Halfspace Matrices

Alexander A. Sherstov

The University of Texas at Austin
Department of Computer Sciences
Austin, TX 78712 USA
sherstov@cs.utexas.edu

Abstract

A *halfspace matrix* is a Boolean 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 demonstrate the potential of halfspace matrices as tools to answer nontrivial open questions.

1. *(Communication complexity)* We exhibit a Boolean function f with discrepancy $\Omega(1/n^4)$ under all product distributions but $O(\sqrt{n}/2^{n/4})$ under a certain non-product distribution. This essentially solves an open problem of Kushilevitz and Nisan [23].

2. *(Complexity of sign matrices)* 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 the results of Forster et al. [12] and Srebro and Shraibman [35] and is close to optimal. As an application to circuit complexity, we prove an $\Omega(2^{n/4}/(d\sqrt{n}))$ circuit lower bound for computing halfspaces by a majority of an arbitrary set of d gates. This complements a result of Goldmann, Håstad, and Razborov [14]. In addition, we prove new results on the complexity measures of sign matrices, complementing a recent study by Linial et al. [25–27].

3. *(Learning theory)* We prove that the statistical-query (SQ) dimension of halfspaces in n dimensions is less than $2(n+1)^2$ under all distributions (with $n+1$ being a trivial lower bound). This improves on the fundamental result of Blum et al. [5]. We are able to give a simple, one-page proof of our result that contrasts with the sophisticated argument of Blum et al.

Finally, we motivate our learning-theoretic result for the complexity community by showing that SQ dimension estimates for natural classes of Boolean functions can resolve major open problems in complexity theory. Specifically, we show that an $\exp(2^{(\log n)^{o(1)}})$ upper bound on the SQ dimension of AC^0 would imply an explicit language in $\mathsf{PSPACE}^{\mathsf{cc}} \setminus \mathsf{PH}^{\mathsf{cc}}$. Viewed differently, this explains the apparent lack of progress on designing distribution-free learning algorithms for AC^0 with a nontrivial running time.

1 Introduction

A *halfspace* is Boolean function f representable as $f(x) = \text{sign}(\sum_{i=1}^n a_i x_i - \theta)$ 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 explore the potential of halfspace matrices in the study of complexity. Specifically, we demonstrate how halfspace matrices can 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 [10] on the sign-representation of Boolean matrices by real ones. Forster's work has had exciting applications, including a linear lower bound [10] on communication complexity in the unbounded-error model, extremal results [10, 12, 13] on Euclidean embeddings, and lower bounds [11] for depth-2 threshold circuits. 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 of communication complexity is the *randomized model* [23, 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 randomization. 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* $\mathrm{R}(f)$ of a function f is cost of the best protocol for f.

The standard approach to proving lower bounds on $\mathrm{R}(f)$ 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 $1/3$ over μ must be high. It can be shown that $\mathrm{R}(f) = \max_{\mu} \mathrm{D}^{\mu}_{1/3}(f)$. 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 $\mathrm{D}^{\mu}_{1/3}(\mathrm{DISJ}) = O(\sqrt{n}\log n)$ (see [23]), although $\mathrm{R}(\mathrm{DISJ}) = \Theta(n)$ (see [16, 32]). This motivates the following intriguing question in communication complexity:

Open Problem (Kushilevitz and Nisan [23, page 37]). Can restricting the distribution μ to be a product distribution affect the resulting lower bound on $\mathrm{R}(f)$ by more than a polynomial factor? Formally, is $\mathrm{R}(f) = (\max_{\mu:\text{product}} \mathrm{D}^{\mu}_{1/3}(f))^{O(1)}$?

Since its formulation 10 years ago, this problem has seen no progress. This motivates us to look at a related question. The chief source of lower bounds on distributional complexity $\mathrm{D}^{\mu}_{1/3}(f)$ and thus randomized complexity $\mathrm{R}(f)$ 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 (explicit) function $f : \{-1,1\}^n \times \{-1,1\}^{n^2} \to \{-1,1\}$ for which $\mathrm{disc}_{\mu}(f) = \Omega(1/n^4)$ under all product distributions 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 $\mathrm{R}(f)$ if restricted to product distributions. This is the first nontrivial discrepancy gap obtained for any function.

