean functions. These algorithms use an iterative procedure to find a function $l^* \in X^\perp$ of bounded norm that maximizes $|\sum_x l(x)g(x)|$. Once the function l^* is found, by the proof of Theorem 7.1, the best approximation h can be found, by solving an interpolation problem on $supp(l^*)$, i.e, by inverting a VanderMonde of size $|supp(l^*)| = dim(X) + 1$.

At the i^{th} iteration of the procedure, the function $l^{i+1} \in X^\perp$ is constructed from $l_i \in X^\perp$, again by solving an interpolation problem on $supp(l_i)$ to construct a pseudo-approximation h_i. The new function l_{i+1} is constructed by removing a point from l_i and including a point outside $supp(l_i)$ where $g - h_i$ is maximal, (if all such points lie inside $supp(l_i)$, the algorithm takes l^* to be l_i and halts, justifiably by the proof of Theorem 7.1).

The rate of convergence of this iterative procedure to the optimum function l^* depends on the function g and the space X, and could take exponentially long. However, the algorithm is deterministic and its sample set $\Sigma = \bigcup_i supp(l_i)$ for the successive $l_i \in X^\perp$.

Improvements of the algorithm, and complexity analyses for the special case of Boolean approximation are open. In the Boolean case, it is not neccessary to find the best approximation: it is sufficient to find an approximation for which $\gamma < 1$ in the framework **A**.

In this context, it should be noted that for bases B and spaces X satisfying certain properties, there is a multidimensional analog of a linear approximation operator called the Korovkin summation operator for arbitrary $||.||_p$ norms. See [35]. This operator provides a general constructive method of approximation which sometimes reduces to well-known methods for specific cases.

◯ The algorithm described above is an exact, deterministic learning algorithm for the class of Boolean functions g for which there is an approximation with the same sign from the space X. ◯

CASE 2 Next, we consider the case of $||.||_\infty$ norm approximations over the whole domain, i.e, for $\sigma = 1$, but with the space $X = span(M)$ varying over all sets M of independent basis fuctions in B, with $dim(X) = |M|$ is at most some fixed value m.

Such problems have been considered by approximation theorists for general, non-Boolean functions, originating from the question of approximation

by splines with "free knots," or by functions with "few harmonics" (i.e, few Fourier coefficients). For the origins of this subject, see [19] and [53]. In the recent literature, such algorithms are studied in the context of approximation by wavelets.

The significance of these results to Boolean function approximation is yet to be investigated.

CASE 3 Next, we consider 2-norm approximation of Boolean functions, over the whole domain, i.e, $\sigma = 1$, for arbitrary distributions $\mathcal{D}$, when the space $X = span(M)$ is fixed. We could assume, if we choose, that the norm is defined based on the inner product $<>_{\mathcal{D}}$. In this case, for all distributions $\mathcal{D}$, the 2-norm approximant for a Boolean function, with error $\gamma < 1/2$ is also a meaningful ∞-norm approximant with error $\gamma' < 1$, over a subdomain S with measure $\sigma' \geq 1 - \gamma$ wit respect to $\mathcal{D}$.

The best 2-norm approximation of g from X is easily described as the projection $g|_{X,\mathcal{D}} = \sum_{\alpha} < g, h_\alpha >_{\mathcal{D}} h_\alpha$, where h_α form an orthonormal basis for X under $<>_{\mathcal{D}}$. Finding this projection is usually achieved by finding an orthonormal basis for X with respect to $<>_{\mathcal{D}}$.

We assume that the distribution $\mathcal{D}$ is fixed to be the uniform distribution, and that the given basis M for X is itself orthonormal with respect to the usual $<>$. Also without loss, we assume the entire basis B is the Fourier basis, and hence the quantities $< g, h >_\alpha$ are nothing but the Fourier coefficients $\hat{g}(\alpha) =< g, \chi_\alpha >$. The uniform distribution and the Fourier basis can be replaced by any "close-to-uniform" distribution $\mathcal{D}$ for which a well-behaved orthonormal basis exists with respect to $<>_{\mathcal{D}}$, where, by "well-behaved," we mean that the basis functions are easily computable, and have well-bounded norms.

