HALFSPACE MATRICES

Alexander A. Sherstov

Abstract. We introduce the notion of a *halfspace matrix,* which is a sign matrix A with rows indexed by linear threshold functions f, columns indexed by inputs $x \in \{-1,1\}^n$, and the entries given by $A_{f,x} = f(x)$. We use halfspace matrices to solve the following problems.

In communication complexity, we exhibit a Boolean function f with discrepancy $\Omega(1/n^4)$ under every product distribution but $O(\sqrt{n}/2^{n/4})$ under a certain non-product distribution. This partially solves an open problem of Kushilevitz and Nisan (1997).

In learning theory, we give a short and simple proof that the statistical-query (SQ) dimension of halfspaces in n dimensions is less than $2(n+1)^2$ under all distributions. This improves on the $n^{O(1)}$ estimate from the fundamental paper of Blum et al. (1998). We show that estimating the SQ dimension of natural classes of Boolean functions can resolve major open problems in complexity theory, such as separating $\mathsf{PSPACE}^{\mathrm{cc}}$ and $\mathsf{PH}^{\mathrm{cc}}$.

Finally, we construct a matrix $A \in \{-1,1\}^{N \times N^{\log N}}$ with dimension complexity $\log N$ but margin complexity $\Omega(N^{1/4}/\sqrt{\log N})$. This gap is an exponential improvement over previous work. We prove several other relationships among the complexity measures of sign matrices, complementing work by Linial et al. (2006, 2007).

Keywords. Linear threshold functions, communication complexity, complexity measures of sign matrices, complexity of learning

Subject classification. 03D15, 68Q15, 68Q17.

1. Introduction

A *linear threshold function*, or *halfspace,* is a Boolean function f representable as $f(x) \equiv \operatorname{sign}(\sum_{i=1}^n a_i x_i - \theta)$ for some reals $a_1, \ldots, a_n, \theta$. We introduce the notion of a *halfspace matrix*, which is a ± 1-valued matrix A with rows indexed by halfspaces, columns indexed by inputs $x \in \{-1,1\}^n$, and the entries given by $A_{f,x} = f(x)$. We demonstrate the potential of halfspace matrices in the study of complexity. Specifically, we use halfspace matrices to answer nontrivial open

questions in communication complexity, the complexity of sign matrices, and the complexity of learning.

Our work is inspired by Forster's groundbreaking result (2002) on the sign-representation of Boolean matrices by real ones. Forster's work has had exciting applications, including a linear lower bound on communication complexity in the unbounded-error model (Forster 2002), extremal results on Euclidean embeddings (Forster 2002; Forster *et al.* 2003; Forster & Simon 2006), and lower bounds for depth-2 threshold circuits (Forster *et al.* 2001). This paper builds on Forster's discovery and related work to illustrate the power of halfspace matrices.

1.1. Communication complexity. Among the primary models in communication complexity is the *randomized model* (Kushilevitz & Nisan 1997, Chapter 3). Two parties, Alice and Bob, have access to disjoint parts $x, y \in \{-1,1\}^n$ of the input to a fixed function $f : \{-1,1\}^n \times \{-1,1\}^n \to \{-1,1\}$ and must therefore communicate to evaluate $f(x,y)$. They are allowed to use shared random bits. On every input, the players must compute the correct value with probability at least $2/3$. The *cost* of a protocol is number of bits exchanged in the worst case. The *randomized complexity* $\mathrm{R}^{\mathrm{pub}}_{1/3}(f)$ of a function f is the cost of the best protocol for f.

The standard approach to proving lower bounds on $\mathrm{R}^{\mathrm{pub}}_{1/3}(f)$ is to analyze the *distributional complexity* $\mathrm{D}^{\mu}_{1/3}(f)$ instead. One defines a probability distribution μ on $\{-1,1\}^n \times \{-1,1\}^n$ and argues that the cost $\mathrm{D}^{\mu}_{1/3}(f)$ of the best deterministic protocol with error at most $1/3$ under μ must be high. Yao's well-known Minimax Principle states that $\mathrm{R}^{\mathrm{pub}}_{1/3}(f) = \max_{\mu}\{\mathrm{D}^{\mu}_{1/3}(f)\}$. The main design question, then, is what distribution μ to consider. While *product* distributions $\mu(x,y) = \mu_X(x)\mu_Y(y)$ are easier to analyze, they do not always yield the optimal lower bounds. A standard example of this phenomenon is the set disjointness function DISJ: every product distribution μ has $\mathrm{D}^{\mu}_{1/3}(\mathrm{DISJ}) = O(\sqrt{n}\log n)$ (Kushilevitz & Nisan 1997), although $\mathrm{R}^{\mathrm{pub}}_{1/3}(\mathrm{DISJ}) = \Theta(n)$ (Kalyanasundaram & Schnitger 1992; Razborov 1992). This motivates the following intriguing question in communication complexity, due to Kushilevitz and Nisan.

Problem (Kushilevitz & Nisan 1997, page 37). *Can restricting the distribution μ to be a product distribution affect the resulting lower bound on $\mathrm{R}^{\mathrm{pub}}_{1/3}(f)$ by more than a polynomial factor? Formally:*

$$\mathrm{R}^{\mathrm{pub}}_{1/3}(f) \quad \stackrel{?}{=} \quad \left(\max_{\mu \text{ product}} \left\{ \mathrm{D}^{\mu}_{1/3}(f) \right\} \right)^{O(1)}.$$

Since its formulation 10 years ago, this problem has seen little progress. The only work known to us is due to Kremer *et al.* (1999), who study the restriction of this problem to *one-way* protocols. They obtain a separation of $O(1)$ vs. $\Omega(n)$ for the "greater than" function GT, in the one-way model. Unfortunately, a function can have vastly different communication complexity in the one-way and normal (two-way multi-round) randomized models. Such is the case of GT, whose one-way randomized complexity is $\Omega(n)$ but normal randomized complexity $O(\log n)$. Therefore, new techniques are needed to answer the Kushilevitz-Nisan question in its original formulation.

This motivates us to pursue a different approach. The chief source of lower bounds on distributional complexity $\mathrm{D}^{\mu}_{1/3}(f)$ and thus randomized complexity $\mathrm{R}^{\mathrm{pub}}_{1/3}(f)$ is the so-called *discrepancy method*. The method lower-bounds $\mathrm{D}^{\mu}_{1/3}(f)$ in terms of a quantity called *discrepancy,* $\mathrm{disc}_{\mu}(f)$. (Small discrepancy implies high communication complexity.) A natural question to ask is whether there can be a large gap between the discrepancy under product and non-product distributions. Our first main result states that, in fact, this gap can be exponential:

THEOREM 1.1 (Discrepancy gap). *There exists an (explicitly given) function* $f : \{-1,1\}^n \times \{-1,1\}^{n^2} \to \{-1,1\}$ *for which* $\mathrm{disc}_{\mu}(f) = \Omega(1/n^4)$ *under every product distribution* μ *but* $\mathrm{disc}_{\lambda}(f) = O(\sqrt{n}/2^{n/4})$ *under a certain non-product distribution* λ.

We thus establish that discrepancy-based methods can yield exponentially worse lower bounds on randomized complexity $\mathrm{R}^{\mathrm{pub}}_{1/3}(f)$ if restricted to product distributions. This is the first nontrivial discrepancy gap obtained for any function.

Recent progress. Building on the ideas in this manuscript, the author has recently solved (Sherstov 2007) the Kushilevitz-Nisan problem in its original formulation. Namely, we have shown the existence of a function $f : \{-1,1\}^n \times \{-1,1\}^n \to \{-1,1\}$ for which $\max_{\mu \text{ product}} \{\mathrm{D}^{\mu}_{\epsilon}(f)\} = O(1)$ but $\mathrm{R}^{\mathrm{pub}}_{1/3}(f) = \Omega(n)$, where $\epsilon > 0$ is an arbitrary constant. Furthermore, the function f in question has discrepancy $O(2^{-n(1/2-\epsilon)})$, which is almost the smallest possible. In particular, f is a hardest function for every major model of communication. Yet, the distributional method restricted to product distributions can certify at best an $\Omega(1)$ communication lower bound for f.

The work in (Sherstov 2007) also yields an essentially optimal separation, $\Omega(1)$ vs. $O(2^{-n(1/2-\epsilon)})$, between discrepancy under product and nonproduct distributions, improving on Theorem 1.1 above.

However, the results of Sherstov (2007) do not supersede Theorem 1.1. Specifically, the Kushilevitz-Nisan problem and the optimal discrepancy gap are settled there using the probabilistic method, for a *nonexplicit* function. All functions in this work, on the other hand, are given explicitly.

1.2. Complexity of sign matrices. A *sign matrix* is any matrix with ± 1 entries. A systematic study of sign matrices from a complexity-theoretic perspective has been recently initiated by Linial *et al.* (2006). Apart from the inherent interest of this subject as a new area of complexity, Linial *et al.* observe that several major problems in theoretical computer science are questions about sign matrices. Indeed, research into sign matrices has already yielded excellent complexity results (Forster 2002; Forster *et al.* 2001).

Our paper continues this investigation, focusing on the two main complexity measures of a sign matrix: dimension and margin complexity. Their formal definition is as follows. A *Euclidean embedding* of a sign matrix $A \in \{-1,1\}^{M \times N}$ is a collection of unit-length vectors $\mathbf{u}_1, \ldots, \mathbf{u}_M \in \mathbb{R}^k$ and $\mathbf{v}_1, \ldots, \mathbf{v}_N \in \mathbb{R}^k$ (for some k) such that $\langle \mathbf{u}_i, \mathbf{v}_j \rangle \cdot A_{ij} > 0$ for all i, j. The integer k is the *dimension* of the embedding. The quantity $\gamma = \min_{i,j} |\langle \mathbf{u}_i, \mathbf{v}_j \rangle|$ is the *margin* of the embedding. The *dimension complexity* $\mathrm{dc}(A)$ is the smallest dimension of an embedding of A. The *margin complexity* $\mathrm{mc}(A)$ is the minimum $1/\gamma$ over all embeddings of A.