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. [25]. 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 [10, 11].

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 [4, 10–13, 25]. In addition to their roles as complexity measures, dimension complexity is a key player in the unbounded-error model of communication complexity [1, 29], and margin complexity is the central notion in the highly successful *kernel methods* [8, 37] of machine learning.

Using the random projection technique of Arriaga and Vempala [2], it is straightforward to show [4] 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 and Servedio [20] 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 [22]. Unfortunately, we give a strong negative answer to this question.

This problem of estimating the gap between dimension and margin complexity has been studied by several researchers. Forster et al. [12] 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 and Shraibman [35] amplified this separation, obtaining $\mathrm{dc}(A) \leqslant 2^p$ and $\mathrm{mc}(A) = (\log N)^{\Theta(p)}$ for any choice of the parameter $1 \leqslant p \leqslant (\log N)^{1-\varepsilon}$. Our second main theorem is an exponential improvement over these results.

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

The exponential separation we obtain above 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 [14] from circuit complexity. Fix arbitrary d 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 [15, 34] that some halfspaces require exponentially large weights, and complements a result due to Goldmann, Håstad, and Razborov [14]. 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 [17]. The SQ model is a restricted version of the standard PAC learning model [36]. Fix a set $\mathscr{C}$ of Boolean functions $\{-1,1\}^n \to \{-1,1\}$ (a *concept class*) and distribution μ over $\{-1,1\}^n$. For each choice of an unknown function $f \in \mathscr{C}$, 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 [6, 7, 18, 22, 39]. In particular, essentially all known PAC learning algorithms can be adapted [18] 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 $\mathscr{C}$ under a given distribution μ is its *statistical query (SQ) dimension*, $\mathrm{sqdim}_\mu(\mathscr{C})$. This complexity measure is essentially *tight*: low SQ dimension implies efficient weak learnability in the SQ model, and high SQ dimension rules it out. Informally, $\mathrm{sqdim}_\mu(\mathscr{C})$ is the size of the largest subset $\mathscr{F} \subseteq \mathscr{C}$ of (almost) mutually orthogonal functions in $\mathscr{C}$ under μ. We put $\mathrm{sqdim}(\mathscr{C}) \rightleftharpoons \max_\mu \mathrm{sqdim}_\mu(\mathscr{C})$. 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 [19–22, 24, 38] 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. [5] gave a polynomial-time algorithm for learning halfspaces in the SQ model under arbitrary distributions. It follows from the work of Blum et al. 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 n arbitrary functions $\phi_1, \ldots, \phi_n : \{-1,1\}^n \to \mathbb{R}$. Let $\mathscr{C}$ be the set of all Boolean functions f representable as $f = \mathrm{sign}(\sum_{i=1}^n a_i \phi_i(x))$ for some reals $a_1, \ldots, a_n$. Then $\mathrm{sqdim}_\mu(\mathscr{C}) < 2n^2$ under all distributions μ.*

Corollary 1.3.1 (SQ dimension of halfspaces). *Let $\mathscr{C}$ be the concept class of halfspaces in n dimensions. Then $\mathrm{sqdim}_\mu(\mathscr{C}) < 2(n+1)^2$ under all distributions μ.*

Since $\mathrm{sqdim}(\text{halfspaces}) \geqslant n+1$ even under the uniform distribution, the quadratic upper bound of Corollary 1.3.1 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 [10]. By contrast, the proof of Blum et al. relies on nontrivial notions from computational geometry and requires a lengthy analysis of robustness under noise. We believe that Theorem 1.3 is additionally valuable in that it brings a novel set of techniques (complexity of sign matrices) to bear on a central problem of computational learning.