In the case of Boolean g, for estimating the coefficients $\hat{g}(\alpha)$ for $\chi_\alpha \in M$ with suitable accuracy, small *random* sample sets Σ, consisting of $poly(|M| = dim(X))$ points can be shown to be sufficient, using Chernoff bounds, see [46].

However, the supports of certain functions $l \in X^\perp$ can be used to give small *deterministic* sample sets as follows: express the function l in $X^\perp$, and $||l||_1 \leq 2$ as $l^* - 1/2^n$ for a positive function $l^* \in (X \setminus One)^\perp$. By Theorem 5.1, since g is approximable from X, it follows that $supp(l^*)$ looks uniformly distributed to g in the sense of expected values, i.e, the quantity

$$|\sum_x l(x)g(x)| = |\sum_x (l^* - 1/2^n)(x)g(x)\chi_{0^n}(x)| \text{ is small.}$$

Therefore, $\hat{g}(0^n)$ can be estimated with reasonable accuracy by sampling g on $supp(l^*)$ and computing $\sum_x l^*(x)g(x)\chi_{0^n}(x)$. This idea was discussed in Section

5 to systematically obtain sets of pseudorandom elements for functions g that are approximable from X.

If l^* additionally satisfies the conditions that

$$|\sum_x (l^* - 1/2^n)(x)g(x)\chi_\alpha(x)| \text{ is small } \forall \chi_\alpha \in M,$$

then $supp(l^*)$ can be used as a deterministic sample set for approximating g.

For distributions $\mathcal{D}$ that are close-to-uniform, the constant function $1/2^n(x)$ (representing the uniform distribution) can be replaced by $\mathcal{D}(x)$, and the functions χ_α by the orthonormal basis functions under $<>_\mathcal{D}$. In fact, different l^*_α's can be found that satisfy the above conditions for each $\chi_\alpha \in M$, whereby the set $\Sigma = \bigcup_{\chi_\alpha \in M} l^*_\alpha$ is a deterministic sample set for approximating g.

The aim, then, is to find such functions l^* with small support, and which, in addition, are easy to generate. This question is, in general, open for the current case of $||.||_2$ norm approximation. In contrast, in the case of uniform approximation, such small sampling distributions always exist by Theorem 7.1, whose support and values can be generated using the Remez iterative process discussed in Case 1. However, when M is a geometrically "nice" set of Fourier basis functions, i.e, if for each χ_α, χ_β in M, the Hamming distance $|\alpha - \beta|$ is small, then the m conditions on l^* imply a minimum (Hamming) distance condition on $supp(\hat{l^*})$. See [64]. This is the idea behind Shannon-Whitaker type sampling theorems [63]. In fact, in the case where $|\alpha| \leq e/2$ for all $\chi_\alpha \in M$, this minimum distance condition is equivalent to saying that $supp(\hat{l^*})$ is a familiar e-wise independent distributions which were already discussed in Section 5 in the context of pseudorandom sets.

◯ The paper [46] showed that AC^0 functions g are approximable in the 2-norm from the space X of low ($polylog$) degree polynomials, i.e, the space spanned by the set M of Fourier basis functions χ_α for small Hamming weight $|\alpha|$. For the special case of AC^0, The random sampling method discussed above was used to estimate the the coefficients $\hat{g}(\alpha)$ for small $|\alpha|$, by sampling at $poly(|M|) = O(n^{polylog}(n))$ points, thereby giving a learning algorithm for AC^0 with respect to the uniform distribution, which was extended to "close-to-uniform" distributions in the paper [22]. The former algorithm was derandomized in [64] by finding deterministic sample sets of the same size using the ideas explained above, thereby illustrating the connection between pseudorandom sets, sample sets for learning, and $polylog$-wise independent distributions. The latter algorithms derandomize using the same ideas. ◯

CASE 4 Next, we consider an algorithm by [40] which works under the same conditions as in Case 3, except that the space X is not fixed, but spanned by a set M of independent basis functions in B with $dim(X) = |M|$ bounded by m. Again, we assume the distribution $\mathcal{D}$ to be uniform and the basis B orthonormal, and as before, we can assume any distribution $\mathcal{D}$ for which an orthonormal basis exists with respect to $<>_{\mathcal{D}}$, which is well-behaved, i.e, the basis functions are easily computable and have a bounded norm.