Both dimension complexity and margin complexity have drawn much interest (Ben-David *et al.* 2003; Forster 2002; Forster *et al.* 2001, 2003; Forster & Simon 2006; Linial *et al.* 2006). In addition to their roles as complexity measures, dimension complexity is a key player in the unbounded-error model of communication (Alon *et al.* 1985; Paturi & Simon 1986), and margin complexity is the central notion in the highly successful *kernel methods* of machine learning (Burges 1998; Vapnik 1995).

Using the random projection technique of Arriaga & Vempala (2006), it is straightforward to show (Ben-David *et al.* 2003) that

$$\mathrm{dc}(A) = O(\mathrm{mc}(A)^2 \log(M+N))$$

for every $M \times N$ sign matrix. This observation has had important algorithmic applications, such as the algorithm of Klivans & Servedio (2004) for learning certain intersections of halfspaces. In this paper, we ask the opposite question: *Can one place an upper bound on margin complexity in terms of dimension complexity?* A suitable upper bound of this type would establish the distribution-free weak learnability of unrestricted intersections of two halfspaces, leading to

a major breakthrough in the area (Klivans & Sherstov 2007). Unfortunately, we give a strong negative answer to this question (Theorem 1.2 below).

This problem of estimating the gap between dimension and margin complexity has been studied by several researchers. Forster *et al.* (2003) constructed a family of matrices $A \in \{-1,1\}^{N\times N}$ for which $\mathrm{dc}(A) = O(1)$ but $\mathrm{mc}(A) = \Theta(\log N)$. Srebro & Shraibman (2005) amplified this separation, obtaining $\mathrm{dc}(A) \leqslant 2^p$ and $\mathrm{mc}(A) = (\log N)^{\Theta(p)}$ for any choice of the parameter $1 \leqslant p \leqslant (\log N)^{1-\epsilon}$. Our second main theorem is an exponential improvement over these results.

THEOREM 1.2 (Margin vs. dimension). *There is an (explicitly given) matrix* $A \in \{-1,1\}^{N\times N^{\log N}}$ *for which* $\mathrm{dc}(A) \leqslant \log N$ *but* $\mathrm{mc}(A) = \Omega(N^{1/4}/\sqrt{\log N})$.

Note. An exponential separation between margin and dimension complexity was independently obtained by Buhrman, Vereshchagin & de Wolf (2007) and appeared in the same conference proceedings as this paper. The proof by Buhrman *et al.* features a different matrix and completely different techniques (approximation theory and quantum communication complexity). Buhrman *et al.* phrase their result as a separation between the classes $\mathsf{PP}^{\mathrm{cc}}$ and $\mathsf{UPP}^{\mathrm{cc}}$ in communication complexity.

The exponential separation in Theorem 1.2 is quite close to optimal since every matrix $A \in \{-1,1\}^{M\times N}$ has $1 \leqslant \mathrm{dc}(A) \leqslant \min\{M,N\}$ and $1 \leqslant \mathrm{mc}(A) \leqslant \min\{\sqrt{M},\sqrt{N}\}$ (see Section 2).

As an application of our analysis in Theorem 1.2, we consider a problem from circuit complexity (Goldmann *et al.* 1992). Fix d arbitrary functions $f_1,\ldots,f_d : \{-1,1\}^n \to \{-1,1\}$. Assume that every halfspace can be computed as a majority vote of gates from among $f_1,\ldots,f_d$. We prove that there are halfspaces that require circuits of size $\Omega(2^{n/4}/(d\sqrt{n}))$ in this model. This generalizes the well-known fact (Håstad 1994; Siu & Bruck 1991) that some halfspaces require exponentially large weights, and complements a result due to Goldmann *et al.* (1992). See Section 6.1 for details.

We prove a number of additional results (see Section 7). In particular, we show that the standard complexity measures ($\mathrm{dc}(A)$, $\mathrm{mc}(A)$, and a new complexity measure $\mathrm{sq}(A)$ that we introduce) form an ordered sequence that spans the continuum between $\mathrm{disc}^{\times}(A)^{-1}$ and $\mathrm{disc}(A)^{-1}$. Here $\mathrm{disc}^{\times}(A)$ is the discrepancy under product distributions, and $\mathrm{disc}(A)$ is general discrepancy. This close interplay between linear-algebraic complexity measures ($\mathrm{dc}(A)$ and $\mathrm{mc}(A)$) and those from communication complexity ($\mathrm{disc}^{\times}(A)$ and $\mathrm{disc}(A)$) is further evidence that the study of sign matrices has much to contribute to complexity theory.

1.3. Learning theory. We adopt the *statistical query* (SQ) model of learning, due to Kearns (1993), which is a restriction of the standard PAC learning model (Valiant 1984). Fix a set $\mathcal{C}$ of Boolean functions $\{-1,1\}^n \to \{-1,1\}$ (a *concept class*) and a distribution μ over $\{-1,1\}^n$. For each choice of an unknown function $f \in \mathcal{C}$, the learner in the SQ model must be able to construct an approximation to f by asking queries of the form, "What is $\mathbf{E}_{x\sim\mu}[G(x,f(x))]$, approximately?" Here $G : \{-1,1\}^n \times \{-1,1\} \to \{-1,1\}$ is any polynomial-time computable predicate of the learner's choosing, distinct for each query. Extensive research has established the SQ model as a powerful and elegant abstraction of learning (Blum *et al.* 1994, 2003; Kearns & Vazirani 1994; Klivans & Sherstov 2007; Yang 2005). In particular, essentially all known PAC learning algorithms can be adapted (Kearns & Vazirani 1994) to work in the SQ model. Furthermore, SQ algorithms are inherently robust to random classification noise since they use *statistics* instead of individual labeled examples.

A measure of the learning complexity of a given concept class $\mathcal{C}$ under a given distribution μ is its *statistical query (SQ) dimension*, $\mathrm{sqdim}_\mu(\mathcal{C})$. This complexity measure captures both the running time and the sample complexity required to weakly learn $\mathcal{C}$ in the SQ model. Informally, $\mathrm{sqdim}_\mu(\mathcal{C})$ is the size of the largest subset $\mathcal{F} \subseteq \mathcal{C}$ of (almost) mutually orthogonal functions in $\mathcal{C}$ under μ. We put $\mathrm{sqdim}(\mathcal{C}) \stackrel{\text{def}}{=} \max_\mu\{\mathrm{sqdim}_\mu(\mathcal{C})\}$. We delay the technical details to Section 2.

Our next result concerns the concept class of halfspaces. This concept class is arguably the most studied one (Klivans *et al.* 2004; Klivans & Servedio 2004; Klivans & Sherstov 2006, 2007; Kwek & Pitt 1998; Vempala 1997) in computational learning theory, with applications in areas as diverse as data mining, artificial intelligence, and computer vision. In a fundamental paper, Blum *et al.* (1998) gave a polynomial-time algorithm for learning halfspaces in the SQ model under arbitrary distributions. It follows from their work that the SQ dimension of halfspaces is $O(n^c)$, where $c > 0$ is a sufficiently large constant. We substantially sharpen this estimate:

THEOREM 1.3 (SQ dimension of generalized halfspaces). *Fix arbitrary functions $\phi_1, \ldots, \phi_k : \{-1,1\}^n \to \mathbb{R}$. Let $\mathcal{C}$ be the set of all Boolean functions f representable as $f(x) \equiv \mathrm{sign}(\sum_{i=1}^k a_i\phi_i(x))$ for some reals $a_1, \ldots, a_k$. Then $\mathrm{sqdim}_\mu(\mathcal{C}) < 2k^2$ under all distributions μ.*

COROLLARY 1.4 (SQ dimension of halfspaces). *Let $\mathcal{C}$ be the concept class of halfspaces in n dimensions. Then $\mathrm{sqdim}_\mu(\mathcal{C}) < 2(n+1)^2$ under all distributions μ.*

Prior to our work, Simon (2006, Cor. 8) proved the special case of Theorem 1.3 for μ uniform. We generalize his result to arbitrary distributions μ.

The SQ dimension of halfspaces is at least $n+1$ under the uniform distribution (consider the functions $x_1, x_2, \ldots, x_n, 1$). Thus, the quadratic upper bound of Corollary 1.4 is not far from optimal. In addition to strengthening the estimate of Blum *et al.*, Theorem 1.3 has a much simpler, one-page proof that builds only on Forster's self-contained theorem (2002). The proof of Blum *et al.* relies on nontrivial notions from computational geometry and requires a lengthy analysis of robustness under noise. That said, our result gives only a nonuniform SQ algorithm for weakly learning halfspaces under an arbitrary (but fixed and known) distribution, whereas Blum *et al.* give an explicit SQ algorithm for strongly learning halfspaces under arbitrary distributions.

To best convey the importance—outside of learning theory—of studying the SQ dimension of natural classes of functions, we establish the following final result.

THEOREM 1.5 (On the conjecture that IP $\in \mathsf{PSPACE}^{\mathrm{cc}} \setminus \mathsf{PH}^{\mathrm{cc}}$). *Let $\mathcal{C}$ be the class of functions $\{-1,1\}^n \to \{-1,1\}$ computable in AC^0. If $\mathrm{sqdim}(\mathcal{C}) \leqslant O\left(2^{2^{(\log n)^{\epsilon}}}\right)$ for every constant $\epsilon > 0$, then* IP $\in \mathsf{PSPACE}^{\mathrm{cc}} \setminus \mathsf{PH}^{\mathrm{cc}}$.

Thus, a suitable upper bound on the SQ dimension of AC^0 circuits would separate the communication-complexity analogues of the polynomial hierarchy (PH) and polynomial space (PSPACE). This latter problem is a major unresolved question in theoretical computer science that dates back to a paper by Babai *et al.* (1986). Viewed from a different standpoint, Theorem 1.5 explains the lack of progress in designing learning algorithms for AC^0: a distribution-free SQ algorithm for weakly learning AC^0 in reasonable time would settle a major open question in complexity theory.