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

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 1989 manuscript by Razborov [31]. Viewed from a different standpoint, Theorem 1.4 explains the lack of progress in designing learning algorithms for AC^0: a distribution-free 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 1.4 grows faster than any quasipolynomial function but slower than any subexponential one. 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 theorem [10], which we show constrains the sign pattens of halfspace matrices in an important way. These structural constraints allow us prove an upper bound on the SQ dimension of halfspaces (Theorem 1.3). Another technical tool we use is a result of Goldmann, Håstad, and Razborov [14] in communication complexity. We apply it to show that halfspace matrices possess considerable structural complexity. It is this contrast between Forster's result and that of Goldmann, Håstad, and Razborov that allows us to obtain 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 [10–12, 25] for explicit matrices are 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 1.4 regarding $\mathsf{PSPACE}^{\mathrm{cc}} \setminus \mathsf{PH}^{\mathrm{cc}}$ builds on the original manuscript of Razborov [31] and a combinatorial observation due to Lokam [28].

The rest of the paper is organized as follows. After the technical preliminaries, we first prove the SQ dimension upper bound (Theorem 1.3) and then use it to establish the discrepancy and margin-dimension gaps (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 1.4).

2 Preliminaries

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 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* $\mathrm{R}_{1/2-\gamma/2}(f)$ 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 $\mu(x,y) = \mu_X(x)\mu_Y(y)$ 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 and Nisan [23, pp. 36–38]). *For every Boolean function $f(x,y)$, every distribution μ, and every $\gamma > 0$,*

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

A definitive resource for further details is the book of Kushilevitz and Nisan [23].

2.2 Sign matrices

We frequently use "generic-entry" notation to specify a matrix succinctly: we write $A = [F(i,j)]_{i,j}$ to mean that 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 $\mathbf{u}_1, \ldots, \mathbf{u}_M \in \mathbb{R}^k$ 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

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

is the *margin* of the embedding. The *dimension complexity* $\mathrm{dc}(A)$ 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.

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 $A = [\mathbf{a}_1 \mid \ldots \mid \mathbf{a}_N] \in \{-1,1\}^{M\times N}$: label the rows by vectors $\mathbf{e}_1, \ldots, \mathbf{e}_M \in \mathbb{R}^M$ and the columns by vectors $\frac{1}{\sqrt{M}}\mathbf{a}_1, \ldots, \frac{1}{\sqrt{M}}\mathbf{a}_N$. 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

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

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. We symbolically write $A = \mathrm{sign}(R)$. 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 $\|R\| = \max_{\|x\|=1} \|Rx\|$. 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

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

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

Theorem 2.2 (Forster [10]). *Let* $A \in \{-1,1\}^{M\times N}$. *Then* $\mathrm{dc}(A) \geqslant \sqrt{MN}/\|A\|$.

Using the random projection technique of Arriaga and Vempala [2], it is straightforward to show the following relationship between dimension and margin complexity.

Proposition 2.3 (Ben-David, Eiron, and Simon [4]). *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,* $\mathrm{dc}(A) \leqslant O(\mathrm{mc}(A)^2\log(N+M))$.

2.3 SQ dimension

A *concept class* $\mathscr{C}$ is a 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 $\mathscr{C}$ under μ, denoted $\mathrm{sqdim}_\mu(\mathscr{C})$, is the largest N for which there are N functions $f_1,\ldots,f_N \in \mathscr{C}$ 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 $\mathrm{sqdim}(\mathscr{C}) \rightleftharpoons \max_\mu\{\mathrm{sqdim}_\mu(\mathscr{C})\}$. 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. The following two theorems make these statements precise.

Theorem 2.4. [6] (Upper Bound) *Let* $\mathscr{C}$ *be a concept class and* μ *a distribution s.t.* $\mathrm{sqdim}_\mu(\mathscr{C}) = N$. *Then there is a non-uniform learning algorithm for* $\mathscr{C}$ *that makes* N *queries, each of tolerance* $1/(3N^3)$*, and finds a hypothesis with error at most* $1/2 - 1/(3N^3)$ *under* μ.

Theorem 2.5. [39, Corollary 1] (Lower Bound) *Let* $\mathscr{C}$ *be a concept class and* μ *a distribution s.t.* $\mathrm{sqdim}_\mu(\mathscr{C}) = N$. *Then if all queries are made with tolerance at least* $1/N^{1/3}$*, at least* $N^{1/3}/2 - 1$ *queries are required to learn* $\mathscr{C}$ *to error* $1/2 - 1/(2N^{1/3})$ *under* μ *in the statistical query model.*

We will need the following folklore fact about the SQ dimension.