This kind of approximation is useful for Boolean functions g for which $||\hat{g}||_1$ is small, for example, $g \in PL_1$. Whenever X is the span of χ_α's for which $|\hat{g}(\alpha)|$ is at least $\gamma/||\hat{g}||_1$, the number of such α's is at most $||\hat{g}||_1^2/\gamma$, and $||g-g|_X||_2 \leq \gamma$, Therefore such functions g can be well approximated in the 2-norm from the span of only a few Fourier basis functions. (The analogous result for ∞-norm approximation, i.e, that $PL_1 \subset \hat{PT}_1$, is harder to prove and was shown by [15]. A more general result is given in Theorem 6.9).

The algorithm in [40] uses a clever search technique to isolate m functions χ_α that form the basis M for a space X such that the 2-norm approximation or the projection $g|_X$ is suitably close to g, or in other words, has a large 2-norm $\sum_{\chi_\alpha \in M} \hat{g}^2(\alpha)$.

Once the set M of χ_α's is found, then either the random samples or the deterministic sample sets Σ of Case 3 can be used to estimate the $\hat{g}(\alpha)$'s, and thus $g|_X$.

The set M of basis functions χ_α is found by a divide and conquer process, estimating the quantity $\sum_{\chi_\alpha \in Q} \hat{g}^2(x)$ for successively smaller sets Q of Fourier basis functions, starting with Q being the entire set B. The estimation of this quantity is carried out by a quasirandom, uniform sampling of g over specific subdomains, (See [40]) for details).

It is an open question whether there are natural deterministic sample sets as in Case 3, that can replace the random samples for this part of the algorithm.

◯ As mentioned earlier, the algorithm [40] is fast for finding 2-norm approximations of functions in PL_1 for example, linear decision trees, since in this case, the dimension of the approximating space X is small. In addition, as mentioned in Case 3, a 2-norm approximation with error $\gamma < 1/2$ over the entire domain gives a meaningful ∞-norm approximation with error $\gamma' < 1$, over a fraction, $\sigma' = 1 - \gamma$, of the domain. (This holds as well for other distributions $\mathcal{D}$, with "fraction" replaced by "measure"). This gives a fast, randomized, learning algorithm for PL_1 functions with respect to the uniform distribution, where the running time grows linearly with σ'.

The extension of this algorithm to close-to-uniform distributions $\mathcal{D}$ (for which a well-behaved orthonormal basis exists w.r.t. $<>_{\mathcal{D}}$) is given in [34]. ◯

CASE 5 Next, we turn to the case when a *close* ∞-norm approximation h exists for a Boolean function g, with $||g-h||_\infty \leq \gamma$, i.e, h has the same sign as g and $||h||_1 \geq 1-\gamma$, where with $h = \sum_{b\in M} a_b b$, and $\sum\limits_b |a_b| \leq 1$. Furthermore, for the first part of this discussion, no restrictions such as orthonormality are placed on the basis functions $b \in M$, and the distribution $\mathcal{D}$ is considered to be arbitrary, as is customary for ∞-norm approximation over the entire domain. Note that here, γ might be specified to be greater than $1/2$, even as large as $1 - poly(n)$, and therefore finding a 2-norm approximation may not provide a meaningful ∞-norm approximation. Such close approximations exist, for example, when $g \in \hat{PT}_1$, which includes PL_1 and a $\hat{LT}_1$ of PL_1 functions as well, by Corollary 6.13.