The SQ upper bound, $\exp\left(2^{(\log n)^{o(1)}}\right)$, for AC^0 assumed in Theorem 1.5 grows faster than any quasipolynomial function but more slowly than any exponential one ($2^{n^{\epsilon}}$). In particular, any quasipolynomial upper bound on the SQ dimension of AC^0 would separate $\mathsf{PH}^{\mathrm{cc}}$ and $\mathsf{PSPACE}^{\mathrm{cc}}$. At present, however, no upper bounds better than $2^{\tilde{O}(n^{1/3})}$ are known on the SQ dimension of polynomial-size DNF formulas, let alone AC^0 circuits. We hope that our observations will draw the community's attention to the SQ dimension as an important notion in complexity theory.

1.4. Our techniques. A common theme of this paper is the use of halfspace matrices to answer the extremal questions at hand. Our first technical tool is

Forster's work (2002), which we show constrains the sign patterns of halfspace matrices in an important way. These structural constraints allow us to prove an upper bound on the SQ dimension of halfspaces (Theorem 1.3). Another technical tool that we use is a theorem of Goldmann *et al.* (1992) in communication complexity. We apply it to show that halfspace matrices still do possess some structural complexity. Together, these contrasting results lead to the discrepancy gap (Theorem 1.1) and the margin-dimension gap (Theorem 1.2).

To prove Theorem 1.2, we use a technique for lower-bounding margin complexity based on communication complexity. By contrast, all previous lower bounds (Forster 2002; Forster *et al.* 2001, 2003; Linial *et al.* 2006) for explicit matrices have been based solely on linear algebra. Most of these previous techniques yield identical bounds on dimension and margin complexity, and thus cannot yield the gap of Theorem 1.2.

Finally, our proof of Theorem 1.5 regarding $\mathsf{PSPACE}^{\mathrm{cc}} \setminus \mathsf{PH}^{\mathrm{cc}}$ builds on a manuscript of Razborov (1989) and a combinatorial observation due to Lokam (2001).

Organization. After some technical preliminaries, we first prove the SQ dimension upper bound (Theorem 1.3) and then use it to establish the discrepancy gap and margin-dimension gap (Theorems 1.1 and 1.2). Additional results on the complexity of sign matrices come next. We conclude the paper with observations concerning $\mathsf{PSPACE}^{\mathrm{cc}} \setminus \mathsf{PH}^{\mathrm{cc}}$ (Theorem 1.5).

2. Preliminaries

This section provides relevant background on communication complexity, sign matrices, and learning theory.

2.1. Communication complexity. We consider Boolean functions $f : X \times Y \to \{-1, 1\}$. Typically $X = Y = \{-1, 1\}^n$, but we also allow X and Y to be arbitrary finite sets, possibly of unequal cardinality. We identify a function f with its *communication matrix* $M = [f(x, y)]_{y,x} \in \{-1, 1\}^{|Y| \times |X|}$. In particular, we use the terms "communication complexity of f" and "communication complexity of M" interchangeably (and likewise for other complexity measures, such as discrepancy). The two communication models of interest to us are the randomized model and the deterministic model, both reviewed in Section 1. The *randomized complexity* $\mathrm{R}^{\mathrm{pub}}_{1/2-\gamma/2}(f)$ of f is the minimum cost of a randomized protocol for f that computes $f(x, y)$ correctly with probability

at least $\frac{1}{2} + \frac{\gamma}{2}$ (equivalently, with *advantage* γ) for each input (x, y). The *distributional complexity* $\mathrm{D}^{\mu}_{1/2-\gamma/2}(f)$ is the minimum cost of a deterministic protocol for f that has error at most $\frac{1}{2} - \frac{\gamma}{2}$ (equivalently, *advantage* γ) with respect to the distribution μ over the inputs.

A distribution μ over $X \times Y$ is called a *product distribution* if it can be represented as $\mu = \mu_X \times \mu_Y$ (meaning that $\mu(x, y) = \mu_X(x)\mu_Y(y)$ for all x, y), where μ_X and μ_Y are distributions over X and Y, respectively. 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 let $\mathrm{disc}^{\times}(f)$ denote the minimum discrepancy of f under *product* distributions. Clearly, $\mathrm{disc}(f) \leqslant \mathrm{disc}^{\times}(f)$, and we will show that there can be an exponential gap between these quantities.

The *discrepancy method* is a powerful technique that lower-bounds the randomized and distributional complexity in terms of the discrepancy:

Proposition 2.1 (Kushilevitz & Nisan 1997, pp. 36–38). *For every Boolean function $f(x, y)$, every distribution μ, and every $\gamma > 0$,*

$$\mathrm{R}^{\mathrm{pub}}_{1/2-\gamma/2}(f) \;\geqslant\; \mathrm{D}^{\mu}_{1/2-\gamma/2}(f) \;\geqslant\; \log_2 \frac{\gamma}{\mathrm{disc}_{\mu}(f)}.$$

2.2. Discrepancy under product distributions. We now recall a useful technique from the literature for estimating the discrepancy under a product distribution. Our starting point is an important observation that arises as a special case in the work of Ford & Gál (2005, Thm. 3.1) on multiparty communication complexity. It is also implicit in an article by Raz (2000, Lem. 5.1).

Lemma 2.2 (Ford & Gál 2005; Raz 2000). *Let $M \in \{-1, 1\}^{|X|\times|Y|}$, and let μ be a probability distribution over $X \times Y$. Then there is a choice of signs $\alpha_x, \beta_y \in \{-1, 1\}$ for all $x \in X$, $y \in Y$ such that*

$$\mathrm{disc}_{\mu}(M) \leqslant \left| \sum_{x,y} \alpha_x \beta_y \mu(x, y) M_{xy} \right|.$$

PROOF (adapted from Raz 2000). Let $R = A \times B$ be the rectangle for which the discrepancy is achieved. Fix $\alpha_x = 1$ for all $x \in A$, and likewise $\beta_y = 1$ for all $y \in B$. Choose the remaining signs α_x, β_y independently and at random. Passing to expectations,

$$\begin{aligned}
&\left| \mathbf{E}\left[\sum_{x,y} \alpha_x \beta_y \mu(x,y) M_{xy} \right] \right| \\
&\qquad = \left| \sum_{(x,y)\in R} \mathbf{E}[\underbrace{\alpha_x \beta_y}_{=1}] \mu(x,y) M_{xy} + \sum_{(x,y)\notin R} \underbrace{\mathbf{E}\left[\alpha_x \beta_y\right]}_{=0} \mu(x,y) M_{xy} \right| \\
&\qquad = \left| \sum_{(x,y)\in R} \mu(x,y) M_{xy} \right| \\
&\qquad = \operatorname{disc}_\mu(M).
\end{aligned}$$

In particular, there exists a setting $\alpha_x, \beta_y \in \{-1, 1\}$ for all x, y with the desired property. □

Ford and Gál used Lemma 2.2 in an elegant way to relate discrepancy to the pairwise correlations of the matrix rows:

LEMMA 2.3 (Ford & Gál 2005). *For every Boolean function $f(x, y)$ and every product distribution $\mu = \mu_X \times \mu_Y$,*

$$\operatorname{disc}_\mu(f) \leqslant \sqrt{\mathop{\mathbf{E}}_{y,y' \sim \mu_Y} \left[\left| \mathop{\mathbf{E}}_{x \sim \mu_X} [f(x,y) f(x,y')] \right| \right]}.$$

PROOF (adapted from Ford & Gál 2005). By Lemma 2.2, there is a choice of values $\alpha_x, \beta_y \in \{-1, 1\}$ for all x and y such that

$$\operatorname{disc}_\mu(f) \leqslant \left| \sum_x \sum_y \mu(x,y) \alpha_x \beta_y f(x,y) \right| = \left| \mathop{\mathbf{E}}_{x \sim \mu_X} \mathop{\mathbf{E}}_{y \sim \mu_Y} [\alpha_x \beta_y f(x,y)] \right|.$$

Thus,

$$\begin{aligned}
\operatorname{disc}_\mu(f)^2 &\leqslant \left(\mathop{\mathbf{E}}_{x\sim\mu_X} \mathop{\mathbf{E}}_{y\sim\mu_Y} [\alpha_x \beta_y f(x,y)] \right)^2 \\
&\leqslant \mathop{\mathbf{E}}_{x} \left[\left(\mathop{\mathbf{E}}_{y} [\alpha_x \beta_y f(x,y)] \right)^2 \right] \\
&= \mathop{\mathbf{E}}_{x} \mathop{\mathbf{E}}_{y,y'} \left[\alpha_x^2 \beta_y \beta_{y'} f(x,y) f(x,y') \right] \\
&\leqslant \mathop{\mathbf{E}}_{y,y'} \left[\left| \mathop{\mathbf{E}}_{x} [f(x,y) f(x,y')] \right| \right] \qquad \text{since } \alpha_x^2 = |\beta_y \beta_{y'}| = 1. \quad \square
\end{aligned}$$

2.3. Sign matrices. 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 $F(i,j)$. We denote vectors by boldface letters ($\mathbf{u}, \mathbf{v}, \mathbf{e}_i$, etc.) and scalars by plain letters (u_i, x_j, etc.) A *(Euclidean) embedding* of a matrix $A \in \{-1,1\}^{M\times N}$ is a collection of vectors $\mathbf{u}_1, \dots, \mathbf{u}_M \in \mathbb{R}^k$ and $\mathbf{v}_1, \dots, \mathbf{v}_N \in \mathbb{R}^k$ (for some k) such that $\langle \mathbf{u}_i, \mathbf{v}_j \rangle \cdot A_{ij} > 0$ for all i, j. The integer k is the *dimension* of the embedding. The quantity

$$\gamma = \min_{i,j} \frac{|\langle \mathbf{u}_i, \mathbf{v}_j \rangle|}{\|\mathbf{u}_i\| \cdot \|\mathbf{v}_j\|}$$

is the *margin* of the embedding. The *dimension complexity* $\operatorname{dc}(A)$ is the smallest dimension of an embedding of A. The *margin complexity* $\operatorname{mc}(A)$ is the minimum $1/\gamma$ over all embeddings of A.