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

In words, Proposition 2.6 says that when the SQ dimension of $\mathscr{C}$ is low, it is possible to select a small number of functions that, collectively, will approximate every function in $\mathscr{C}$. See Appendix A for a proof.

When analyzing the SQ dimension of a concept class under arbitrary distributions, it is often helpful (e.g., Klivans and Sherstov [22]) to consider a modified concept class in order to keep the distribution in the analysis uniform:

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

We omit the simple proof of this fact; see [22] 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. 3). *Fix n arbitrary functions $\phi_1, \ldots, \phi_n : \{-1,1\}^n \to \mathbb{R}$. Let $\mathscr{C}$ be the set of all Boolean functions f representable as $f = \operatorname{sign}(\sum_{i=1}^n a_i \phi_i(x))$ for some reals $a_1, \ldots, a_n$. Then $\operatorname{sqdim}_\mu(\mathscr{C}) < 2n^2$ under all distributions μ.*

Corollary 1.3.1 (Restated from p. 3). *Let $\mathscr{C}$ be the concept class of halfspaces in n dimensions. Then $\operatorname{sqdim}_\mu(\mathscr{C}) < 2(n+1)^2$ under all distributions μ.*

Proof of Theorem 1.3. Let μ 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 $N = \operatorname{sqdim}_\mu(\mathscr{C})$. Then there is a set $\mathscr{F} \subseteq \mathscr{C}$ of $|\mathscr{F}| = N$ functions with $|\mathbf{E}_\mu[f \cdot g]| \leqslant 1/N$ for all distinct $f, g \in \mathscr{F}$. Consider the matrix $A \in \{-1,1\}^{N \times M}$ whose rows are indexed by the functions in $\mathscr{F}$, 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 2.2,

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

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

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

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

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

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

We now turn to $\|A\|$. The entries of the $N \times N$ matrix AA^t are given by $AA^t = [M \cdot \mathbf{E}_\mu[f \cdot g]]_{f,g}$. Thus,

$$\|A\|^2 = \|AA^t\| \leqslant \|M \cdot I\| + \|AA^t - M \cdot I\| \leqslant \|M \cdot I\| + \|AA^t - M \cdot I\|_F \leqslant M + M\sqrt{\frac{N(N-1)}{N^2}} < 2M.$$

We have shown that

$$\|A\| < \sqrt{2M}. \tag{3.3}$$

Substituting the estimates (3.2) and (3.3) into (3.1) completes the proof. To extend the analysis to irrational distributions μ, one considers a rational distribution μ' that approximates μ closely enough and follows the same reasoning. We omit these simple manipulations. □

Remark 3.1. 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 $\{f_1, \ldots, f_N\} \subseteq \mathscr{C}$ with *average* (not maximum!) pairwise correlations 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 $\mathscr{C}$. Theorem 1.3 establishes an upper bound on this larger quantity: $N < 2n^2$.

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

Theorem 3.2 (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 3.2, we implicitly extended the notion of the SQ dimension of sets of *Boolean functions* to sets of *arbitrary 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 and margin-dimension gaps (Theorems 1.1 and 1.2) in the next two sections, we recall a result from communication complexity. Consider the Boolean function $\mathrm{GHR} : \{-1,1\}^{4n^2} \times \{-1,1\}^{2n} \to \{-1,1\}$, defined as

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

This function was constructed and studied by Goldmann, Håstad, and Razborov [14] 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, Håstad, and Razborov [14, Theorem 6 and its proof]). *There is an (explicit) non-product distribution λ such that any 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 the combinatorial rectangle over which the discrepancy $\mathrm{disc}_\lambda(\mathrm{GHR})$ is achieved:

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

Then there is a deterministic one-way protocol for $\mathrm{GHR}(x,y)$ with advantage at least $\mathrm{disc}_\lambda(\mathrm{GHR})$ and constant cost. Namely, if $(x,y) \in R$, the players output $\mathrm{sign}(\sum_{(x,y)\in R} \lambda(x,y)\mathrm{GHR}(x,y))$. If $(x,y) \notin R$, the players analogously output $\mathrm{sign}(\sum_{(x,y)\notin R} \lambda(x,y)\mathrm{GHR}(x,y))$. 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 $\mathscr{H} \subseteq \{\mathrm{GHR}_x\}_x$ of $|\mathscr{H}| = O(n^2)$ functions such that each GHR_x has

