LOWER BOUNDS FOR AGNOSTIC LEARNING VIA APPROXIMATE RANK

Adam R. Klivans and Alexander A. Sherstov

Abstract. We prove that the concept class of disjunctions cannot be pointwise approximated by linear combinations of any small set of *arbitrary* real-valued functions. That is, suppose that there exist functions $\phi_1, \ldots, \phi_r : \{-1,1\}^n \to \mathbb{R}$ with the property that every disjunction f on n variables has $\|f - \sum_{i=1}^r \alpha_i \phi_i\|_\infty \leqslant 1/3$ for some reals $\alpha_1, \ldots, \alpha_r$. We prove that then $r \geqslant \exp\{\Omega(\sqrt{n})\}$, which is tight. We prove an incomparable lower bound for the concept class of decision lists. For the concept class of majority functions, we obtain a lower bound of $\Omega(2^n/n)$, which almost meets the trivial upper bound of 2^n for *any* concept class. These lower bounds substantially strengthen and generalize the polynomial approximation lower bounds of Paturi (1992) and show that the regression-based agnostic learning algorithm of Kalai *et al.* (2005) is optimal.

Keywords. Agnostic learning, approximate rank, matrix analysis, communication complexity

Subject classification. 03D15, 68Q32, 68Q17.

1. Introduction

Approximating Boolean functions by linear combinations of small sets of features is a fundamental area of study in machine learning. Well-known algorithms such as linear regression, support vector machines, and boosting attempt to learn concepts as linear functions or thresholds over a fixed set of real-valued features. In particular, much work in learning theory has centered around approximating various concept classes, with respect to a variety of distributions and metrics, by *low-degree polynomials* (Bshouty & Tamon 1996; Jackson 1995; Klivans *et al.* 2004; Klivans & Servedio 2004; Kushilevitz & Mansour 1993; Linial *et al.* 1993; Mansour 1995; O'Donnell & Servedio 2008). In this case, the features mentioned above are simply monomials. For example, Linial *et al.* (1993) gave a celebrated uniform-distribution algorithm for

learning constant-depth circuits by proving that any such circuit can be approximated by a low-degree Fourier polynomial, with respect to the uniform distribution and ℓ_2 norm.

A more recent application of the polynomial paradigm is due to Kalai *et al.* (2008), who considered the well-studied problem of agnostically learning disjunctions (Decatur 1993; Kearns & Li 1993; Mansour & Parnas 1996; Valiant 1985). Kalai *et al.* recalled that a disjunction on n variables can be approximated pointwise by a degree-$O(\sqrt{n})$ polynomial (Nisan & Szegedy 1994; Paturi 1992). They then used linear regression to obtain the first subexponential ($2^{\tilde{O}(\sqrt{n})}$-time) algorithm for *agnostically* learning disjunctions with respect to *any* distribution (Kalai *et al.* 2008, Thm. 2). More generally, Kalai *et al.* used ℓ_∞-norm approximation to give subexponential-time algorithms for distribution-free agnostic learning.

Before stating our results formally, we briefly describe our notation. A Boolean function is a mapping $f : \{-1, 1\}^n \to \{-1, 1\}$, where -1 corresponds to "true." A *feature* is any function $\phi : \{-1, 1\}^n \to \mathbb{R}$. We say that ϕ *approximates* f *pointwise within* ϵ, denoted

$$\|f - \phi\|_\infty \leqslant \epsilon,$$

if $|f(x) - \phi(x)| \leqslant \epsilon$ for all x. We say that *a linear combination of features* $\phi_1, \ldots, \phi_r$ *approximates* f *pointwise within* ϵ if $\|f - \sum_{i=1}^r \alpha_i \phi_i\|_\infty \leqslant \epsilon$ for some reals $\alpha_1, \ldots, \alpha_r$.

Our results. Let $\mathcal{C}$ be a concept class. Suppose that $\phi_1, \ldots, \phi_r$ are features whose linear combinations can pointwise approximate every function in $\mathcal{C}$. We first observe that the algorithm of Kalai *et al.*—assuming that $\phi_1, \ldots, \phi_r$ can be evaluated efficiently—learns $\mathcal{C}$ agnostically under any distribution in time polynomial in r and n.

To put our lower bounds in context, we note that current methods for agnostically learning a concept class $\mathcal{C}$ involve solving an empirical risk minimization problem using polynomials. That is, all algorithms for agnostic learning that we are aware of work by finding the best fitting polynomial (with respect to some metric) to a training set of labeled examples and taking a threshold. Kalai *et al.* (2008) proved that if polynomials can pointwise approximate the concept class, this method is guaranteed to solve the empirical risk minimization problem (and hence the agnostic learning problem) for $\mathcal{C}$. We will give scenarios where linear combinations of *any* small number of features fail to approximate an unknown concept, thus giving us no guarantee that we are solving the

empirical risk minimization problem. We believe that these scenarios demonstrate the limits of the polynomial-minimization approach for distribution-free agnostic learning.

We begin with the concept class of disjunctions:

THEOREM 1.1 (Disjunctions). *Let $\mathcal{C} = \{\bigvee_{i \in S} x_i : S \subseteq [n]\}$ be the concept class of disjunctions. Let $\phi_1, \ldots, \phi_r : \{-1,1\}^n \to \mathbb{R}$ be arbitrary functions whose linear combinations can pointwise approximate every $f \in \mathcal{C}$ within $\epsilon = 1/3$. Then $r \geqslant 2^{\Omega(\sqrt{n})}$.*

Theorem 1.1 shows the optimality of using monomials as features for approximating disjunctions. In particular, it rules out the possibility of using the algorithm of Kalai *et al.* with other, cleverly constructed features to obtain an improved agnostic learning result for disjunctions. The same result of course holds for the concept class of *conjunctions*.

We obtain an incomparable result against decision lists (and hence linear-size DNF formulas).

THEOREM 1.2 (Decision lists). *Let $\mathcal{C}$ be the concept class of decision lists. Let $\phi_1, \ldots, \phi_r : \{-1,1\}^n \to \mathbb{R}$ be arbitrary functions whose linear combinations can pointwise approximate every $f \in \mathcal{C}$ within $\epsilon = 1 - 2^{-cn^{1/3}}$, where $c > 0$ is a sufficiently small absolute constant. Then $r \geqslant 2^{\Omega(n^{1/3})}$.*

Theorems 1.1 and 1.2 both give exponential lower bounds on r. Comparing the two, we see that Theorem 1.1 gives a better bound on r against a simpler concept class. On the other hand, Theorem 1.2 remains valid for a particularly weak success criterion: when the approximation quality is exponentially close to trivial ($\epsilon = 1$).

The last concept class that we study is that of majority functions. Here we prove our best lower bound, $r = \Omega(2^n/n)$, that essentially meets the trivial upper bound of 2^n for any concept class. Put differently, we show that the concept class of majorities is essentially as hard to approximate as *any* concept class at all. In particular, this shows that the polynomial-minimization paradigm cannot yield any nontrivial ($2^{o(n)}$-time) distribution-free algorithm for agnostically learning majority functions.