Let $\mathbf{e}_i$ denote the vector with 1 in the ith component and zeroes elsewhere. The following is a trivial embedding of a sign matrix $A = [\,\mathbf{a}_1 \,|\, \dots \,|\, \mathbf{a}_N\,] \in \{-1,1\}^{M\times N}$: label the rows by vectors $\mathbf{e}_1, \dots, \mathbf{e}_M \in \mathbb{R}^M$ and the columns by vectors $\frac{1}{\sqrt{M}}\mathbf{a}_1, \dots, \frac{1}{\sqrt{M}}\mathbf{a}_N$. It is easy to see that this embedding has dimension M and margin $1/\sqrt{M}$. By interchanging the roles of the rows and columns, we see that

$$1 \leqslant \operatorname{dc}(A) \leqslant \min\{M, N\}, \qquad 1 \leqslant \operatorname{mc}(A) \leqslant \min\{\sqrt{M}, \sqrt{N}\}$$

for every matrix $A \in \{-1,1\}^{M\times N}$. We say that a matrix $R \in \mathbb{R}^{M\times N}$ *sign-represents* a matrix $A \in \{-1,1\}^{M\times N}$ if $A_{ij} R_{ij} > 0$ for all i, j. In this case, we symbolically write $A = \operatorname{sign}(R)$. Observe that the dimension complexity of a sign matrix is the minimum rank of any real matrix that sign-represents it.

The *spectral* norm of $R \in \mathbb{R}^{M\times N}$ is defined as

$$\|R\| = \max_{\|x\|=1} \|Rx\|,$$

where the vector norm $\|\cdot\|$ is the Euclidean norm. The *Frobenius* norm of R is defined as

$$\|R\|_F = \sqrt{\sum_{i,j} R_{ij}^2}.$$

For all $R \in \mathbb{R}^{M\times N}$, we have

$$\|R\|_F \geqslant \|R\| = \sqrt{\|RR^{\mathsf{T}}\|} = \sqrt{\|R^{\mathsf{T}}R\|}.$$

A fundamental result, due to Forster, gives a lower bound on the dimension complexity of a matrix in terms of its spectral norm:

THEOREM 2.4 (Forster 2002). *Let $A \in \{-1,1\}^{M\times N}$. Then*

$$\mathrm{dc}(A) \geqslant \frac{\sqrt{MN}}{\|A\|}.$$

Using the random projection technique of Arriaga & Vempala (2006), it is straightforward to show the following relationship between dimension and margin complexity.

PROPOSITION 2.5 (Ben-David *et al.* 2003). *Let $A \in \{-1,1\}^{M\times N}$. If A has an embedding with margin γ (in arbitrarily high dimension), then A has an embedding with margin $\gamma/2$ and dimension $O(\frac{1}{\gamma^2}\log(N+M))$. In particular,* $\mathrm{dc}(A) \leqslant O(\mathrm{mc}(A)^2 \log(N+M))$.

2.4. SQ dimension. A *concept class* $\mathcal{C}$ is any set of Boolean functions $\{-1,1\}^n \to \{-1,1\}$. Let μ be a probability distribution over $\{-1,1\}^n$. The *statistical query (SQ) dimension* of $\mathcal{C}$ under μ, denoted $\mathrm{sqdim}_\mu(\mathcal{C})$, is the largest N for which there are N functions $f_1, \ldots, f_N \in \mathcal{C}$ with

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

for all $i \neq j$. We denote $\mathrm{sqdim}(\mathcal{C}) \stackrel{\text{def}}{=} \max_\mu\{\mathrm{sqdim}_\mu(\mathcal{C})\}$. The SQ dimension of a concept class fully characterizes its weak learnability in the statistical query model: a low SQ dimension implies an efficient weak-learning algorithm, and a high SQ dimension rules out such an algorithm; see Blum *et al.* (1994) and Yang (2005, Cor. 1). A folklore fact is that when the SQ dimension of concept class $\mathcal{C}$ is low, it is possible to select a small number of functions that, collectively, will approximate every function in $\mathcal{C}$. For completeness, we state this observation with a proof.

PROPOSITION 2.6 (SQ dimension and approximation). *Let* $\mathrm{sqdim}_\mu(\mathcal{C}) = N$. *Then there is a set* $\mathcal{H} \subseteq \mathcal{C}$ *with* $|\mathcal{H}| = N$ *such that each* $f \in \mathcal{C}$ *has* $|\mathbf{E}_\mu [f \cdot h]| > 1/(N+1)$ *for some* $h \in \mathcal{H}$.

PROOF. For a set $\mathcal{F} \subseteq \mathcal{C}$, define

$$\gamma(\mathcal{F}) \stackrel{\text{def}}{=} \max_{f_1 \neq f_2 \in \mathcal{F}} \{|\mathbf{E}_\mu [f_1 \cdot f_2]|\},$$

the largest correlation between any two functions in $\mathcal{F}$. Let γ^* be the minimum $\gamma(\mathcal{F})$ over all N-element subsets $\mathcal{F} \subseteq \mathcal{C}$. Let $\mathcal{H}$ be a set of N functions in $\mathcal{C}$ such that $\gamma(\mathcal{H}) = \gamma^*$ and the number of function pairs in $\mathcal{H}$ with correlation γ^* is the smallest possible (over all N-element subsets $\mathcal{F}$ with $\gamma(\mathcal{F}) = \gamma^*$).

We claim that each $f \in \mathcal{C}$ has $|\mathbf{E}_\mu [f \cdot h]| > 1/(N+1)$ for some $h \in \mathcal{H}$. If $f \in \mathcal{H}$, the claim is trivially true. Thus, assume that $f \notin \mathcal{H}$. There are two cases to consider.

$\gamma(\mathcal{H}) \leqslant 1/(N+1)$. Then f must have correlation more than $1/(N+1)$ with some member of $\mathcal{H}$: otherwise we would have $\gamma(\mathcal{H} \cup \{f\}) \leqslant 1/(N+1)$ and $\mathrm{sqdim}_\mu(\mathcal{C}) \geqslant N+1$.

$\gamma(\mathcal{H}) > 1/(N+1)$. Again, f must have correlation more than $1/(N+1)$ with some member of $\mathcal{H}$: otherwise we could improve on the number of function pairs in $\mathcal{H}$ with correlation γ^* by replacing a suitably chosen element of $\mathcal{H}$ with f. □

When analyzing the SQ dimension of a concept class under arbitrary distributions, it is often helpful (Klivans & Sherstov 2007) to consider a modified concept class in order to keep the distribution in the analysis uniform:

PROPOSITION 2.7 (Distribution change by composition). *Let* $\mathcal{C} = \{f_1, \ldots, f_t\}$ *be a concept class of functions* $\{-1,1\}^n \to \{-1,1\}$. *Define a related class* $\mathcal{C}' = \{f_1 \circ g, \ldots, f_t \circ g\}$, *where* $g : \{-1,1\}^m \to \{-1,1\}^n$ *is an arbitrary function for some* m. *Then* $\mathrm{sqdim}(\mathcal{C}) \geqslant \mathrm{sqdim}_U(\mathcal{C}')$, *where* U *denotes the uniform distribution over* $\{-1,1\}^m$.

We omit the simple proof of this fact; see Klivans & Sherstov (2007) for details.

3. SQ dimension of halfspaces

This section establishes an SQ upper bound for halfspaces, which plays a key role in further development.

THEOREM 1.3 (Restated from p. 6). *Fix arbitrary functions* $\phi_1, \ldots, \phi_k : \{-1,1\}^n \to \mathbb{R}$. *Let* $\mathcal{C}$ *be the set of all Boolean functions* f *representable as* $f(x) \equiv \operatorname{sign}(\sum_{i=1}^k a_i \phi_i(x))$ *for some reals* $a_1, \ldots, a_k$. *Then* $\operatorname{sqdim}_\mu(\mathcal{C}) < 2k^2$ *under all distributions* μ.

COROLLARY 1.4 (Restated from p. 6). *Let* $\mathcal{C}$ *be the concept class of halfspaces in* n *dimensions. Then* $\operatorname{sqdim}_\mu(\mathcal{C}) < 2(n+1)^2$ *under all distributions* μ.

PROOF (of Theorem 1.3). We shall use the same technical tool—Forster's work on Euclidean embeddings—as Simon (2006), who proved this claim for μ uniform.

Let μ be an arbitrary distribution. Assume for simplicity that μ is rational (extension to the general case is straightforward). Then the weight $\mu(x)$ of each point x is an integral multiple of $1/M$, where M is a suitably large integer.

Let $N = \operatorname{sqdim}_\mu(\mathcal{C})$. Then there is a set $\mathcal{F} \subseteq \mathcal{C}$ of $|\mathcal{F}| = N$ functions with $|\mathbf{E}_\mu [f \cdot g]| \leqslant 1/N$ for all distinct $f, g \in \mathcal{F}$. Consider the matrix $A \in \{-1,1\}^{N \times M}$ whose rows are indexed by the functions in $\mathcal{F}$, whose columns are indexed by inputs $x \in \{-1,1\}^n$ (an input x indexes exactly $\mu(x) \cdot M$ columns), and whose entries are given by $A = [f(x)]_{f,x}$. By Theorem 2.4,

$$N \leqslant \frac{\operatorname{dc}(A)^2 \, \|A\|^2}{M}. \tag{3.1}$$

We complete the proof by obtaining upper bounds on $\operatorname{dc}(A)$ and $\|A\|$.