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

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

This yields the following protocol for evaluating $\mathrm{GHR}(x,y)$. Alice, who knows x, sends Bob the index of the function $f \in \mathscr{H}$ that is best correlated with GHR_x. This costs $\log |\mathscr{H}|$ bits. Bob, who knows y, announces $f(y)$ as the output of the protocol. For every fixed x, this protocol achieves advantage $1/(|\mathscr{H}|+1)$ over the choice y. As a result, the protocol achieves overall advantage $1/(|\mathscr{H}|+1)$ with respect to any distribution μ_X on the x's. Since only $1 + \log|\mathscr{H}|$ bits are exchanged, we obtain the sought bound on the discrepancy by Proposition 2.1:

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

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

Theorem 1.1 (Restated from p. 1). *There exists an (explicit) function $f : \{-1,1\}^n \times \{-1,1\}^{n^2} \to \{-1,1\}$ for which $\mathrm{disc}_\mu(f) = \Omega(1/n^4)$ under all product distributions 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 and Shraibman [26] 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 and Shraibman [26]). *For every matrix $A \in \{-1,1\}^{M \times N}$,*

$$\frac{1}{4K_G \cdot \mathrm{mc}(A)} \leqslant \mathrm{disc}(A) \leqslant \frac{1}{\mathrm{mc}(A)},$$

where $K_G \in [1.67, 1.79]$ is the Grothendieck constant.

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 facts relating margin complexity and discrepancy (see, e.g., Paturi and Simon [29], Forster et al. [11]). Proposition 6.1 and Lemma 6.3 readily imply the main result of this section:

Theorem 1.2 (Restated from p. 2). *There is an (explicit) 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 dc and margin complexity mc, there can be no such gap between the dimension complexity dc and "average margin." Specifically, a powerful lemma due to Forster [10] 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 $\mathbf{u}_1, \ldots, \mathbf{u}_M \in \mathbb{S}^{k-1}$ and $\mathbf{v}_1, \ldots, \mathbf{v}_N \in \mathbb{S}^{k-1}$ of A that has high *average* 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 our margin analysis in Lemma 6.3, we study a question [14] from circuit complexity. Recall that every halfspace on n variables can be represented as

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

where $a_1, a_2, \ldots, a_n, \theta$ 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 [33]. A short and simple $2^{\Omega(n)}$ lower bound for an explicit halfspace is due to Siu and Bruck [34]. Håstad [15] 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 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

$$\mathrm{sign}(a_1f_1(x) + a_2f_2(x) + \cdots + a_df_d(x)),$$

where $a_1, a_2, \ldots, a_d$ are integers bounded by a polynomial in n. Unfortunately, the three above approaches do not yield weight lower bounds for 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* $\mathrm{sign}(\sum_{i=1}^d a_i f_i(x))$, *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, Håstad, and Razborov [14, Corollary 9] proved a related result, in which the set of the base functions can be arbitrarily large but they must have low randomized communication complexity (e.g., majority gates, mod 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

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

Since for each fixed x, the function $\mathrm{GHR}_x(y) = \mathrm{GHR}(x,y)$ 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 coefficients bounded 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 $|a_i(x)| \leqslant W$. The margin of this embedding is

$$\gamma \geqslant \frac{1}{\sqrt{\sum_{i=1}^{D} W^2} \cdot \sqrt{\sum_{i=1}^{D} 1^2}} = \frac{1}{DW}.$$

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 in this section is to unify them into a coherent picture and better demonstrate how they relate to previous work.