We are interested in algorithms that find an ∞-norm approximation to a Boolean function g with respect to a large class of distributions, over a subdomain of large measure.

By 6.1(2), we know that for *every* distribution $\mathcal{D}$, there is a basis function $b \in M$ such that $| < g, b >_{\mathcal{D}} | \geq 1-\gamma$. Such a function is called a **weak hypothesis** in learning theory terms. *Assuming* an oracle provides such a basis function $b \in M$ for any distribution $\mathcal{D}$, then [34] describes a randomized "boosting" algorithm, due originally to [21], that, with a high probability, finds a (close) ∞-norm approximation h', called a **strong hypothesis**, from $X = span(M)$: for any distribution $\mathcal{D}$, the approximation h' approximates over a subdomain that has large measure σ with respect to $\mathcal{D}$, and the running time of the algorithm increases linearly with σ.

To simulate the oracle above, however, restrictions on the distribution $\mathcal{D}$ are required. Using the algorithms described and partially derandomized in Case 4, the oracle above can be simulated efficiently in the case where the basis functions in M are almost orthonormal, and the distributions $\mathcal{D}$ are close to the uniform distribution. Thus, for these special sets M, and distributions $\mathcal{D}$, a basis function $b \in M$ such that $| < g, b >_{\mathcal{D}} | \geq 1-\gamma$ can be found efficiently by a quasi-random sampling of g. This, however, represents only a partial simulation of the oracle required by the boosting algorithm of [21], described in the previous paragraph.

To surmount this difficulty, [34] gives an adaptation of the boosting algorithm of [21]. The modified boosting algorithm makes do with information about g on close-to-uniform distributions, and provides a uniform approxima-

tion over a set of large measure σ, only with respect to such distributions. The algorithm uses the basis functions b in the almost orthonormal basis M, for which $| < g, b >_{\mathcal{D}} | \geq 1 - \gamma$, which are given by the simulated oracle for close-to-uniform distributions $\mathcal{D}$. The running time of the algorithm increases linearly with σ.

In fact, notice that the algorithm described above works as long as there is a function b in an almost orthonormal basis M such that $| < g, b >_{\mathcal{D}} | \geq 1 - \gamma$, for close-to-uniform distributions $\mathcal{D}$. This does not require g to be closely approximable from $X = span(M)$. It is sufficient that g have a high energy approximation h from X, i.e, with $< g, h > \geq 1 - \gamma$.

◯ The partially derandomized procedure described above provides a learning algorithm with respect to almost uniform distributions for any function in $\hat{PT}_1$, or for any function that has a high-energy approximation from the space spanned by a few Fourier basis functions which clearly consist of functions that are closely approximable from the space spanned by a few Fourier basis functions.

It is a natural open question whether the observation above concerning high-energy approximations has an application in learning theory. ◯

Acknowledgements

This work was partially supported by NSF grant CCR 9409809.

References

[1] N. Alon, "Threshold gates, coin weighing, and indecomposable hypergraphs," *Manuscript*, 1996.

[2] N. Alon, R.Boppana, "The monotone circuit complexity of Boolean functions," *Combinatorica* 7, pp. 1-22, 1987.

[3] J. Aspnes, R. Beigel, M. Furst, S. Rudich, "The expressive power of voting polynomials," *Combinatorica* 14, pp. 1-14, 1994.

[4] L. Babai, N. Nisan, M. Szegedy, "Multiparty protocols and pseudorandom sequences," *Proc.* 21^{st} *Ann. ACM Symp. on Theory of Computing,* pp. 1-11, 1989.

[5] M. Ben-Or, P. Tiwari, " A deterministic algorithm for sparse multivariate polynomial interpolation," *Proc.* 20^{th} *Ann. ACM Symp. Theory of Comput,* pp. 301-309, May 1988.

[6] D.A. Barrington, R. Beigel, S. Rudich, "Representing Boolean functions as polynomials modulo composite numbers," 24^{th} *Ann. ACM Symp. on Theory of Computing,* pp. 455-461, 1992.