THEOREM 1.3 (Majority functions). *Let $\mathcal{C} = \{\mathrm{MAJ}_n(\pm x_1, \ldots, \pm x_n)\}$ be the concept class of majority functions. Let $\phi_1, \ldots, \phi_r : \{-1,1\}^n \to \mathbb{R}$ be arbitrary functions whose linear combinations can pointwise approximate every $f \in \mathcal{C}$ within $\epsilon = c/\sqrt{n}$, where c is a sufficiently small absolute constant. Then $r \geqslant \Omega(2^n/n)$. For approximation to within $\epsilon = 1/3$, we obtain $r \geqslant 2^{\Omega(n/\log n)}$.*

We also relate our inapproximability results to the notions of *dimension complexity* and *statistical query dimension* (Sections 5–7). Among other things, we show that the types of approximation lower bounds that we study are prerequisites for lower bounds on dimension complexity and the SQ dimension.

Additional applications. The preceding discussion has emphasized the implications of Theorems 1.1–1.3 in learning theory. Our results also have consequences in approximation theory. In a classic result, Paturi (1992) constructed polynomials of degree $\Theta(\sqrt{n})$ and $\Theta(n)$ that pointwise approximate disjunctions and majority functions, respectively. He also showed that these *degree* results are optimal for polynomials. This, of course, does not exclude polynomials that are *sparse*, i.e., contain few monomials. Our lower bounds strengthen Paturi's result by showing that the approximating polynomials cannot be sparse. In addition, our analysis remains valid when monomials are replaced by *arbitrary* features. As anticipated, our techniques differ significantly from Paturi's.

It is also useful to examine our work from the standpoint of matrix analysis. As will become apparent in later sections, the quantity of interest to us is the *ϵ-approximate rank* of a Boolean matrix M. This quantity is defined as the least rank of a real matrix A that differs from M by at most ϵ in any entry: $\|M - A\|_\infty \leqslant \epsilon$. Apart from being a natural matrix-analytic notion with applications to learning theory, ϵ-approximate rank arises in quantum communication complexity (Buhrman & Wolf 2001). While ϵ-approximate rank remains difficult to analyze in general, our paper shows several techniques that prove to be successful in concrete cases.

Our techniques. We obtain our main theorems in two steps. First, we show how to place a lower bound on the quantity of interest (the size of feature sets that pointwise approximate a concept class $\mathcal{C}$) using the *discrepancy* and the *ϵ-approximate trace norm* of the characteristic matrix of $\mathcal{C}$. The latter two quantities have been extensively studied. In particular, the discrepancy estimate that we need is a recent result of Buhrman *et al.* (2007b). For estimates of the ϵ-approximate trace norm, we turn to the pioneering work of Razborov (2003) on quantum communication complexity, as well as classical results on matrix perturbation and Fourier analysis.

2. Preliminaries

The notation $[n]$ stands for the set $\{1, 2, \ldots, n\}$, and $\binom{[n]}{k}$ stands for the family of all k-element subsets of $[n] = \{1, 2, \ldots, n\}$. The symbol $\mathbb{R}^{n \times m}$ refers to the family of all $n \times m$ matrices with real entries. The (i, j)th entry of a matrix

A is denoted by A_{ij} or $A(i,j)$. We frequently use "generic-entry" notation to specify a matrix succinctly: we write $A = [F(i,j)]_{i,j}$ to mean that the (i,j)th entry of A is given by the expression $F(i,j)$.

A *concept class* $\mathcal{C}$ is any set of Boolean functions $f : \{-1,1\}^n \to \{-1,1\}$. The *characteristic matrix* of $\mathcal{C}$ is the matrix $M = [f(x)]_{f\in\mathcal{C},\, x\in\{-1,1\}^n}$. In words, the rows of M are indexed by functions $f \in \mathcal{C}$, the columns are indexed by inputs $x \in \{-1,1\}^n$, and the entries are given by $M_{f,x} = f(x)$.

A *decision list* is a Boolean function $f\colon \{-1,1\}^n \to \{-1,1\}$ specified by a fixed permutation $\sigma\colon [n] \to [n]$, a fixed vector $a \in \{-1,1\}^{n+1}$, and a fixed vector $b \in \{-1,1\}^n$. The computation of f on input $x \in \{-1,1\}^n$ proceeds as follows. If $x_{\sigma(i)} \neq b_i$ all $i = 1,2,\ldots,n$, then one outputs a_{n+1}. Otherwise, one outputs a_i, where $i \in \{1,2,\ldots,n\}$ is the least integer with $x_{\sigma(i)} = b_i$.

2.1. Agnostic learning. The agnostic learning model was defined by Kearns *et al.* (1994). It gives the learner access to arbitrary example-label pairs with the requirement that the learner output a hypothesis competitive with the best hypothesis from some fixed concept class. Specifically, let D be a distribution on $\{-1,1\}^n \times \{-1,1\}$ and let $\mathcal{C}$ be a concept class. For a Boolean function f, define its *error* as $\mathrm{err}(f) = \mathbf{P}_{(x,y)\sim D}[f(x) \neq y]$. Define the *optimal error* of $\mathcal{C}$ as $\mathsf{opt} = \min_{f\in\mathcal{C}} \mathrm{err}(f)$.

A concept class $\mathcal{C}$ is *agnostically learnable* if there exists an algorithm which takes as input δ, ϵ, and access to an example oracle $\mathsf{EX}(D)$, and outputs with probability at least $1-\delta$ a hypothesis $h : \{-1,1\}^n \to \{-1,1\}$ such that $\mathrm{err}(h) \leqslant \mathsf{opt} + \epsilon$. We say $\mathcal{C}$ is *agnostically learnable in time t* if the running time, including calls to the example oracle, is bounded by $t(\epsilon, \delta, n)$.

The following proposition relates pointwise approximation by linear combinations of features to efficient agnostic learning. It is a straightforward generalization of the ℓ_1 polynomial-regression algorithm of Kalai *et al.* (2008).

PROPOSITION 2.1. *Fix a constant $\epsilon \in (0,1)$ and a concept class $\mathcal{C}$. Assume that there are functions $\phi_1, \ldots, \phi_r : \{-1,1\}^n \to \mathbb{R}$ whose linear combinations can pointwise approximate every $f \in \mathcal{C}$ within ϵ. Assume further that each $\phi_i(x)$ is computable in polynomial time. Then $\mathcal{C}$ is agnostically learnable to accuracy ϵ in time polynomial in r and n.*

PROOF. Let $\mathcal{C}'$ stand for the family of functions $f\colon \{-1,1\}^n \to \{-1,1\}$ representable as $f(x) = \mathrm{sgn}(a_1\phi_1(x) + \ldots + a_r\phi_r(x) - a_{r+1})$ for some reals $a_1, \ldots, a_{r+1}$. Since halfspaces in n dimensions have VC dimension $n+1$, the VC dimension of $\mathcal{C}'$ is at most $r+1$. For m labeled examples $(x^1, y^1), \ldots, (x^m, y^m)$ drawn independently from distribution D, one can minimize the quantity

$\frac{1}{m}\sum_{j=1}^{m}|\sum_{i=1}^{r}a_i\phi_i(x^j)-y^j|$ over the reals $a_1,\ldots,a_r$ in polynomial time (in r and n) using an efficient algorithm for linear programming. Since linear combinations of $\phi_1,\ldots,\phi_r$ can pointwise approximate every $f\in\mathcal{C}$ within ϵ, we have that for every $f\in\mathcal{C}$, there exist $a_1,\ldots,a_r$ such that $\mathbf{E}_{x\sim D}[(a_1\phi_1(x)+\ldots+a_r\phi_r(x)-f(x))^2]\leqslant\epsilon^2$. Applying Theorem 5 of Kalai *et al.* (2008) finishes the proof. □

2.2. Fourier transform. Consider the vector space of functions $\{-1,1\}^n\to\mathbb{R}$, equipped with the inner product $\langle f,g\rangle=2^{-n}\sum_{x\in\{-1,1\}^n}f(x)g(x)$. The parity functions $\chi_S(x)=\prod_{i\in S}x_i$, where $S\subseteq[n]$, form an orthonormal basis for this inner product space. As a result, every Boolean function f can be uniquely written as

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

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

$$\|\hat{f}\|_1=\sum_{S\subseteq[n]}|\hat{f}(S)|.$$

2.3. Matrix analysis. We draw freely on basic notions from matrix analysis; a standard reference on the subject is Golub & Loan (1996). This section only reviews the notation and the more substantial results.