We start with the observation that the SQ dimension, so far viewed as a property of *concept classes*, is just as naturally viewed as a property of *sign matrices*. 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, $\mathrm{sq}(A) = \mathrm{sqdim}(\{f_1, \ldots, f_M\})$, 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 $\sqrt{\frac{1}{2}\mathrm{sq}(A)} < \mathrm{disc}^{\times}(A)^{-1} < (2\,\mathrm{sq}(A))^2$.*

Since $\mathrm{disc}^{\times}(A) = \mathrm{disc}^{\times}(A^t)$, 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 7.1.1. *Let $A \in \{-1,1\}^{M \times N}$. Then $(\frac{1}{32}\mathrm{sq}(A))^{1/4} < \mathrm{sq}(A^t) < 32\,\mathrm{sq}(A)^4$.*

We defer the proof of Theorem 7.1 to Appendix C. At this point, we can summarize much of this paper and relevant previous work in the following succinct diagram:

$$\mathrm{disc}^{\times}(A)^{-1} =_{\mathsf{poly}} \mathrm{sq}(A) \underbrace{\qquad \leqslant_{\mathrm{poly}} \qquad}_{\text{unknown gap}} \mathrm{dc}(A) \underbrace{\qquad \leqslant_{\mathrm{poly}} \qquad}_{\text{exponential gap achievable}} \mathrm{mc}(A) \approx \mathrm{disc}(A)^{-1}$$

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 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 3.2.
- It is unknown how large the gap between $\mathrm{sq}(A)$ and $\mathrm{dc}(A)$ can be.
- Dimension complexity lower-bounds margin complexity: $\mathrm{mc}(A) \geqslant \Omega(\sqrt{\mathrm{dc}(A)/\log(N+M)})$. This well-known result is easily proved using random projections (see, e.g., Ben-David, Eiron, and Simon [4]).
- 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: $\mathrm{mc}(A) \leqslant \mathrm{disc}(A)^{-1} \leqslant 4K_G \cdot \mathrm{mc}(A)$, where $K_G \in [1.67, 1.79]$ is the Grothendieck constant. This is a recent result due to Linial and Shraibman [26].

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.

We conclude this section by proposing an approach to separating $\mathrm{sq}(A)$ and $\mathrm{dc}(A)$. For p prime, consider the n-dimensional vector space $\mathbb{F}_p^n$. Consider the symmetric matrix A of size $\frac{p^n-1}{p-1} \times \frac{p^n-1}{p-1}$ whose rows and columns are indexed by nonzero one-dimensional subspaces of $\mathbb{F}_p^n$ and whose entries are given by

$$A_{S,T} = \begin{cases} 1 & \text{if } S \text{ and } T \text{ are orthogonal,} \\ -1 & \text{otherwise.} \end{cases}$$

This matrix is known as a *projective space* matrix and is a straightforward generalization of the inner-product-mod-2 matrix IP to higher moduli p. Forster et al. [11] have shown that A has exponentially high dimension complexity: $\mathrm{dc}(A) \geqslant p^{n/2-1}(1-o(1))$. At the same time, it seems plausible that $\mathrm{disc}^{\times}(A)^{-1}$ (and thus $\mathrm{sq}(A)$) is small: it is unclear how to construct a product distribution under which A would have low discrepancy. In this light, A is a promising candidate for separating $\mathrm{sq}(A)$ and $\mathrm{dc}(A)$.

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, Frankl, and Simon [3] as analogues of the polynomial hierarchy PH and polynomial space PSPACE in computational complexity. For our purposes, it will be more convenient to view $\mathsf{PH}^{\mathrm{cc}}$ and $\mathsf{PSPACE}^{\mathrm{cc}}$ as classes of $N \times N$ matrices computed by certain circuits rather than by protocols. 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 $\operatorname{rank}(A - J) \leqslant 1$, where J is the all-ones matrix. The *output* of a circuit is a matrix $\{-1,1\}^{N \times N}$ computed entry-wise from the inputs.

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 $\{A_N\}$ that are 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 $\{A_N\}$ 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.

It is clear that $\mathsf{PH}^{\mathrm{cc}} \subseteq \mathsf{PSPACE}^{\mathrm{cc}}$, and separating these classes is a major open problem [28, 31]. Razborov [31, Remark 3] argues that "the most natural candidate for $\mathsf{PSPACE}^{\mathrm{cc}} \setminus \mathsf{PH}^{\mathrm{cc}}$ is the INNER PRODUCT MOD 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 1.4 (Restated from p. 3). *Let $\mathscr{C}$ be the class of functions $\{-1,1\}^n \to \{-1,1\}$ computable in AC^0. If $\operatorname{sqdim}(\mathscr{C}) \leqslant O\left(2^{2^{(\log n)^{\varepsilon}}}\right)$ for every constant $\varepsilon > 0$, 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 $2^{2^{(\log n)^{\varepsilon}}}$ under some distribution.