We analyze $\operatorname{dc}(A)$ first. Recall that each $f \in \mathcal{F}$ has the form $f(x) = \operatorname{sign}(\sum_{i=1}^k a_{f,i} \phi_i(x))$, where $a_{f,1}, \ldots, a_{f,k}$ are reals specific to f. Therefore,

$$A = [f(x)]_{f,x} = \operatorname{sign}\left(\left[\sum_{i=1}^k a_{f,i} \phi_i(x) \right]_{f,x} \right) = \operatorname{sign}\left(\sum_{i=1}^k [a_{f,i} \phi_i(x)]_{f,x} \right).$$

The last equation shows that A is sign-representable by the sum of k matrices of rank 1, i.e.,

$$\operatorname{dc}(A) \leqslant k. \tag{3.2}$$

We now turn to $\|A\|$. The $N \times N$ matrix AA^{T} is given by

$$AA^{\mathsf{T}} = \Big[M \cdot \mathop{\mathbf{E}}_{\mu} [f \cdot g] \Big]_{f,g}.$$

This means that the diagonal entries of $AA^\intercal$ are equal to M, whereas the off-diagonal entries do not exceed M/N in absolute value. As a consequence,

$$\begin{aligned} \|A\|^2 &= \|AA^\intercal\| \\ &\leqslant \|M \cdot I\| + \|AA^\intercal - M \cdot I\| \\ &\leqslant \|M \cdot I\| + \|AA^\intercal - M \cdot I\|_F \\ &\leqslant M + M\sqrt{\frac{N(N-1)}{N^2}}. \end{aligned} \tag{3.3}$$

Substituting the estimates (3.2) and (3.3) in (3.1) yields the inequality

$$N \leqslant k^2 \left(1 + \sqrt{1 - \frac{1}{N}}\right),$$

which leads to

$$N \leqslant 2k^2 - \frac{1}{4}.$$

This completes the proof for rational μ. To extend the analysis to irrational distributions μ, one considers a rational distribution $\tilde{\mu}$ that approximates μ closely enough and follows the same reasoning. We omit these simple manipulations. □

Remark 3.4. *An easy inspection of the proof of Theorem 1.3 reveals the following stronger result. For a distribution μ, let N be the size of the largest set $\{f_1, \ldots, f_N\} \subseteq \mathcal{C}$ with* average *(not maximum!) pairwise correlation at most $\frac{1}{N}$, i.e., $\frac{1}{N(N-1)} \sum_{i \neq j} (\mathbf{E}_\mu [f_i \cdot f_j])^2 \leqslant \frac{1}{N^2}$. Clearly, N is at least the SQ dimension of $\mathcal{C}$. Theorem 1.3 establishes an upper bound on this larger quantity: $N < 2k^2$.*

We will also need a version of Theorem 1.3 in a slightly different terminology.

Theorem 3.5 (SQ dimension and dimension complexity). *Let $A \in \{-1, 1\}^{M \times N}$ be an arbitrary matrix. View the rows $f_1, \ldots, f_M \in \{-1, 1\}^N$ of A as Boolean functions. Then* $\mathrm{sqdim}(\{f_1, \ldots, f_M\}) < 2\,\mathrm{dc}(A)^2$.

In stating Theorem 3.5, we implicitly extended the notion of the SQ dimension from sets of *Boolean functions* to sets of *vectors* with ± 1 components. This extension is natural since every Boolean function can be viewed as a vector with ± 1 components, and vice versa.

4. A result from communication complexity

To obtain the discrepancy gap and margin-dimension gap (Theorems 1.1 and 1.2) in the next two sections, we recall a result from communication complexity. Consider the Boolean function $\mathrm{GHR} : \{-1,1\}^{4n^2} \times \{-1,1\}^{2n} \to \{-1,1\}$ given by

$$\mathrm{GHR}(x,y) = \operatorname{sign}\left(1 + \sum_{j=0}^{2n-1} y_j \sum_{i=0}^{n-1} 2^i (x_{i,2j} + x_{i,2j+1})\right).$$

This function was defined and studied by Goldmann *et al.* (1992) in the context of separating classes of threshold circuits. Their analysis exhibits a non-product distribution with respect to which $\mathrm{GHR}(x,y)$ has high distributional complexity:

THEOREM 4.1 (Goldmann *et al.* 1992, Thm. 6 and its proof). *There is an (explicitly given) non-product distribution λ such that every deterministic one-way protocol for* GHR *with advantage γ with respect to λ has cost at least* $\log(\gamma\, 2^{n/2}/\sqrt{n}) - O(1)$.

A key consequence of Theorem 4.1 for our purposes is the following result.

LEMMA 4.2 (Discrepancy under non-product distributions). *There is a non-product distribution λ for which* $\mathrm{disc}_\lambda(\mathrm{GHR}) = O(\sqrt{n}/2^{n/2})$.

PROOF. Consider the distribution λ from Theorem 4.1. Let R be a rectangle for which the discrepancy $\mathrm{disc}_\lambda(\mathrm{GHR})$ is achieved:

$$\mathrm{disc}_\lambda(\mathrm{GHR}) = \left| \sum_{(x,y)\in R} \lambda(x,y)\mathrm{GHR}(x,y) \right|.$$

We claim that there is a deterministic one-way protocol for $\mathrm{GHR}(x,y)$ with constant cost and with advantage at least $\mathrm{disc}_\lambda(\mathrm{GHR})$ with respect to λ. Namely, define

$$a = \operatorname{sign}\left(\sum_{(x,y)\in R} \lambda(x,y)\mathrm{GHR}(x,y)\right), \quad b = \operatorname{sign}\left(\sum_{(x,y)\notin R} \lambda(x,y)\mathrm{GHR}(x,y)\right).$$

Consider the protocol P that outputs a if the input is in R, and outputs b otherwise. By definition, the advantage of P with respect to λ is

$$\begin{aligned}
&\sum_{x,y} \lambda(x,y)P(x,y)\mathrm{GHR}(x,y) \\
&\qquad = a \sum_{(x,y)\in R} \lambda(x,y)\mathrm{GHR}(x,y) + b \sum_{(x,y)\notin R} \lambda(x,y)\mathrm{GHR}(x,y) \\
&\qquad \geqslant \left| \sum_{(x,y)\in R} \lambda(x,y)\mathrm{GHR}(x,y) \right| \\
&\qquad = \mathrm{disc}_\lambda(\mathrm{GHR}).
\end{aligned}$$

But by Theorem 4.1, every one-way constant-cost protocol achieves advantage at most $O(\sqrt{n}/2^{n/2})$. Thus, $\mathrm{disc}_\lambda(\mathrm{GHR}) = O(\sqrt{n}/2^{n/2})$. □

5. Discrepancy gap

Lemma 4.2 of the previous section established that $\mathrm{GHR}(x,y)$ has exponentially small discrepancy under a certain *non-product* distribution. Using the SQ upper bound of Section 3, we now prove that the discrepancy of $\mathrm{GHR}(x,y)$ under all *product* distributions is $\Omega(1/n^4)$.

Lemma 5.1 (Product distributions). *Let $\mu = \mu_X \times \mu_Y$ be a product distribution. Then* $\mathrm{disc}_\mu(\mathrm{GHR}) = \Omega(1/n^4)$.

Proof. For each fixed x, denote $\mathrm{GHR}_x(y) = \mathrm{GHR}(x,y)$. Since each GHR_x is a halfspace in the $2n$ variables $y_0, \ldots, y_{2n-1}$, Theorem 1.3 implies that

$$\mathrm{sqdim}_{\mu_Y}(\{\mathrm{GHR}_x\}_x) \leqslant \mathrm{sqdim}_{\mu_Y}(\{\text{halfspaces in } 2n \text{ dimensions}\}) = O(n^2).$$

Thus, by Proposition 2.6, there is a set $\mathcal{H} \subseteq \{\mathrm{GHR}_x\}_x$ of $|\mathcal{H}| = O(n^2)$ functions such that each GHR_x has

$$\left| \mathop{\mathbf{E}}_{y\sim\mu_Y} [\mathrm{GHR}_x(y) \cdot f(y)] \right| > \frac{1}{|\mathcal{H}|+1}$$

for some $f \in \mathcal{H}$.

This yields the following protocol for evaluating $\mathrm{GHR}(x,y)$. Alice, who knows x, sends Bob the index of the function $f \in \mathcal{H}$ whose correlation with GHR_x is the greatest (in absolute value). Alice additionally sends Bob the sign

$\sigma \in \{+1, -1\}$ of the correlation of f and GHR_x. This communication costs $\lceil \log |\mathcal{H}| \rceil + 1$ bits. Bob, who knows y, announces $\sigma \cdot f(y)$ as the output of the protocol.

For every fixed x, the described protocol achieves advantage greater than $1/(|\mathcal{H}|+1)$ over the choice y. As a result, the protocol achieves overall advantage greater than $1/(|\mathcal{H}|+1)$ with respect to any distribution μ_X on the x's. Since only $\lceil \log |\mathcal{H}| \rceil + 2$ bits are exchanged, we obtain the sought bound on the discrepancy by Proposition 2.1:

$$\mathrm{disc}_\mu(\mathrm{GHR}) > \frac{1}{(|\mathcal{H}|+1) \cdot 2^{\lceil \log |\mathcal{H}| \rceil + 2}} = \Omega\left(\frac{1}{n^4}\right). \qquad \square$$

Lemmas 4.2 and 4.2 immediately imply the main result of this section:

THEOREM 1.1 (Restated from p. 3). *There exists an (explicitly given) function $f : \{-1,1\}^n \times \{-1,1\}^{n^2} \to \{-1,1\}$ for which $\mathrm{disc}_\mu(f) = \Omega(1/n^4)$ under every product distribution μ but $\mathrm{disc}_\lambda(f) = O(\sqrt{n}/2^{n/4})$ under a certain non-product distribution λ.*

6. Margin-dimension gap

To exhibit a large gap between margin complexity and dimension complexity, we consider the function $\mathrm{GHR}(x, y)$ from the previous section. We first note that its dimension complexity is low.