Let $A\in\mathbb{R}^{m\times n}$. We let $\|A\|_\infty=\max_{ij}|A_{ij}|$, the largest absolute value of an entry of A. We denote the singular values of A by $\sigma_1(A)\geqslant\sigma_2(A)\geqslant\ldots\geqslant\sigma_{\min\{m,n\}}(A)\geqslant 0$. Recall that $\|A\|_\Sigma=\sum_{i=1}^{\min\{m,n\}}\sigma_i(A)$ and $\|A\|_F=(\sum_{i=1}^{m}\sum_{j=1}^{n}A_{ij}^2)^{1/2}$ are the trace norm and Frobenius norm of A. We will also need the *ϵ-approximate trace norm,* defined as

$$\|A\|_\Sigma^\epsilon=\min\{\|B\|_\Sigma:\|A-B\|_\infty\leqslant\epsilon\}.$$

Our analysis requires the Hoffman-Wielandt inequality (see Golub & Loan 1996, Theorem 8.6.4). In words, it states that small perturbations to the entries of a matrix result in small perturbations to its singular values.

THEOREM 2.2 (Hoffman-Wielandt inequality). *Fix matrices $A,B\in\mathbb{R}^{m\times n}$. Then*

$$\sum_{i=1}^{\min\{m,n\}}(\sigma_i(A)-\sigma_i(B))^2\leqslant\|A-B\|_F^2.$$

In particular, if $\operatorname{rank}(B) = k$ *then*

$$\sum_{i \geqslant k+1} \sigma_i(A)^2 \leqslant \|A - B\|_F^2.$$

The Hoffman-Wielandt inequality is used in the following lemma, which allows us to easily construct matrices with high approximate trace norm.

LEMMA 2.3. *Let* $M = [f(x \oplus y)]_{x,y}$, *where* $f\colon \{0,1\}^n \to \{-1,1\}$ *is a given function and the indices* x, y *range over* $\{0,1\}^n$. *Then for all* $\epsilon \geqslant 0$,

$$\|M\|_\Sigma^\epsilon \geqslant 2^n(\|\hat{f}\|_1 - \epsilon 2^{n/2}).$$

PROOF. Let $N = 2^n$ be the order of M. Fix a matrix A with $\|A - M\|_\infty \leqslant \epsilon$. By the Hoffman-Wielandt inequality,

$$N^2\epsilon^2 \geqslant \|A - M\|_F^2 \geqslant \sum_{i=1}^{N} (\sigma_i(A) - \sigma_i(M))^2 \geqslant \frac{1}{N}(\|A\|_\Sigma - \|M\|_\Sigma)^2,$$

so that $\|A\|_\Sigma \geqslant \|M\|_\Sigma - N^{3/2}\epsilon$. Since the choice of A was arbitrary, we conclude that

$$\|M\|_\Sigma^\epsilon \geqslant \|M\|_\Sigma - N^{3/2}\epsilon. \tag{2.4}$$

It is well-known (Linial *et al.* 2007, p. 458) that the singular values of M/N are precisely the absolute values of the Fourier coefficients of f. Indeed,

$$M = Q \begin{bmatrix} N\hat{f}(\varnothing) & & \\ & \ddots & \\ & & N\hat{f}([n]) \end{bmatrix} Q^{\mathsf{T}},$$

where $Q = N^{-1/2}[\chi_S(x)]_{x,S}$ is an orthogonal matrix. In particular, $\|M\|_\Sigma = N\|\hat{f}\|_1$. Together with (2.4), this completes the proof. □

A *sign matrix* is any matrix with ± 1 entries.

2.4. Communication complexity. We consider functions $f : X \times Y \to \{-1,1\}$. Typically $X = Y = \{-1,1\}^n$, but we also allow X and Y to be arbitrary sets, possibly of unequal cardinality. A *rectangle* of $X \times Y$ is any set

$R = A \times B$ with $A \subseteq X$ and $B \subseteq Y$. For a fixed distribution μ over $X \times Y$, the *discrepancy* of f is defined as

$$\mathrm{disc}_\mu(f) = \max_R \left| \sum_{(x,y) \in R} \mu(x,y) f(x,y) \right|,$$

where the maximum is taken over all rectangles R. We define $\mathrm{disc}(f) = \min_\mu \{\mathrm{disc}_\mu(f)\}$. We identify the function f with its *communication matrix* $M = [f(x,y)]_{x,y}$ and define $\mathrm{disc}_\mu(M) = \mathrm{disc}_\mu(f)$. A definitive resource for further details on communication complexity is the book of Kushilevitz & Nisan (1997).

2.5. Statistical query dimension. The statistical query (SQ) model of learning, due to Kearns (1998), is a restriction of Valiant's PAC model. See Kearns & Vazirani (1994) for a comprehensive treatment. The *SQ dimension* of $\mathcal{C}$ under μ, denoted $\mathrm{sqdim}_\mu(\mathcal{C})$, is the largest d for which there are d functions $f_1, \ldots, f_d \in \mathcal{C}$ with

$$\left| \mathop{\mathbf{E}}_{x \sim \mu} [f_i(x) f_j(x)] \right| \leqslant \frac{1}{d}$$

for all $i \neq j$. We denote

$$\mathrm{sqdim}(\mathcal{C}) = \max_\mu \{\mathrm{sqdim}_\mu(\mathcal{C})\}.$$

The SQ dimension is a tight measure (Blum *et al.* 1994) of the learning complexity of a given concept class $\mathcal{C}$ in the SQ model.

3. Approximate rank: definition and properties

For a real matrix A, its *ϵ-approximate rank* is defined as

$$\mathrm{rank}_\epsilon(A) = \min_B \{\mathrm{rank}(B) : B \text{ real}, \|A - B\|_\infty \leqslant \epsilon\}.$$

This notion is a natural one and has been studied before. In particular, Buhrman & Wolf (2001) show that the approximate rank of a sign matrix implies lower bounds on its quantum communication complexity (in the bounded-error model without prior entanglement). In Section 6, we survey two other related concepts: matrix rigidity and dimension complexity.

We define the ϵ-*approximate rank* of a concept class $\mathcal{C}$ as

$$\mathrm{rank}_\epsilon(\mathcal{C}) = \mathrm{rank}_\epsilon(M),$$

where M is the characteristic matrix of $\mathcal{C}$. For example, $\mathrm{rank}_0(\mathcal{C}) = \mathrm{rank}(M)$ and $\mathrm{rank}_1(\mathcal{C}) = 0$. It is thus the behavior of $\mathrm{rank}_\epsilon(\mathcal{C})$ for intermediate values of ϵ that is of primary interest. The following proposition follows trivially from our definitions.