We start with an insight due to Lokam [28], 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 [28] 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 $n \rightleftharpoons 2^{(\log\log N)^c}$, we conclude that the class of constant-depth, size-n circuits (a subclass of AC^0) has SQ dimension at least $2^{2^{(\log n)^{1/c}}}$. □

Acknowledgments

I would like to thank Anna Gál, Adam Klivans, and Sasha Razborov for helpful discussions and feedback on an earlier version of this manuscript.

References

[1] N. Alon, P. Frankl, and V. Rödl. Geometrical realization of set systems and probabilistic communication complexity. In *FOCS*, pages 277–280, 1985.

[2] R. I. Arriaga and S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *FOCS '99: Proceedings of the 40th Annual Symposium on Foundations of Computer Science*, page 616, Washington, DC, USA, 1999. IEEE Computer Society.

[3] L. Babai, P. Frankl, and J. Simon. Complexity classes in communication complexity theory. In *FOCS*, pages 337–347, 1986.

[4] S. Ben-David, N. Eiron, and H. U. Simon. Limitations of learning via embeddings in Euclidean half spaces. *J. Mach. Learn. Res.*, 3:441–461, 2003.

[5] A. Blum, A. M. Frieze, R. Kannan, and S. Vempala. A polynomial-time algorithm for learning noisy linear threshold functions. *Algorithmica*, 22(1/2):35–52, 1998.

[6] A. Blum, M. Furst, J. Jackson, M. Kearns, Y. Mansour, and S. Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *STOC '94: Proceedings of the twenty-sixth annual ACM symposium on Theory of computing*, pages 253–262, New York, NY, USA, 1994. ACM Press.

[7] A. Blum, A. Kalai, and H. Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *J. ACM*, 50(4):506–519, 2003.

[8] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, 1998.

[9] J. Ford and A. Gál. Hadamard tensors and lower bounds on multiparty communication complexity. In *ICALP*, pages 1163–1175, 2005.

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

[11] J. Forster, M. Krause, S. V. Lokam, R. Mubarakzjanov, N. Schmitt, and H.-U. Simon. Relations between communication complexity, linear arrangements, and computational complexity. In *FST TCS '01: Proceedings of the 21st Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 171–182, London, UK, 2001. Springer-Verlag.

[12] J. Forster, N. Schmitt, H. U. Simon, and T. Suttorp. Estimating the optimal margins of embeddings in euclidean half spaces. *Mach. Learn.*, 51(3):263–281, 2003.

[13] J. Forster and H. U. Simon. On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes. *Theor. Comput. Sci.*, 350(1):40–48, 2006.

[14] M. Goldmann, J. Håstad, and A. A. Razborov. Majority gates vs. general weighted threshold gates. *Computational Complexity*, 2:277–300, 1992.

[15] J. Håstad. On the size of weights for threshold gates. *SIAM J. Discret. Math.*, 7(3):484–492, 1994.

[16] B. Kalyanasundaram and G. Schintger. The probabilistic communication complexity of set intersection. *SIAM J. Discret. Math.*, 5(4):545–557, 1992.

[17] M. Kearns. Efficient noise-tolerant learning from statistical queries. In *STOC '93: Proceedings of the twenty-fifth annual ACM symposium on theory of computing*, pages 392–401, New York, NY, USA, 1993. ACM Press.

[18] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, MA, USA, 1994.

[19] A. R. Klivans, R. O'Donnell, and R. A. Servedio. Learning intersections and thresholds of halfspaces. *J. Comput. Syst. Sci.*, 68(4):808–840, 2004.

[20] A. R. Klivans and R. A. Servedio. Learning intersections of halfspaces with a margin. In *COLT*, pages 348–362, 2004.