PROPOSITION 6.1. *The dimension complexity of $[\mathrm{GHR}(x,y)]_{x,y}$ is at most $2n+1$.*

PROOF. By definition of GHR, the sign matrix $[\mathrm{GHR}(x,y)]_{x,y}$ is sign-represented by the real matrix

$$M = \left[1 + \sum_{j=0}^{2n-1} y_j \sum_{i=0}^{n-1} 2^i (x_{i,2j} + x_{i,2j+1})\right]_{x,y}.$$

It is easy to verify that M has rank at most $2n+1$. $\square$

It remains to show that the margin complexity of $[\mathrm{GHR}(x,y)]_{x,y}$ is high. We do so using the discrepancy estimate for $\mathrm{GHR}(x,y)$. By appealing to Grothendieck's inequality and linear-programming duality, Linial & Shraibman (2006) have recently given a short, elegant proof that the margin complexity and discrepancy of a matrix are equivalent up to a small multiplicative constant:

THEOREM 6.2 (Linial & Shraibman 2006). *For every matrix* $A \in \{-1,1\}^{M\times N}$,

$$\frac{1}{8\,\mathrm{mc}(A)} \leqslant \mathrm{disc}(A) \leqslant \frac{1}{\mathrm{mc}(A)}.$$

Lemma 4.2 and Theorem 6.2 immediately yield an estimate of the margin complexity of $[\mathrm{GHR}(x,y)]_{x,y}$.

LEMMA 6.3. *The margin complexity of* $[\mathrm{GHR}(x,y)]_{x,y}$ *is* $\Omega(2^{n/2}/\sqrt{n})$.

Thus, Linial and Shraibman's subtle result allows us to obtain a particularly good lower bound on the margin complexity. For completeness, we note that a slightly worse bound can be obtained using well-known and more elementary ideas relating margin complexity and discrepancy (Forster *et al.* 2001; Paturi & Simon 1986). Proposition 6.1 and Lemma 6.3 readily imply the main result of this section:

THEOREM 1.2 (Restated from p. 5). *There is an (explicitly given) matrix* $A \in \{-1,1\}^{N\times N^{\log N}}$ *for which* $\mathrm{dc}(A) \leqslant \log N$ *but* $\mathrm{mc}(A) = \Omega(N^{1/4}/\sqrt{\log N})$.

REMARK 6.4. *While Theorem 1.2 exhibits an exponential gap between the dimension complexity and margin complexity, there can be no such gap between the dimension complexity and* average squared margin. *Specifically, a powerful lemma due to Forster (2002) shows that any* k*-dimensional embedding of a given matrix* $A \in \{-1,1\}^{M\times N}$ *can be converted into another* k*-dimensional embedding* $\mathbf{u}_1,\dots,\mathbf{u}_M \in \mathbb{S}^{k-1}$ *and* $\mathbf{v}_1,\dots,\mathbf{v}_N \in \mathbb{S}^{k-1}$ *of* A *that has high average squared margin:* $\frac{1}{MN}\sum_{i,j}\langle \mathbf{u}_i, \mathbf{v}_j\rangle^2 \geqslant \frac{1}{k}$. *(Here* $\mathbb{S}^{k-1}$ *denotes the* k*-dimensional real unit sphere.)*

6.1. Application: circuit complexity of halfspaces. As an application of Lemma 6.3, we study a question from circuit complexity (Goldmann *et al.* 1992). Recall that every halfspace in n variables can be represented as

$$\mathrm{sign}(a_1x_1 + a_2x_2 + \cdots + a_nx_n - \theta),$$

where $a_1, a_2, \dots, a_n, \theta$ are integers called *weights*. A fundamental fact is that there are halfspaces that require weights of magnitude $2^{\Omega(n)}$. This fact can be deduced by an easy counting argument, since there are $2^{\Theta(n^2)}$ distinct halfspaces (Saks 1993). A short and simple $2^{\Omega(n)}$ lower bound for an explicit

halfspace is due to Siu & Bruck (1991). Håstad (1994) improves on that construction, obtaining an explicit halfspace that requires weight $2^{\Theta(n \log n)}$. The $2^{\Theta(n \log n)}$ lower bound is best possible for any halfspace.

Consider now a slightly modified question. Instead of expressing a halfspace as a weighted sum of the singletons $x_1, x_2, \ldots, x_n, 1$, we get to choose an arbitrary set of Boolean functions $f_1, \ldots, f_d : \{-1,1\}^n \to \{-1,1\}$. We would like to know if there is a way to choose $d = \mathsf{poly}(n)$ such functions such that *every* halfspace is expressible as

$$\operatorname{sign}(a_1 f_1(x) + a_2 f_2(x) + \cdots + a_d f_d(x)),$$

where $a_1, a_2, \ldots, a_d$ are integers bounded by a polynomial in n. Unfortunately, the three approaches above do not yield weight lower bounds in this more general setting. Lemma 6.3, on the other hand, yields a simple solution to the problem.

THEOREM 6.5 (Weights of halfspaces over generalized bases). *Let $f_1, \ldots, f_d : \{-1,1\}^n \to \{-1,1\}$ be arbitrary functions. Assume that each halfspace in n dimensions can be expressed as $\operatorname{sign}(\sum_{i=1}^d a_i f_i(x))$, where $a_1, a_2, \ldots, a_d$ are integers bounded in absolute value by w. Then $dw \geqslant \Omega(2^{n/4}/\sqrt{n})$.*

Theorem 6.5 shows a continuous trade-off between the number d of base functions and the magnitude w of the weights. In particular, the weights cannot be polynomially bounded unless there are exponentially many base functions. Goldmann *et al.* (1992, Cor. 9) proved a related result, in which the set of base functions can be arbitrarily large but they must have low randomized communication complexity (e.g., majority gates, modulo gates). Theorem 6.5 complements that result.

PROOF (of Theorem 6.5). It will be convenient to view $d = d(n)$ and $w(n) = w$ as functions of n, and set $D = d(2n)$ and $W = w(2n)$. Fix the Boolean functions $f_1, \ldots, f_D$ satisfying the premise of the theorem. Consider the function

$$\operatorname{GHR}(x,y) = \operatorname{sign}\left(1 + \sum_{j=0}^{2n-1} y_j \sum_{i=0}^{n-1} 2^i (x_{i,2j} + x_{i,2j+1})\right).$$

Since for each fixed x, the function $\operatorname{GHR}_x(y) = \operatorname{GHR}(x,y)$ is a halfspace in the $2n$ variables $y_0, \ldots, y_{2n-1}$, it is representable as a weighted sum of

$f_1(y), \ldots, f_D(y)$ with integer coefficients bounded in absolute value by W. This yields the following embedding of the matrix $[\mathrm{GHR}(x,y)]_{x,y}$:

$$A = \left[\sum_{i=1}^{D} a_i(x) f_i(y)\right]_{x,y},$$

where each $a_i(x)$ is an integer in $[-W, W]$. This embedding has margin

$$\begin{aligned}\gamma &\geqslant \frac{\min\limits_{x,y} |A_{xy}|}{\max\limits_x \left\{\sqrt{\sum_{i=1}^{D} a_i(x)^2}\right\} \cdot \max\limits_y \left\{\sqrt{\sum_{i=1}^{D} f_i(y)^2}\right\}} \\ &\geqslant \frac{1}{\sqrt{\sum_{i=1}^{D} W^2} \cdot \sqrt{\sum_{i=1}^{D} 1^2}} \\ &= \frac{1}{DW}.\end{aligned}$$

At the same time, $\gamma \leqslant O(\sqrt{n}/2^{n/2})$ by Lemma 6.3. Combining these two bounds on γ yields

$$DW = d(2n)w(2n) \geqslant \Omega\left(\frac{2^{n/2}}{\sqrt{n}}\right),$$

and thus $dw = d(n)w(n) \geqslant \Omega(2^{n/4}/\sqrt{n})$. □

7. Complexity measures of sign matrices: An integrated view

The results of the previous sections can be interpreted, in particular, as a study of the complexity measures of sign matrices. Our goal here is to unify them into a coherent picture and better demonstrate how they relate to previous work. We start by establishing one last fact in Section 7.1. A holistic view of this research is then given in Section 7.2.

7.1. SQ Dimension and Discrepancy. The SQ dimension, so far viewed as a complexity measure of *concept classes*, is just as naturally viewed as a complexity measure of *sign matrices*, as follows. Given a matrix $A \in \{-1,1\}^{M \times N}$, view its rows as Boolean functions. We define the *statistical-query complexity* $\mathrm{sq}(A)$ of A as the SQ dimension of its rows. Formally, $\mathrm{sq}(A) \stackrel{\text{def}}{=} \mathrm{sqdim}(\{f_1, \ldots, f_M\})$, where $f_1, \ldots, f_M$ are the rows of A. We prove that the SQ complexity of a matrix is essentially equivalent to the minimum discrepancy of the matrix under product distributions:

THEOREM 7.1 (SQ complexity vs. discrepancy under product distributions). *Let* $A \in \{-1,1\}^{M\times N}$. *Then*

$$\sqrt{\frac{1}{2}\operatorname{sq}(A)} < \frac{1}{\operatorname{disc}^{\times}(A)} < 8\operatorname{sq}(A)^2.$$

Theorem 7.1 establishes a link between learning theory and communication under product distributions. Another such link was discovered earlier by Kremer *et al.* (1999), who showed that the VC dimension of a sign matrix is asymptotically equal to its maximum one-way distributional communication complexity under product distributions. The two results can be compared in a straightforward way: Theorem 7.1 relates small-bias communication to small-bias learning, whereas Kremer *et al.* relate high-accuracy communication to high-accuracy learning. Perhaps surprisingly, the proofs turn out to be quite different in the two cases.

We split the proof of Theorem 7.1 in two parts: the upper bound on $\operatorname{disc}^{\times}(A)$ and the lower bound.