PROPOSITION 3.1 (Approximate rank reinterpreted). *Let $\mathcal{C}$ be a concept class. Then $\mathrm{rank}_\epsilon(\mathcal{C})$ is the smallest integer r such that there exist real functions $\phi_1, \ldots, \phi_r : \{-1,1\}^n \to \mathbb{R}$ with the property that each $f \in \mathcal{C}$ has $\|f - \sum_{i=1}^r \alpha_i \phi_i\|_\infty \leqslant \epsilon$ for some reals $\alpha_1, \ldots, \alpha_r$.*

3.1. Improving the quality of the approximation. We now take a closer look at $\mathrm{rank}_\epsilon(M)$ as a function of ϵ. Suppose that we have an estimate of $\mathrm{rank}_E(M)$ for some $0 < E < 1$. Can we use this information to obtain a nontrivial upper bound on $\mathrm{rank}_\epsilon(M)$, where $0 < \epsilon < E$? It turns out that we can. We first recall that the sign function can be approximated well by a real polynomial:

FACT 3.2 (Buhrman *et al.* 2007a). *Let E be given, $0 < E < 1$. Then for each integer $d \geqslant 1$, there exists a degree-d real univariate polynomial p such that*

$$|p(t) - \operatorname{sgn} t| \leqslant 2 \exp\left\{-\frac{d}{2}\left(\frac{1-E}{1+E}\right)^2\right\} \qquad (1 - E \leqslant |t| \leqslant 1 + E).$$

PROOF (adapted from Buhrman *et al.* 2007a). Consider the univariate polynomial

$$q(t) = \sum_{i=\lceil d/2 \rceil}^{d} \binom{d}{i} t^i (1-t)^{d-i}.$$

By definition, $q(t)$ is the probability of observing at least $d/2$ heads in a sequence of d independent coin flips, each coming up heads with probability t. For $0 \leqslant \gamma \leqslant 1/2$, the Chernoff bound (Chernoff 1952) implies that q sends $[0, \frac{1}{2} - \gamma] \to [0, \mathrm{e}^{-2d\gamma^2}]$ and $[\frac{1}{2} + \gamma, 1] \to [1 - \mathrm{e}^{-2d\gamma^2}, 1]$. Letting

$$\gamma = \frac{1-E}{2(1+E)}, \qquad p(t) = 2q\left(\frac{1}{2} + \frac{1}{2(1+E)}\, t\right) - 1,$$

we see that p has the desired behavior. □

THEOREM 3.3. *Let M be a sign matrix, and let $0 < \epsilon < E < 1$. Then*

$$\operatorname{rank}_\epsilon(M) \leqslant \operatorname{rank}_E(M)^d,$$

where d is any positive integer with $2\exp\left\{-\frac{d}{2}\left(\frac{1-E}{1+E}\right)^2\right\} \leqslant \epsilon$.

PROOF. Let d be as stated. By Fact 3.2, there is a degree-d polynomial $p(t)$ with

$$|p(t) - \operatorname{sgn} t| \leqslant \epsilon \qquad (1 - E \leqslant |t| \leqslant 1 + E).$$

Let A be a real matrix with $\|A - M\|_\infty \leqslant E$ and $\operatorname{rank}(A) = \operatorname{rank}_E(M)$. Then the matrix $B = [p(A_{ij})]_{i,j}$ approximates M to the desired accuracy: $\|B - M\|_\infty \leqslant \epsilon$. Since p is a polynomial of degree d, elementary linear algebra shows that $\operatorname{rank}(B) \leqslant \operatorname{rank}(A)^d$. □

Note. The key idea in the proof of Theorem 3.3 is to improve the quality of the approximating matrix by applying a suitable polynomial to its entries. This idea is not new. For example, Alon (2003) uses the same method in the simpler setting of *one-sided* errors.

We will mainly need the following immediate consequences of Theorem 3.3.

COROLLARY 3.4. *Let M be a sign matrix. Let ϵ, E be constants with $0 < \epsilon < E < 1$. Then*

$$\operatorname{rank}_\epsilon(M) \leqslant \operatorname{rank}_E(M)^c,$$

where $c = c(\epsilon, E)$ is a constant.

COROLLARY 3.5. *Let M be a sign matrix. Let ϵ be a constant with $0 < \epsilon < 1$. Then*

$$\operatorname{rank}_{1/n^c}(M) \leqslant \operatorname{rank}_\epsilon(M)^{O(\log n)}$$

for every constant $c > 0$.

By Corollary 3.4, the choice of the constant ϵ affects $\operatorname{rank}_\epsilon(M)$ by at most a polynomial factor. When such factors are unimportant, we will adopt $\epsilon = 1/3$ as a canonical setting.

3.2. Estimating the approximate rank. We will use two methods to estimate the approximate rank. The first uses the ϵ-approximate trace norm of the same matrix, and the second uses its discrepancy.

LEMMA 3.6 (Lower bound via approximate trace norm). *Fix a matrix* $M \in \{-1,1\}^{N\times N}$. *Then*

$$\operatorname{rank}_\epsilon(M) \geqslant \left(\frac{\|M\|_\Sigma^\epsilon}{(1+\epsilon)N}\right)^2.$$

PROOF. Let A be an arbitrary matrix with $\|M-A\|_\infty \leqslant \epsilon$. We have:

$$\begin{aligned}(\|M\|_\Sigma^\epsilon)^2 \leqslant (\|A\|_\Sigma)^2 = \left(\sum_{i=1}^{\operatorname{rank}(A)} \sigma_i(A)\right)^2 &\leqslant \left(\sum_{i=1}^{\operatorname{rank}(A)} \sigma_i(A)^2\right)\operatorname{rank}(A)\\ &= (\|A\|_F)^2 \operatorname{rank}(A) \leqslant (1+\epsilon)^2 N^2 \operatorname{rank}(A),\end{aligned}$$

as claimed. □

Our second method is as follows.

LEMMA 3.7 (Lower bound via discrepancy). *Let* M *be a sign matrix and* $0 \leqslant \epsilon < 1$. *Then*

$$\operatorname{rank}_\epsilon(M) \geqslant \frac{1-\epsilon}{1+\epsilon}\cdot\frac{1}{64\operatorname{disc}(M)^2}.$$

The proof of Lemma 3.7 requires several definitions and facts that we do not use elsewhere in this paper. For this reason, we defer it to Appendix A.

4. Approximate rank of specific concept classes

We proceed to prove our main results (Theorems 1.1–1.3), restated here as Theorems 4.2, 4.6, and 4.8.

4.1. Disjunctions. We recall a breakthrough result of Razborov (2003) on the quantum communication complexity of disjointness. The crux of that work is the following theorem.

THEOREM 4.1 (Razborov 2003, Sec. 5.3). *Let n be an integer multiple of 4. Let M be the $\binom{n}{n/4} \times \binom{n}{n/4}$ matrix whose rows and columns are indexed by sets in $\binom{[n]}{n/4}$ and entries given by*

$$M_{S,T} = \begin{cases} 1 & \text{if } S \cap T = \varnothing, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\|M\|_{\Sigma}^{1/4} = 2^{\Omega(\sqrt{n})} \binom{n}{n/4}.$$

We can now prove an exponential lower bound on the approximate rank of disjunctions, a particularly simple concept class.