[7] W. Beckner, "Inequalities in Fourier analysis," *Annals of Mathematics,* 102, pp. 159-182, 1975.

[8] R. Beigel, "The polynomial method in circuit complexity," 8^{th} *Ann. IEEE conf. on Struct. in Compl. Theory,* pp. 82-95, 1993.

[9] R. Beigel, J.Tarui, "On ACC," *Proc.* 32^{nd} *Ann. IEEE Symp. on Foundations of CS,* pp. 783-792, 1991.

[10] R. Beimel, A. Gal, M. Paterson, "Lower bounds for monotone span programs," *Proc.* 36^{th} *Ann. IEEE Symp. on Foundations of CS,* pp. 674-681, 1995.

[11] J. Bruck, "Harmonic analysis of polynomial threshold functions," *SIAM Journal of Discrete Mathematics,* 3 (2), pp. 168-177, 1990.

[12] N. Bshouty, Y. Mansour, "Simple learning algorithms for decision trees and multivariate polynomials," *Proc.* 36^{th} *Ann. IEEE Symp. Foundations of CS,* pp. 304-311, 1995.

[13] R.C. Buck, "Applications of duality in approximation theory," in *Approximation of functions, Elsevier,* H.L. Garabedian ed., pp. 27-42, 1964.

[14] P.L. Butzer, "Fourier analysis and approximation," *Academic Press,* 1971.

[15] J. Bruck, R. Smolensky, "Polynomial threshold functions, AC^0 functions, and spectral norms," 31^{st} *Ann. IEEE Symp. Foundations of CS,* pp. 632-641, 1990.

[16] C. de Boor, "Topics in multivariate approximation theory," *Lect. Notes in Math,* 965, pp. 39-78, 1981; "Part II," *Acta Numerica*, 65-109, 1993.

[17] C.K. Chui "Multivariate Splines," *SIAM CBMS-NSF regional conference,* Philadelphia, 1988.

[18] R. DeVore "One sided approximation of functions," *J. Approx. theory,* 1, pp.11-25, 1968.

[19] R. DeVore, V.A. Popov, "Interpolation spaces and nonlinear approximation," In *Function spaces and applications*, Cwickel, Peetre, Sagher, Wallin, ed.s, *Springer Verlag Lecture Notes in Mathematics,* 1302, pp. 191-205, 1988.

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

[21] Y. Freund, "An improved boosting algorithms and its applications on learning compelxity," 5^{th} *Conf. on Computational Learning Theory,* pp. 391-398, 1992.

[22] M. Furst, J. Jackson, S. Smith, "Improved learning of AC^0 functions," 4^{th} *Conf. on Computational Learning Theory,* pp. 317-325, 1991.

[23] M. Goldman, "On the power of a threshold gate at the top," *Manuscript*, 1995.

[24] M. Goldman, J. Hastad, "On the power of small depth threshold circuits," 31^{st} *Ann. IEEE Symp. Foundations of CS*, pp. 610-618, 1990.

[25] M. Goldman, J. Hastad, "A simple lower bound for monotone clique using a communication game," *Information Processing Letters* 41, 221-226, 1991.

[26] M. Goldman, J. Hastad, A.A. Razborov, "Majority gates vs. general weighted threshold gates," 32^{nd} *Ann. IEEE Symp. Foundations of CS*, 1991.

[27] C. Gotsman, N. Linial, "Equivalence of two problems on the cube - A note," *J. Comb. Theory, Ser. A*, 61, pp. 142-146, 1992.

[28] V. Grolmusz, Harmonic analysis, real approximation and communication complexity of Boolean functions, *Manuscript*, 1994.

[29] M. Grigni, M. Sipser, "Monotone separation of logspace from NC^1," *Proc.* 6^{th} *IEEE Conf. on Structure in Complexity Theory,* pp. 294-298, 1991.