LEMMA 7.2 (Upper bound). *Let* $A \in \{-1,1\}^{M\times N}$. *Then*

$$\operatorname{disc}^{\times}(A) < \sqrt{\frac{2}{\operatorname{sq}(A)}}.$$

PROOF. Assume $\operatorname{sq}(A) = d$. Then there are d rows $f_1, \ldots, f_d \in \{-1,1\}^N$ of A and a distribution μ on $\{1,\ldots,N\}$ such that $|\mathbf{E}_{x\sim\mu}[f_i(x)f_j(x)]| \leqslant 1/d$ for all $i \neq j$. Let U be the uniform distribution over the d rows $f_1, \ldots, f_d$ of A. We prove the lemma by showing that $\operatorname{disc}_{\mu\times U}(A) < \sqrt{2/d}$:

$$\begin{aligned}
\operatorname{disc}_{\mu\times U}(A) &\leqslant \sqrt{\mathbf{E}_{i,j\sim U}\left[\left|\mathbf{E}_{x\sim\mu}[f_i(x)f_j(x)]\right|\right]} && \text{by Lemma 2.3}\\
&\leqslant \sqrt{\frac{1}{d}\cdot 1 + \frac{d-1}{d}\cdot\frac{1}{d}}\\
&< \sqrt{\frac{2}{d}}.
\end{aligned}$$

□

LEMMA 7.3 (Lower bound). *Let* $A \in \{-1,1\}^{M\times N}$. *Then*

$$\operatorname{disc}^{\times}(A) > \frac{1}{8\operatorname{sq}(A)^2}.$$

PROOF. The proof is closely analogous to that of Lemma 5.1; indeed, Lemma 5.1 can be deduced from this lemma. Let $\mu\times\lambda$ be an arbitrary product distribution over $[N]\times[M]$. We will obtain a lower bound on $\operatorname{disc}_{\mu\times\lambda}(A)$ by constructing an efficient protocol for A with a suitable advantage.

Let $\operatorname{sq}(A) = d$. Then by Proposition 2.6, there are d rows $f_1,\dots,f_d \in \{-1,1\}^N$ in A such that each of the remaining rows f has $|\mathbf{E}_{x\sim\mu}[f(x)f_i(x)]| > 1/(d+1)$ for some $i = 1,\dots,d$. This yields the following protocol for evaluating A_{yx}. Bob, who knows the row index y, sends Alice the index i of the function f_i whose correlation with the yth row of A is the greatest (in absolute value). Bob additionally sends Alice the sign $\sigma\in\{+1,-1\}$ of the correlation of f_i and the yth row of A. This communication costs $\lceil\log d\rceil + 1$ bits. Alice, who knows the column index x, announces $\sigma\cdot f_i(x)$ as the output of the protocol.

For every fixed y, the described protocol achieves advantage greater than $1/(d+1)$ over the choice x. As a result, the protocol achieves overall advantage greater than $1/(d+1)$ with respect to any distribution λ on the rows of A. Since only $\lceil\log d\rceil + 2$ bits are exchanged, we obtain the sought bound on the discrepancy by Proposition 2.1:

$$\operatorname{disc}_{\mu\times\lambda}(A) > \frac{1}{(d+1)\cdot 2^{2+\lceil\log d\rceil}} \geqslant \frac{1}{8d^2},$$

where the second inequality holds because d is an integer. □

Lemmas 7.2 and 7.3 immediately yield Theorem 7.1.

Since $\operatorname{disc}^{\times}(A) = \operatorname{disc}^{\times}(A^{\mathsf{T}})$, Theorem 7.1 has the interesting corollary that the rows and columns of a matrix have the same SQ dimension, up to a polynomial factor:

COROLLARY 7.4. *Let* $A \in \{-1,1\}^{M\times N}$. *Then*

$$\left(\frac{\operatorname{sq}(A)}{128}\right)^{1/4} < \operatorname{sq}(A^{\mathsf{T}}) < 128\operatorname{sq}(A)^4.$$

7.2. A unified picture. At this point, we can summarize much of this paper and relevant previous work in the following succinct diagram:

$$\frac{1}{\mathrm{disc}^{\times}(A)} \;=_{\mathsf{poly}}\; \mathrm{sq}(A) \;\leqslant_{\mathsf{poly}}\; \mathrm{dc}(A) \;\leqslant_{\mathsf{poly}}\; \mathrm{mc}(A) \;\approx\; \frac{1}{\mathrm{disc}(A)}$$

exponential gap achievable

The purpose of this schematic is to show that the standard complexity measures $(\mathrm{sq}(A), \mathrm{dc}(A), \mathrm{mc}(A))$ of sign matrices form an ordered spectrum that extends from $\mathrm{disc}^{\times}(A)^{-1}$ to $\mathrm{disc}(A)^{-1}$. In what follows, we let $A \in \{-1,1\}^{M\times N}$ be an arbitrary matrix. We shall traverse the diagram from left to right, giving precise quantitative statements.

- The smallest discrepancy of a matrix under product distributions, $\mathrm{disc}^{\times}(A)$, and the SQ complexity of that matrix, $\mathrm{sq}(A)$, are within a polynomial factor of each other: $\Theta(\sqrt{\mathrm{sq}(A)}) \leqslant \mathrm{disc}^{\times}(A)^{-1} \leqslant \Theta(\mathrm{sq}(A)^2)$. We establish this fact in Theorem 7.1.

- SQ complexity puts a lower bound on dimension complexity: $\mathrm{dc}(A) > \sqrt{\mathrm{sq}(A)/2}$. We prove this relationship in Theorem 3.5.

- Dimension complexity places a lower bound the margin complexity: $\mathrm{mc}(A) \geqslant \Omega(\sqrt{\mathrm{dc}(A)/\log(N+M)})$. This well-known result is easy to prove using random projections (Ben-David *et al.* 2003).

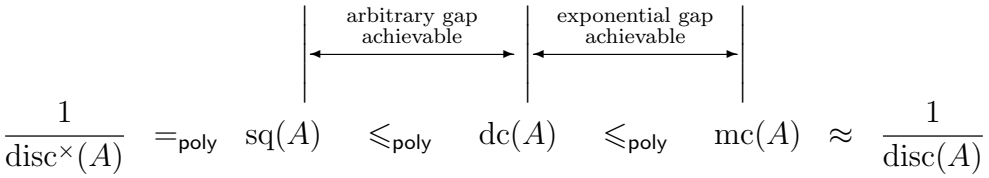
- In Theorem 1.2, we show that the gap between $\mathrm{dc}(A)$ and $\mathrm{mc}(A)$ can be exponentially large. In particular, we exhibit a matrix $A \in \{-1,1\}^{N\times N^{\log N}}$ for which $\mathrm{dc}(A) \leqslant \log N$ but $\mathrm{mc}(A) = \Omega(N^{1/4}/\sqrt{\log N})$.

- Margin complexity is within a multiplicative constant of the discrepancy: $\mathrm{mc}(A) \leqslant \mathrm{disc}(A)^{-1} \leqslant 8\,\mathrm{mc}(A)$. This is a recent result due to Linial & Shraibman (2006).

In summary, this paper refines the current understanding of the complexity measures of sign matrices by proving new relationships among them and analyzing the gaps. A particularly interesting fact is that the standard complexity measures $(\mathrm{sq}(A), \mathrm{dc}(A), \mathrm{sq}(A))$ form an ordered sequence that spans the continuum between product-distribution discrepancy and general discrepancy. This close interplay between linear-algebraic complexity measures ($\mathrm{dc}(A)$ and

$\mathrm{mc}(A)$) and those from communication complexity ($\mathrm{disc}^{\times}(A)$ and $\mathrm{disc}(A)$) is further evidence that the study of sign matrices has much to contribute to complexity theory.

The only piece missing from the above schematic is the gap between $\mathrm{sq}(A)$ and $\mathrm{dc}(A)$. This gap, stated as an open problem in the preliminary version of this work, has been recently settled by the author (Sherstov 2007). Specifically, for any constant $\epsilon > 0$, we have shown the existence of a matrix $A \in \{-1,1\}^{N\times N}$ with $\mathrm{sq}(A) = O(1)$ and $\mathrm{dc}(A) = \Omega(N^{1-\epsilon})$. This gap is essentially the best possible by definition. It completes the above taxonomy to the following picture:



8. Application of the SQ dimension to complexity theory

This final section demonstrates that estimating the SQ dimension of natural classes of Boolean functions is an important task in complexity theory. Specifically, we show that a suitable estimate of the SQ dimension of AC^0 would solve a long-standing problem, that of separating $\mathsf{PH}^{\mathrm{cc}}$ from $\mathsf{PSPACE}^{\mathrm{cc}}$.

These classes in communication complexity were introduced by Babai *et al.* (1986) by analogy with computational complexity. For our purposes, it will be convenient to view $\mathsf{PH}^{\mathrm{cc}}$ and $\mathsf{PSPACE}^{\mathrm{cc}}$ as classes of matrices $M_N \in \{-1,1\}^{N\times N}$ computed by certain circuits rather than by protocols; cf. Razborov (1989). We consider only circuits with $\mathrm{AND}, \mathrm{OR}, \mathrm{NOT}$ gates. The *inputs* to a circuit are arbitrary matrices $A \in \{-1,1\}^{N\times N}$ whose "-1" entries form a combinatorial rectangle; this is equivalent to requiring that $\mathrm{rank}(A - J) \leqslant 1$, where J is the all-ones matrix. The *output* of a circuit is an $N \times N$ sign matrix computed entry-wise from the input matrices (with the usual identification of -1 with "true," and 1 with "false").

Definition 8.1 (Complexity classes $\mathsf{PH}^{\mathrm{cc}}$ and $\mathsf{PSPACE}^{\mathrm{cc}}$). $\mathsf{PH}^{\mathrm{cc}}$ *is the class of all matrix families* $\{M_N\}$ *computable by circuits of size* $\exp((\log\log N)^{O(1)})$ *and constant depth, for some choice of the input matrices.* $\mathsf{PSPACE}^{\mathrm{cc}}$ *is the*

class of all matrix families $\{M_N\}$ *that are computable by circuits of size* $\exp((\log\log N)^{O(1)})$ *and depth* $(\log\log N)^{O(1)}$, *for some choice of the input matrices.*

The relevance of $\mathsf{PH}^{\mathrm{cc}}$ and $\mathsf{PSPACE}^{\mathrm{cc}}$ to communication complexity becomes apparent when one identifies a sign matrix with its corresponding communication problem. See Babai *et al.* (1986) for several other, equivalent definitions of these classes.