THEOREM 4.2 (Approximate rank of disjunctions). *Let $\mathcal{C} = \{\bigvee_{i \in S} x_i : S \subseteq [n]\}$ be the concept class of disjunctions. Then*

$$\operatorname{rank}_{1/3}(\mathcal{C}) = 2^{\Omega(\sqrt{n})}.$$

PROOF. Without loss of generality, we may assume that n is a multiple of 4. One easily verifies that the characteristic matrix of $\mathcal{C}$ is $M_{\mathcal{C}} = [\bigvee_{i=1}^{n}(x_i \wedge y_i)]_{x,y}$. We can equivalently view $M_{\mathcal{C}}$ as the $2^n \times 2^n$ sign matrix whose rows and columns indexed by sets in $[n]$ and entries given by:

$$M_{\mathcal{C}}(S,T) = \begin{cases} 1 & \text{if } S \cap T = \varnothing, \\ -1 & \text{otherwise.} \end{cases}$$

Now let A be a real matrix with $\|M_{\mathcal{C}} - A\|_{\infty} \leqslant 1/3$. Let $Z_{\mathcal{C}} = \frac{1}{2}(M_{\mathcal{C}} + J)$, where J is the all-ones matrix. We immediately have $\|Z_{\mathcal{C}} - \frac{1}{2}(A + J)\|_{\infty} \leqslant 1/6$, and thus

$$\operatorname{rank}_{1/6}(Z_{\mathcal{C}}) \leqslant \operatorname{rank}\left(\tfrac{1}{2}(A + J)\right) \leqslant \operatorname{rank}(A) + 1. \tag{4.3}$$

However, $Z_{\mathcal{C}}$ contains as a submatrix the matrix M from Theorem 4.1. Therefore,

$$\begin{aligned} \operatorname{rank}_{1/6}(Z_{\mathcal{C}}) &\geqslant \operatorname{rank}_{1/6}(M) \\ &\geqslant \left(\frac{\|M\|_{\Sigma}^{1/4}}{(1 + 1/4)\binom{n}{n/4}} \right)^2 && \text{by Lemma 3.6} \\ &\geqslant 2^{\Omega(\sqrt{n})} && \text{by Theorem 4.1.} \end{aligned} \tag{4.4}$$

The theorem follows immediately from (4.3) and (4.4). □

4.2. Decision lists. We recall a recent result due to Buhrman *et al.* (2007b):

THEOREM 4.5 (Buhrman *et al.* 2007b, Sec. 3). *There is a Boolean function* $f : \{-1,1\}^n \times \{-1,1\}^n \to \{-1,1\}$ *computable by an* AC^0 *circuit of depth* 3 *such that*

$$\operatorname{disc}(f) = 2^{-\Omega(n^{1/3})}.$$

Moreover, for each fixed y, *the function* $f_y(x) = f(x,y)$ *is a decision list.*

We can now analyze the approximate rank of decision lists.

THEOREM 4.6 (Approximate rank of decision lists). *Let* $\mathcal{C}$ *denote the concept class of functions* $f : \{-1,1\}^n \to \{-1,1\}$ *computable by decision lists. Then*

$$\operatorname{rank}_\epsilon(\mathcal{C}) = 2^{\Omega(n^{1/3})}$$

for $0 \leqslant \epsilon \leqslant 1 - 2^{-cn^{1/3}}$, *where* $c > 0$ *is a sufficiently small absolute constant.*

PROOF. Let M be the characteristic matrix of $\mathcal{C}$, and let $f(x,y)$ be the function from Theorem 4.5. Since $[f(x,y)]_{y,x}$ is a submatrix of M, we have $\operatorname{rank}_\epsilon(M) \geqslant \operatorname{rank}_\epsilon([f(x,y)]_{y,x})$. The claim now follows from Lemma 3.7. □

Comparing the results of Theorems 4.2 and 4.6 for small constant ϵ, we see that Theorem 4.2 is stronger in that it gives a better lower bound against a simpler concept class. On the other hand, Theorem 4.6 is stronger in that it remains valid for the broad range $0 \leqslant \epsilon \leqslant 1 - 2^{-\Theta(n^{1/3})}$, whereas the ϵ-approximate rank in Theorem 4.2 is easily seen to be at most n for all $\epsilon \geqslant 1 - \frac{1}{2n}$.

4.3. Majority functions. As a final application, we consider the concept class of majority functions. Here we prove a lower bound of $\Omega(2^n/n)$ on the approximate rank, which is the best of our three constructions.

We start by analyzing the ℓ_1 norm of the Fourier spectrum of the majority function.

THEOREM 4.7. *The majority function* $\mathrm{MAJ}_n \colon \{-1,1\}^n \to \{-1,1\}$ *satisfies*

$$\|\widehat{\mathrm{MAJ}_n}\|_1 = \Theta\left(\frac{2^{n/2}}{\sqrt{n}}\right).$$

The tight estimate in Theorem 4.7 is an improvement on an earlier lower bound of $\Omega(2^{n/2}/n)$ due to Linial *et al.* (2007).

PROOF (of Theorem 4.7). Since $\|\widehat{\mathrm{MAJ}_n}\|_1 \geqslant \|\widehat{\mathrm{MAJ}_{n-1}}\|_1$, we may assume without loss of generality that n is odd. Bernasconi (1998) showed that for an odd integer $n = 2m+1$, the even-order Fourier coefficients of MAJ_n are zero, whereas the Fourier coefficients of MAJ_n of odd order $2i+1$ have absolute value

$$4^{-m}\binom{2i}{i}\binom{2m-2i}{m-i}\binom{m}{i}^{-1}.$$

Summing over all Fourier coefficients of odd order, we obtain

$$\begin{aligned}
\|\widehat{\mathrm{MAJ}_n}\|_1 &= 4^{-m}\sum_{i=0}^{m}\binom{2i}{i}\binom{2m-2i}{m-i}\binom{m}{i}^{-1}\binom{2m+1}{2i+1} \\
&= 4^{-m}\binom{2m}{m}\sum_{i=0}^{m}\frac{2m+1}{2i+1}\binom{m}{i} \\
&= \Theta\left(\frac{2^{n/2}}{\sqrt{n}}\right),
\end{aligned}$$

as claimed. □

THEOREM 4.8 (Approximate rank of majority functions). *Let $\mathcal{C}$ denote the concept class of majority functions, $\mathcal{C} = \{\mathrm{MAJ}_n(\pm x_1, \ldots, \pm x_n)\}$. Then*

$$\mathrm{rank}_{c/\sqrt{n}}(\mathcal{C}) \geqslant \Omega(2^n/n)$$

for a sufficiently small absolute constant $c > 0$. Also,

$$\mathrm{rank}_{1/3}(\mathcal{C}) = 2^{\Omega(n/\log n)}.$$

PROOF. The characteristic matrix of $\mathcal{C}$ is $M = [\mathrm{MAJ}_n(x \oplus y)]_{x,y}$. Taking $\epsilon = c/\sqrt{n}$ for a suitably small constant $c > 0$, we obtain:

$$\begin{aligned}
\mathrm{rank}_{c/\sqrt{n}}(M) &\geqslant \left(\frac{\|M\|_\Sigma^{c/\sqrt{n}}}{(1+c/\sqrt{n})2^n}\right)^2 && \text{by Lemma 3.6} \\
&\geqslant \frac{1}{4}\left(\|\widehat{\mathrm{MAJ}_n}\|_1 - \frac{c2^{n/2}}{\sqrt{n}}\right)^2 && \text{by Lemma 2.3} \\
&\geqslant \Omega\left(\frac{2^n}{n}\right) && \text{by Theorem 4.7.}
\end{aligned}$$

Finally,

$$\mathrm{rank}_{1/3}(\mathcal{C}) \geqslant [\mathrm{rank}_{c/\sqrt{n}}(\mathcal{C})]^{1/O(\log n)} \geqslant 2^{\Omega(n/\log n)}$$

by Corollary 3.5. □

5. Approximate rank versus SQ dimension

This section relates the approximate rank of a concept class $\mathcal{C}$ to its SQ dimension, a fundamental quantity in learning theory. In short, we prove that the SQ dimension is a lower bound on the approximate rank, and that the gap between the two quantities can be exponential. A starting point in our analysis is the relationship between the SQ dimension of $\mathcal{C}$ and ℓ_2-norm approximation of $\mathcal{C}$, which might be of some independent interest.