[30] J. Hastad, "Computational limitations of small depth circuits," *Ph. D thesis, MIT press,* 1986.

[31] J. Hastad, "On the size of weights for threshold gates," *SIAM J. Disc. Math,* pp. 484-492, 1994.

[32] A. Hajnal, W. Maass, G. Turán, "On the communication complexity of graph properties," 20^{th} *Ann. ACM Symp. on Theory of Computing*, pp. 186-191, 1988.

[33] A. Hajnal, W. Maass, P. Pudlák, M. Szegedy, G. Turán, "Threshold circuits of bounded depth," 28^{th} *Ann. IEEE Symp. Foundations of CS*, pp. 99-110, 1987.

[34] J. Jackson, "An efficient membership query al gorithm for learning DNF with respect to the uniform distribution, " *Proc.* 35^{th} *Ann. IEEE Symp. Foundations of CS,* pp. 42-53, 1994.

[35] V.A. Judin, "Approximation of functions of several variables by trigonometric polynomials," *Matem. Zametki and Math Notes* 20, 1976.

[36] S. Karlin, W.J. Studden, "Tchebycheff systems and applications to analysis and statistics," *Interscience,* 1966.

[37] J. Kahn, J. Kalai, N. Linial, "The influence of variables on Boolean functions," 29^{th} *Ann. IEEE Symp. Foundations of CS,* pp. 68-80, 1988.

[38] M. Karchmer, A. Wigderson, "Monotone circuits for connectivity require super-logarithmic depth," 21^{st} *Ann. ACM Symp. on Theory of Computing*, pp. 539-550, 1988, *SIAM J. Disc. Math* 3, 255-265, 1990.

[39] M. Karchmer, A. Wigderson, "On span programs," 8^{th} *Ann. IEEE conf. on Struct. in Compl. Theory,* pp. 102-111, 1993.

[40] E. Kushilevitz, Y. Mansour, "Learning decision trees using the Fourier transform," 32^{nd} *Ann. IEEE Symp. Foundations of CS,* pp. 455-464 1991.

[41] M. Krause, "Geometric arguments yield better bounds for threshold circuits and distributed computing," 6^{th} *IEEE conf. on Struct. in Compl. Theory,* pp. 314-322, 1991.

[42] M. Krause, P. Pudlák "On the power of depth 2 circuits with threshold and modulo gates," 26^{th} *Ann. Symp. Theory of Comput.*, pp. 48-58, 1994.

[43] M. Krause, P. Pudlák "On Computing Boolean functions by sparse real polynomials," 36^{th} *Ann.IEEE Symp. Foundations of CS*, pp. 682-691, 1995.

[44] M. Krause, S. Waack, "Variation ranks of communication matrices and lower bounds for depth two circuits having symmetric gates and unbounded fan-in," 32^{nd} *Ann. IEEE Symp. Foundations of CS*, pp. 777-782, 1991.

[45] N. Linial, *Personal communication.*

[46] N. Linial, Y. Mansour, N. Nisan, "Constant depth circuits, Fourier transforms, and learnability," 30^{th} *Ann. IEEE Symp. Foundations of CS,* pp. 574-579, 1989.

[47] N. Linial, N. Nisan, "Approximate inclusion-exclusion," 22^{nd} *Ann. ACM Symp. on Theory of Computing,* pp. 260-270, 1990.

[48] Alexis Maciel, "Threshold circuits of small majority depth," *Ph.D thesis, McGill University*, Montreal, Canada, 1995.

[49] Alexis Maciel, "A hierarchy in TC^0," *Manuscript*, 1995.

[50] N. Nisan, A.W. Widgerson, "Hardness vs. randomness," 29^{th} *Ann. IEEE Symp. Foundations of CS,* pp. 2-12, 1988.

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

[52] R. Paturi, "On the degree of polynomials that approximate symmetric Boolean functions," 24^{th} *Ann. ACM Symp. on Theory of Computing,* pp. 468-474, 1992.