It is clear that $\mathsf{PH}^{\mathrm{cc}} \subseteq \mathsf{PSPACE}^{\mathrm{cc}}$, and separating these classes is a major open problem (Lokam 2001; Razborov 1989). Razborov (1989, Rem. 3) argues that "the most natural candidate for $\mathsf{PSPACE}^{\mathrm{cc}} \setminus \mathsf{PH}^{\mathrm{cc}}$ is the INNER PRODUCT MODULO 2 predicate," defined as

$$\mathrm{IP}_N = [(x_1 \wedge y_1) \oplus \cdots \oplus (x_{\log N} \wedge y_{\log N})]_{x,y \in \{-1,1\}^{\log N}}.$$

Indeed, it is easy to see that $\mathrm{IP} \in \mathsf{PSPACE}^{\mathrm{cc}}$. However, neither IP nor any other explicit family of matrices is currently known to be outside $\mathsf{PH}^{\mathrm{cc}}$. We now show that a suitable estimate of the SQ dimension of AC^0 would prove the conjecture that $\mathrm{IP} \notin \mathsf{PH}^{\mathrm{cc}}$.

THEOREM 1.5 (Restated from p. 7). *Let* $\mathcal{C}$ *be the class of functions* $\{-1,1\}^n \to \{-1,1\}$ *computable in* AC^0. *If* $\mathrm{sqdim}(\mathcal{C}) \leqslant O\left(2^{2^{(\log n)^{\epsilon}}}\right)$ *for every constant* $\epsilon > 0$, *then* $\mathrm{IP} \in \mathsf{PSPACE}^{\mathrm{cc}} \setminus \mathsf{PH}^{\mathrm{cc}}$.

PROOF. It will be convenient to prove the contrapositive: if $\mathrm{IP} \in \mathsf{PH}^{\mathrm{cc}}$, then the SQ dimension of AC^0 is at least $2^{2^{(\log n)^{\epsilon}}}$ under some distribution.

We start with an insight due to Lokam (2001), restated in a terminology suitable to our proof. Let C be the assumed constant-depth circuit of size $s = 2^{(\log\log N)^c}$ that computes $\mathrm{IP} \in \{-1,1\}^{N \times N}$. Then C has at most s inputs, which we denote by $A_1, \ldots, A_s \in \{-1,1\}^{N \times N}$. View the rows of the input and output matrices as Boolean functions $\{-1,1\}^{\log N} \to \{-1,1\}$. Since the "$-1$" entries in each $A_1, \ldots, A_s$ form a combinatorial rectangle, each A_i features at most one such row function (call it f_i) that is not identically false. As a result, the rth row of the output matrix IP can be computed as $C_r(f_1(x), \ldots, f_s(x))$, where C_r is a constant-depth circuit of size $s = 2^{(\log\log N)^c}$. See Lokam (2001) for interesting other uses of this observation.

We now return to our proof. Since the rows of IP are mutually orthogonal, the N functions $\{C_r(f_1(x), \ldots, f_s(x))\}_{r=1,\ldots,N}$ that form the rows of the IP matrix are mutually orthogonal under the uniform distribution on x. By Proposition 2.7, this implies that the class of constant-depth, size-s circuits

has SQ dimension at least N. Setting $n \stackrel{\text{def}}{=} 2^{(\log\log N)^c}$, we conclude that the class of constant-depth, size-n circuits (a subclass of AC^0) has SQ dimension at least $2^{2^{(\log n)^{1/c}}}$. □

Acknowledgements

I would like to thank Anna Gál, Adam Klivans, and Sasha Razborov for helpful discussions and feedback on an earlier version of this manuscript. Thanks to Harry Buhrman, Nikolai Vereshchagin, and Ronald de Wolf for useful comments.

References

Noga Alon, Peter Frankl & Vojtech Rödl (1985). Geometrical Realization of Set Systems and Probabilistic Communication Complexity. In *Proc. of the 26th Symposium on Foundations of Computer Science (FOCS)*, 277–280.

Rosa I. Arriaga & Santosh Vempala (2006). An algorithmic theory of learning: Robust concepts and random projection. *Mach. Learn.* **63**(2), 161–182.

László Babai, Peter Frankl & Janos Simon (1986). Complexity classes in communication complexity theory. In *Proc. of the 27th Symposium on Foundations of Computer Science (FOCS)*, 337–347.

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.

Avrim Blum, Alan M. Frieze, Ravi Kannan & Santosh Vempala (1998). A Polynomial-Time Algorithm for Learning Noisy Linear Threshold Functions. *Algorithmica* **22**(1/2), 35–52.

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.

Avrim Blum, Adam Kalai & Hal Wasserman (2003). Noise-tolerant learning, the parity problem, and the statistical query model. *J. ACM* **50**(4), 506–519.

Harry Buhrman, Nikolai K. Vereshchagin & Ronald de Wolf (2007). On Computation and Communication with Small Bias. In *Proc. of the 22nd Conf. on Computational Complexity (CCC)*, 24–32.

Christopher J. C. Burges (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov.* **2**(2), 121–167.

Jeff Ford & Anna Gál (2005). Hadamard Tensors and Lower Bounds on Multiparty Communication Complexity. In *Proc. of the 32nd International Colloquium on Automata, Languages and Programming (ICALP)*, 1163–1175.

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, Niels Schmitt, Hans Ulrich Simon & Thorsten Suttorp (2003). Estimating the Optimal Margins of Embeddings in Euclidean Half Spaces. *Mach. Learn.* **51**(3), 263–281.

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.

Mikael Goldmann, Johan Håstad & Alexander A. Razborov (1992). Majority Gates vs. General Weighted Threshold Gates. *Computational Complexity* **2**, 277–300.

Johan Håstad (1994). On the Size of Weights for Threshold Gates. *SIAM J. Discret. Math.* **7**(3), 484–492.

Bala Kalyanasundaram & Georg Schnitger (1992). The Probabilistic Communication Complexity of Set Intersection. *SIAM J. Discrete Math.* **5**(4), 545–557.

Michael Kearns (1993). Efficient noise-tolerant learning from statistical queries. In *Proc. of the 25th Symposium on Theory of Computing (STOC)*, 392–401.

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 Intersections of Halfspaces with a Margin. In *Proc. of the 17th Conf. on Learning Theory (COLT)*, 348–362.

Adam R. Klivans & Alexander A. Sherstov (2006). Cryptographic Hardness for Learning Intersections of Halfspaces. In *Proc. of the 47th Symposium on Foundations of Computer Science (FOCS)*, 553–562.

Adam R. Klivans & Alexander A. Sherstov (2007). Unconditional lower bounds for learning intersections of halfspaces. *Machine Learning* **69**(2–3), 97–114.

Ilan Kremer, Noam Nisan & Dana Ron (1999). On Randomized One-Round Communication Complexity. *Computational Complexity* **8**(1), 21–49.

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

Stephen Kwek & Leonard Pitt (1998). PAC Learning Intersections of Halfspaces with Membership Queries. *Algorithmica* **22**(1/2), 53–75.

N. Linial, S. Mendelson, G. Schechtman & A. Shraibman (2006). Complexity Measures of Sign Matrices. *Combinatorica* To appear. Manuscript at `http://www.cs.huji.ac.il/~nati/PAPERS/complexity_matrices.ps.gz`.

Nati Linial & Adi Shraibman (2006). Learning Complexity vs. Communication Complexity. Manuscript at `http://www.cs.huji.ac.il/~nati/PAPERS/lcc.pdf`.

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.

Ramamohan Paturi & Janos Simon (1986). Probabilistic communication complexity. *J. Comput. Syst. Sci.* **33**(1), 106–123.

Ran Raz (2000). The BNS-Chung criterion for multi-party communication complexity. *Comput. Complex.* **9**(2), 113–122.

Alexander A. Razborov (1989). Ob ustoichivyh matritsah. Research report, Steklov Mathematical Institute, Moscow, Russia. In Russian. *Engl. title:* "On rigid matrices".

Alexander A. Razborov (1992). On the distributional complexity of disjointness. *Theor. Comput. Sci.* **106**(2), 385–390.

Michael E. Saks (1993). Slicing the hypercube. *Surveys in combinatorics, 1993* 211–255.

Alexander A. Sherstov (2007). Communication Complexity under Product and Nonproduct Distributions. In *Electronic Colloquium on Computational Complexity (ECCC)*. Report TR07-072.

Hans Ulrich Simon (2006). Spectral Norm in Learning Theory: Some Selected Topics. In *Proc. of the 17th Conf. on Algorithmic Learning Theory (ALT)*, 13–27.

Kai-Yeung Siu & Jehoshua Bruck (1991). On the Power of Threshold Circuits with Small Weights. *SIAM J. Discrete Math.* **4**(3), 423–435.

Nathan Srebro & Adi Shraibman (2005). Rank, Trace-Norm and Max-Norm. In *Proc. of the 18th Conf. on Learning Theory (COLT)*, 545–560.

Leslie G. Valiant (1984). A theory of the learnable. *Commun. ACM* **27**(11), 1134–1142.

Vladimir N. Vapnik (1995). *The nature of statistical learning theory.* Springer-Verlag, New York.

Santosh Vempala (1997). A Random Sampling based Algorithm for Learning the Intersection of Halfspaces. In *Proc. of the 38th Symposium on Foundations of Computer Science (FOCS)*, 508–513.

Ke Yang (2005). New lower bounds for statistical query learning. *J. Comput. Syst. Sci.* **70**(4), 485–509.

Manuscript received 28 August 2007

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