THEOREM 5.1 (SQ dimension and ℓ_2 approximation). *Let $\mathcal{C}$ be a concept class, and let μ be a distribution over $\{-1,1\}^n$. Suppose that there exist functions $\phi_1, \ldots, \phi_r : \{-1,1\}^n \to \mathbb{R}$ such that each $f \in \mathcal{C}$ has*

$$\mathop{\mathbf{E}}_{x\sim\mu}\left[\left(f(x) - \sum_{i=1}^{r} \alpha_i \phi_i(x)\right)^2\right] \leqslant \epsilon$$

for some reals $\alpha_1, \ldots, \alpha_r$. Then

$$r \geqslant (1-\epsilon)d - \sqrt{d},$$

where $d = \mathrm{sqdim}_{\mu}(\mathcal{C})$.

PROOF. By definition of the SQ dimension, there exist $f_1, \ldots, f_d \in \mathcal{C}$ with $|\mathbf{E}_{\mu}[f_i \cdot f_j]| \leqslant 1/d$ for all $i \neq j$. For simplicity, assume that μ is a distribution with rational weights (extension to the general case is straightforward). Then there is an integer $k \geqslant 1$ such that each $\mu(x)$ is an integral multiple of $1/k$. Construct the $d \times k$ sign matrix

$$M = [f_i(x)]_{i,x}\,,$$

whose rows are indexed by the functions $f_1, \ldots, f_d$ and whose columns are indexed by inputs $x \in \{-1,1\}^n$ (a given input x indexes exactly $k\mu(x)$ columns). It is easy to verify that $MM^{\mathsf{T}} = [k\mathbf{E}_{\mu}[f_i \cdot f_j]]_{i,j}$, and thus

$$\|MM^{\mathsf{T}} - k \cdot I\|_F < k. \tag{5.2}$$

The existence of $\phi_1, \ldots, \phi_r$ implies the existence of a rank-r real matrix A with $\|M - A\|_F^2 \leqslant \epsilon kd$. On the other hand, the Hoffman-Wielandt inequality (Theorem 2.2) guarantees that $\|M - A\|_F^2 \geqslant \sum_{i=r+1}^{d} \sigma_i(M)^2$. Combining these two inequalities yields:

$$\begin{aligned}
\epsilon kd &\geqslant \sum_{i=r+1}^{d} \sigma_i(M)^2 = \sum_{i=r+1}^{d} \sigma_i(MM^{\mathsf{T}}) \\
&\geqslant k(d-r) - \sum_{i=r+1}^{d} |\sigma_i(MM^{\mathsf{T}}) - k| \\
&\geqslant k(d-r) - \sqrt{\sum_{i=r+1}^{d} (\sigma_i(MM^{\mathsf{T}}) - k)^2}\ \sqrt{d-r} && \text{by Cauchy-Schwarz} \\
&\geqslant k(d-r) - \|MM^{\mathsf{T}} - k \cdot I\|_F \sqrt{d-r} && \text{by Hoffman-Wielandt} \\
&\geqslant k(d-r) - k\sqrt{d} && \text{by (5.2).}
\end{aligned}$$

We have shown that $\epsilon d \geqslant (d-r) - \sqrt{d}$, which is precisely what the theorem claims. To extend the proof to irrational distributions μ, one considers a sequence of rational distributions that converges to μ. □

We are now in a position to relate the SQ dimension to the approximate rank.

THEOREM 5.3 (SQ dimension vs. approximate rank). *Let $\mathcal{C}$ be a concept class. Then for $0 \leqslant \epsilon < 1$,*

$$\operatorname{rank}_\epsilon(\mathcal{C}) \geqslant (1-\epsilon^2) \operatorname{sqdim}(\mathcal{C}) - \sqrt{\operatorname{sqdim}(\mathcal{C})}. \tag{5.4}$$

Moreover, there exists a concept class $\mathcal{A}$ with

$$\begin{aligned}
\operatorname{sqdim}(\mathcal{A}) &\leqslant O(n^2), \\
\operatorname{rank}_{1/3}(\mathcal{A}) &\geqslant 2^{\Omega(n/\log n)}.
\end{aligned}$$

PROOF. Let $r = \operatorname{rank}_\epsilon(\mathcal{C})$. Then there are functions $\phi_1, \ldots, \phi_r$ such that each $f \in \mathcal{C}$ has $\|f - \sum_{i=1}^{r} \alpha_i \phi_i\|_\infty \leqslant \epsilon$ for some reals $\alpha_1, \ldots, \alpha_r$. As a result,

$$\mathop{\mathbf{E}}_{\mu} \left[\left(f - \sum_{i=1}^{r} \alpha_i \phi_i \right)^2 \right] \leqslant \epsilon^2$$

for every distribution μ. By Theorem 5.1,

$$r \geqslant (1-\epsilon^2)\,\mathrm{sqdim}_\mu(\mathcal{C}) - \sqrt{\mathrm{sqdim}_\mu(\mathcal{C})}.$$

Maximizing over μ establishes (5.4).

To prove the second part, let $\mathcal{A} = \{\mathrm{MAJ}_n(\pm x_1, \ldots, \pm x_n)\}$. Theorem 4.8 shows that $\mathcal{A}$ has the stated approximate rank. To bound its SQ dimension, note that each function in $\mathcal{A}$ can be pointwise approximated within error $1-1/n$ by a linear combination of the functions $x_1, \ldots, x_n$. Therefore, (5.4) implies that $\mathrm{sqdim}(\mathcal{A}) \leqslant O(n^2)$. □

Remark. It was shown earlier (Sherstov 2008b) that every concept class $\mathcal{C}$ obeys

$$\lim_{\epsilon \nearrow 1} \mathrm{rank}_\epsilon(\mathcal{C}) \geqslant \sqrt{\frac{1}{2}\,\mathrm{sqdim}(\mathcal{C})}.$$

This lower bound is stronger than (5.4) for all sufficiently large $\epsilon < 1$. On the other hand, the proof in this paper gives a quadratically better bound for constant $0 < \epsilon < 1$ and is technically simpler.

6. Related work

Approximate rank and dimension complexity. Dimension complexity is a fundamental and well-studied notion (Forster 2002; Forster & Simon 2006; Linial *et al.* 2007). It is defined for a sign matrix M as

$$\mathrm{dc}(M) = \min_A \{\mathrm{rank}(A) : A \text{ real}, A_{ij}M_{ij} > 0 \text{ for all } i, j\}.$$

In words, the dimension complexity of M is the smallest rank of a real matrix A that has the same sign pattern as M. Thus, $\mathrm{rank}_\epsilon(M) \geqslant \mathrm{dc}(M)$ for each sign matrix M and $0 \leqslant \epsilon < 1$. The dimension complexity of a concept class is defined as the dimension complexity of its characteristic matrix.