[53] P. Petrushev, "Direct and converse theorems for spline approximation and Besov spaces," In *Function spaces and applications*, Cwickel, Peetre, Sagher, Wallin, ed.s, *Springer Verlag Lecture Notes in Mathematics,* 1302, pp. 363-377, 1988.

[54] S.A. Picugov, "On the multidimensional Jackson theorem for linear polynomial approximations," *Math. Notes 24,* 1978.

[55] M.J.D. Powell "Approximation theory and methods," *Cambridge university press,* 1981.

[56] R. Raz, A. Wigderson, "Monotone circuits for matching require linear depth," *Proc.* 22^{nd} *Ann. ACM Symp. Theory of Computing,* pp. 287-292, 1990.

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

[58] A.A. Razborov, "Lower bounds on the size of circuits of bounded depth with basis $\{\wedge, \oplus\}$," *Math. notes of the Aca. of Science of the USSR,* 41 (4), pp. 333-338, 1987.

[59] A.A. Razborov, "One the method of approximation," *Proc.* 21^{st} *Ann. ACM Symp. Theory of Computing,* pp. 167-176, 1989.

[60] K.W. Regan, "Polynomials and other combinatorial definitions of languages," *Manuscript,* 1995.

[61] T.J. Rivlin, "An introduction to the approximation of functions," *Blaisdell Publishing,* 1969.

[62] H.S. Shapiro, "Topics in Approximation theory," *Lecture notes in Mathematics, Springer Verlag,* 1971.

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

[64] M. Sitharam. "Pseudorandom generators and learning algorithms for AC^0," *Ann. ACM Symp. on Theory of Computing,* pp. 478-488, 1994.

[65] M. Sitharam. "Evaluating spectral norms for constant depth circuits with symmetric gates," *J. computational complexity,* 6, pp. 167-189, 1995.

[66] R. Smolensky, "Algebraic methods in the theory of lower bounds for boolean circuit complexity," *Proc.* 19^{th} *Ann. ACM Symp. Theory of Computing,* pp. 77-82, 1987.

[67] R. Smolensky, "On interpolation by analytic functions with special properties and some weak lower bounds on the size of circuits with symmetric gates," complexity," 31^{st} *Ann. IEEE Symp. Foundations of CS,* pp.??-??, 1990.

[68] R. Smolensky, "On representations by low degree polynomials," 34^{th} *Ann. IEEE Symp. Foundations of CS,* pp.??-??, 1993.

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

[70] M. Szegedy "Algebraic methods in computational models with limited communication," *Ph.D thesis U. Chicago,* 1989.

[71] S.-C. Tsai, "Lower bounds on representing Boolean function sas polynomials in $\mathbb{Z}_m$," *Proc.* 8^{th} *Ann. IEEE Conf. on Structure in Complexity Theory,* pp. 96-101, 1993.

[72] V.H. Vu, "On ill-conditioned Boolean matrices and their applications," *Manuscript,* 1996.

[73] J.H. Weaver, "Theory of discrete and continuous Fourier series," *Wiley,* 1989.

[74] A.C. Yao, "Lower bounds by probabilistic arguments," 24^{th} *Ann. IEEE Symp. Foundations of CS,* pp. 420-428, 1983.

[75] A.C. Yao, "Separating the polynomial time hierarchy by oracles," 26^{th} *Ann. IEEE Symp. Foundations of CS,* pp. 1-10, 1985.

[76] A.C. Yao, "Circuits and local computation," 21^{st} *Ann. ACM Symp. on Theory of Computing,* pp. 186-196, 1989.

[77] A.C. Yao, "On ACC and threshold circuits," *Proc.* 31^{st} *Ann. IEEE Symp. Foundations of CS,* pp. 619-627, 1990.

[78] A.C. Yao, "A lower bound for the monotone depth of connectivity," *Proc.* 35^{th} *Ann. IEEE Symp. Foundations of CS,* pp. 302-307, 1994.

Manuscript received March, 1996

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