Ben-David *et al.* (2003) showed that almost all concept classes with constant VC dimension have dimension complexity $2^{\Omega(n)}$; recall that $\mathrm{dc}(\mathcal{C}) \leqslant 2^n$ always. No lower bounds were known for any explicit concept class until the breakthrough work of Forster (2002), who showed that any sign matrix with small spectral norm has high dimension complexity. Several extensions and refinements of Forster's method were proposed in subsequent work (Forster *et al.* 2001; Forster & Simon 2006; Linial *et al.* 2007).

However, this rich body of work is not readily applicable to our problem. The three matrices that we study have trivial dimension complexity, and we derive lower bounds on the approximate rank that are exponentially larger. Furthermore, in Theorem 1.3 we are able to exhibit an explicit concept class with approximate rank $\Omega(2^n/n)$, whereas the highest dimension complexity proved for any explicit concept class is Forster's lower bound of $2^{n/2}$. The key to our results is to bring out, through a variety of techniques, the additional structure in approximation that is not present in sign-representation.

Approximate rank and rigidity. Approximate rank is also closely related to *ϵ-rigidity*, a variant of matrix rigidity introduced by Lokam (2001). For a fixed real matrix A, its ϵ-rigidity function is defined as

$$R_A(r,\epsilon) = \min_B\{\text{weight}(A-B) : \text{rank}(B) \leqslant r,\ \|A-B\|_\infty \leqslant \epsilon\},$$

where $\text{weight}(A-B)$ stands for the number of nonzero entries in $A-B$. In words, $R_A(r,\epsilon)$ is the minimum number of entries of A that must be perturbed to reduce its rank to r, provided that the perturbation to any single entry is at most ϵ. We immediately have:

$$\text{rank}_\epsilon(A) = \min\{r : R_A(r,\epsilon) \leqslant mn\} \qquad (A \in \mathbb{R}^{m\times n}).$$

As a result, lower bounds on ϵ-rigidity translate into lower bounds on approximate rank. In particular, ϵ-rigidity is a more complicated and nuanced quantity. Nontrivial lower bounds on ϵ-rigidity are known for some special matrix families, most notably the Hadamard matrices (Kashin & Razborov 1998; Lokam 2001). Unfortunately, these results are not applicable to the matrices in our work (see Section 4). To obtain near-optimal lower bounds on approximate rank, we use specialized techniques that target approximate rank without attacking the harder problem of ϵ-rigidity.

Recent progress. In recent work on communication complexity, a technique called the *pattern matrix method* (Sherstov 2010) was developed that converts lower bounds on the approximate degree of Boolean functions into lower bounds on the communication complexity of the corresponding Boolean matrices. To illustrate, fix an arbitrary function $f\colon \{-1,1\}^n \to \{-1,1\}$ and let A_f be the matrix whose columns are each an application of f to some subset of the variables $x_1, x_2, \ldots, x_{4n}$. The pattern matrix method shows that A_f has *bounded-error* communication complexity $\Omega(d)$, where d is the approximate degree of f, i.e., the least degree of a real polynomial p with $\|f-p\|_\infty \leqslant 1/3$. In the same

way, the pattern matrix method converts lower bounds on the approximate degree of Boolean functions into lower bounds on the approximate rank of the corresponding matrices. These new results generalize and strengthen the lower bounds in Section 4.

In another paper (Sherstov 2008a), existence was proved for a concept class $\mathcal{C}$ of functions $\{-1,1\}^n \to \{-1,1\}$ such that $\mathrm{sqdim}(\mathcal{C}) = O(1)$ but $\mathrm{rank}_{1/3}(\mathcal{C}) \geqslant \mathrm{dc}(\mathcal{C}) \geqslant 2^{(1-\epsilon)n}$, for any desired constant $\epsilon > 0$. This separation is essentially optimal and improves on Theorem 5.3 of this paper, although the new concept class is no longer explicitly given.

7. Conclusions and open problems

This paper studies the ϵ-approximate rank of a concept class $\mathcal{C}$, defined as the minimum size of a set of features whose linear combinations can pointwise approximate each $f \in \mathcal{C}$ within ϵ. Our main results give exponential lower bounds on the approximate rank even for the simplest concept classes. These in turn establish exponential lower bounds on the running time of the known algorithms for distribution-free agnostic learning. An obvious open problem is to develop an approach to agnostic learning that does not rely on pointwise approximation by a small set of features.

Another open problem is to prove strong lower bounds on the dimension complexity and SQ dimension of natural concept classes. We have shown that

$$\mathrm{rank}_{1/3}(\mathcal{C}) \geqslant \frac{1}{2}\,\mathrm{sqdim}(\mathcal{C}) - O(1)$$

for each concept class $\mathcal{C}$, and it is further clear that $\mathrm{rank}_{\epsilon}(\mathcal{C}) \geqslant \mathrm{dc}(\mathcal{C})$. In this sense, lower bounds on the approximate rank are prerequisites for lower bounds on dimension complexity and the SQ dimension. Of particular interest in this respect are polynomial-size DNF formulas and, more broadly, AC^0 circuits. While this paper obtains strong lower bounds on their approximate rank, it remains a hard open problem to prove an exponential lower bound on their SQ dimension. An exponential lower bound on the dimension complexity of polynomial-size DNF formulas has recently been obtained (Razborov & Sherstov 2008).

References

Noga Alon (2003). Problems and results in extremal combinatorics, Part I. *Discrete Mathematics* **273**(1-3), 31–53.

Shai Ben-David, Nadav Eiron & Hans Ulrich Simon (2003). Limitations of learning via embeddings in Euclidean half spaces. *J. Mach. Learn. Res.* **3**, 441–461.

Anna Bernasconi (1998). *Mathematical techniques for the analysis of Boolean functions*. Ph.D. thesis, Institute for Computational Mathematics, Pisa.

Avrim Blum, Merrick Furst, Jeffrey Jackson, Michael Kearns, Yishay Mansour & Steven Rudich (1994). Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *Proc. of the 26th Symposium on Theory of Computing (STOC)*, 253–262.

Nader H. Bshouty & Christino Tamon (1996). On the Fourier spectrum of monotone functions. *J. ACM* **43**(4), 747–770.

Harry Buhrman, Ilan Newman, Hein Röhrig & Ronald de Wolf (2007a). Robust polynomials and quantum algorithms. *Theory Comput. Syst.* **40**(4), 379–395.

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

Harry Buhrman & Ronald de Wolf (2001). Communication complexity lower bounds by polynomials. In *Proc. of the 16th Conf. on Computational Complexity (CCC)*, 120–130.

Herman Chernoff (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.* **23**(4), 493–507.

Scott E. Decatur (1993). Statistical queries and faulty PAC oracles. In *Proc. of the 6th Conf. on Computational Learning Theory (COLT)*, 262–268.

Jürgen Forster (2002). A linear lower bound on the unbounded error probabilistic communication complexity. *J. Comput. Syst. Sci.* **65**(4), 612–625.

Jürgen Forster, Matthias Krause, Satyanarayana V. Lokam, Rustam Mubarakzjanov, Niels Schmitt & Hans-Ulrich Simon (2001). Relations between communication complexity, linear arrangements, and computational complexity. In *Proc. of the 21st Conf. on Foundations of Software Technology and Theoretical Computer Science (FST TCS)*, 171–182.

Jürgen Forster & Hans Ulrich Simon (2006). On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes. *Theor. Comput. Sci.* **350**(1), 40–48.

Gene H. Golub & Charles F. Van Loan (1996). *Matrix computations*. Johns Hopkins University Press, Baltimore, 3rd edition.

Jeffrey Charles Jackson (1995). *The harmonic sieve: A novel application of Fourier analysis to machine learning theory and practice*. Ph.D. thesis, Carnegie Mellon University.

Adam Tauman Kalai, Adam R. Klivans, Yishay Mansour & Rocco A. Servedio (2008). Agnostically learning halfspaces. *SIAM J. Comput.* **37**(6), 1777–1805.

Boris S. Kashin & Alexander A. Razborov (1998). Improved lower bounds on the rigidity of Hadamard matrices. *Matematicheskie zametki* **63**(4), 535–540. In Russian.

Michael J. Kearns (1998). Efficient noise-tolerant learning from statistical queries. *J. ACM* **45**(6), 983–1006.

Michael J. Kearns & Ming Li (1993). Learning in the presence of malicious errors. *SIAM J. Comput.* **22**(4), 807–837.

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

Michael J. Kearns & Umesh V. Vazirani (1994). *An Introduction to Computational Learning Theory*. MIT Press, Cambridge.

Adam R. Klivans, Ryan O'Donnell & Rocco A. Servedio (2004). Learning intersections and thresholds of halfspaces. *J. Comput. Syst. Sci.* **68**(4), 808–840.

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

Eyal Kushilevitz & Yishay Mansour (1993). Learning decision trees using the Fourier spectrum. *SIAM J. Comput.* **22**(6), 1331–1348.

Eyal Kushilevitz & Noam Nisan (1997). *Communication complexity*. Cambridge University Press, New York.

Nathan Linial, Yishay Mansour & Noam Nisan (1993). Constant depth circuits, Fourier transform, and learnability. *J. ACM* **40**(3), 607–620.

Nathan Linial & Adi Shraibman (2009a). Learning complexity vs communication complexity. *Combinatorics, Probability & Computing* **18**(1-2), 227–245.

Nati Linial, Shahar Mendelson, Gideon Schechtman & Adi Shraibman (2007). Complexity measures of sign matrices. *Combinatorica* **27**(4), 439–463.

Nati Linial & Adi Shraibman (2009b). Lower bounds in communication complexity based on factorization norms. *Random Struct. Algorithms* **34**(3), 368–394.

Satyanarayana V. Lokam (2001). Spectral methods for matrix rigidity with applications to size-depth trade-offs and communication complexity. *J. Comput. Syst. Sci.* **63**(3), 449–473.

Yishay Mansour (1995). An $O(n^{\log\log n})$ learning algorithm for DNF under the uniform distribution. *J. Comput. Syst. Sci.* **50**(3), 543–550.

Yishay Mansour & Michal Parnas (1996). On learning conjunctions with malicious noise. In *Proc. of the 4th Israel Symposium on Theory of Computing and Systems (ISTCS)*, 170–175.

Noam Nisan & Mario Szegedy (1994). On the degree of Boolean functions as real polynomials. *Computational Complexity* **4**, 301–313.

Ryan O'Donnell & Rocco A. Servedio (2008). Extremal properties of polynomial threshold functions. *J. Comput. Syst. Sci.* **74**(3), 298–312.

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

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

Alexander A. Razborov & Alexander A. Sherstov (2008). The sign-rank of AC^0. In *Proc. of the 49th Symposium on Foundations of Computer Science (FOCS)*, 57–66.

Alexander A. Sherstov (2008a). Communication complexity under product and nonproduct distributions. In *Proc. of the 23rd Conf. on Computational Complexity (CCC)*, 64–70.

Alexander A. Sherstov (2008b). Halfspace matrices. *Comput. Complex.* **17**(2), 149–178. Preliminary version in 22nd CCC, 2007.

Alexander A. Sherstov (2010). The pattern matrix method. *SIAM J. Comput.* To appear. Preliminary version in 40th STOC, 2008.

Leslie G. Valiant (1985). Learning disjunction of conjunctions. In *Proc. of the 9th International Joint Conference on Artificial Intelligence (IJCAI)*, 560–566.

A. Discrepancy and approximate rank

The purpose of this section is to prove the relationship between the discrepancy and approximate rank needed in Section 4. We start with several definitions and auxiliary results due to Linial *et al.* (2007) and Linial & Shraibman (2009a,b).

For a real matrix A, let $\|A\|_{1\to 2}$ denote the largest Euclidean norm of a column of A, and let $\|A\|_{2\to\infty}$ denote the largest Euclidean norm of a row of A. Define

$$\gamma_2(A) = \min_{XY=A} \|X\|_{2\to\infty}\|Y\|_{1\to 2}.$$

For a sign matrix M, its *margin complexity* is defined as

$$\mathrm{mc}(M) = \min\{\gamma_2(A) : A \text{ real}, \;\; A_{ij}M_{ij} \geqslant 1 \text{ for all } i,j\}.$$

Lemma A.1 (Linial *et al.* 2007, Lem. 4.2). *Let A be a real matrix. Then*

$$\gamma_2(A) \leqslant \sqrt{\mathrm{rank}(A) \cdot \|A\|_\infty}.$$

Theorem A.2 (Linial & Shraibman 2009a, Thm. 3.1). *Let M be a sign matrix. Then*

$$\mathrm{mc}(M) \geqslant \frac{1}{8\,\mathrm{disc}(M)}.$$

Putting these pieces together yields the desired result:

Lemma 3.7 (Restated from Sec. 3.2). *Let M be a sign matrix and $0 \leqslant \epsilon < 1$. Then*

$$\mathrm{rank}_\epsilon(M) \geqslant \frac{1-\epsilon}{1+\epsilon} \cdot \frac{1}{64\,\mathrm{disc}(M)^2}.$$

PROOF. Let A be any real matrix with $\|A - M\|_\infty \leqslant \epsilon$. Put $B = \frac{1}{1-\epsilon}A$. We have:

$$\begin{aligned}
\operatorname{rank}(A) &= \operatorname{rank}(B) \\
&\geqslant \frac{\gamma_2(B)^2}{\|B\|_\infty} && \text{by Lemma A.1} \\
&\geqslant \frac{\operatorname{mc}(M)^2}{\|B\|_\infty} \\
&\geqslant \frac{1}{\|B\|_\infty} \cdot \frac{1}{64 \operatorname{disc}(M)^2} && \text{by Theorem A.2} \\
&\geqslant \frac{1-\epsilon}{1+\epsilon} \cdot \frac{1}{64 \operatorname{disc}(M)^2},
\end{aligned}$$

as claimed. □

Manuscript received January 26, 2008

ADAM R. KLIVANS
Department of Computer Sciences
The University of Texas at Austin
1 University Station C0500
Austin, TX 78712-0233 USA
klivans@cs.utexas.edu

ALEXANDER A. SHERSTOV
Department of Computer Sciences
The University of Texas at Austin
1 University Station C0500
Austin, TX 78712-0233 USA
sherstov@cs.utexas.edu