[21] A. R. Klivans and A. A. Sherstov. Cryptographic hardness for learning intersections of halfspaces. In *FOCS '06: Proceedings of the 47th Annual Symposium on Foundations of Computer Science*, Berkeley, CA, October 2006.

[22] A. R. Klivans and A. A. Sherstov. Improved lower bounds for learning intersections of halfspaces. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT)*, Pittsburg, USA, June 2006.

[23] E. Kushilevitz and N. Nisan. *Communication complexity*. Cambridge University Press, New York, NY, USA, 1997.

[24] S. Kwek and L. Pitt. PAC learning intersections of halfspaces with membership queries. *Algorithmica*, 22(1/2):53–75, 1998.

[25] N. Linial, S. Mendelson, G. Schechtman, and A. Shraibman. Complexity measures of sign matrices. *Combinatorica*, 2006. To appear. Manuscript at `http://www.cs.huji.ac.il/~nati/PAPERS/complexity_matrices.ps.gz`.

[26] N. Linial and A. Shraibman. Learning complexity vs. communication complexity. Manuscript at `http://www.cs.huji.ac.il/~nati/PAPERS/lcc.pdf`, December 2006.

[27] N. Linial and A. Shraibman. Lower bounds in communication complexity based on factorization norms. Manuscript at `http://www.cs.huji.ac.il/~nati/PAPERS/ccfn.pdf`, December 2006.

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

[29] R. Paturi and J. Simon. Probabilistic communication complexity. *J. Comput. Syst. Sci.*, 33(1):106–123, 1986.

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

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

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

[33] M. E. Saks. Slicing the hypercube. *Surveys in combinatorics, 1993*, pages 211–255, 1993.

[34] K.-Y. Siu and J. Bruck. On the power of threshold circuits with small weights. *SIAM J. Discrete Math.*, 4(3):423–435, 1991.

[35] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In *COLT*, pages 545–560, 2005.

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

[37] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[38] S. Vempala. A random sampling based algorithm for learning the intersection of halfspaces. In *Proceedings of the 38th Annual Symposium on Foundations of Computer Science*, pages 508–513, 1997.

[39] K. Yang. New lower bounds for statistical query learning. *J. Comput. Syst. Sci.*, 70(4):485–509, 2005.

A Statistical Query Dimension

This section presents a folklore result that is needed in the proofs of Theorems 1.1 and 7.1.

Proposition 2.6 (Restated from p. 6). *Let* $\mathrm{sqdim}_\mu(\mathscr{C}) = N$. *Then there is a set* $\mathscr{H}$ *of* $|\mathscr{H}| = N$ *functions in* $\mathscr{C}$ *such that each* $f \in \mathscr{C}$ *has* $|\mathbf{E}_\mu[f \cdot h]| > 1/(N+1)$ *for some* $h \in \mathscr{H}$.

Proof. For a set $\mathscr{F} \subseteq \mathscr{C}$, define

$$\gamma(\mathscr{F}) \rightleftharpoons \max_{f_1 \neq f_2 \in \mathscr{F}} \{|\mathbf{E}_\mu[f_1 \cdot f_2]|\},$$

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

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

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

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

B Discrepancy

This section reviews tools needed in the discrepancy calculation of Theorem 7.1, in Appendix C. We start with an important observation that arises as a special case in the work of Ford and Gál [9, Theorem 3.1] on multiparty communication complexity. It is also implicit in an article by Raz [30, Lemma 5.1].

Lemma B.1 (Ford and Gál [9], Raz [30]). *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 [30]). Let $R = A \times B$ be the rectangle over 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,

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

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 B.1 in an elegant way to relate the discrepancy to the pairwise correlations of the matrix rows:

Lemma B.2 (Ford and Gál [9]). *For every Boolean function $f(x,y)$ and every product distribution $\mu = \mu_X \times \mu_Y$,*

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

Proof (adapted from Ford and Gál [9]). By Lemma B.1, there is a choice of values $\alpha_x, \beta_y \in \{-1,1\}$ for all x and y such that

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

Thus,

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

C SQ Dimension and Discrepancy

Our goal in this section is to prove Theorem 7.1:

Theorem 7.1 (Restated from p. 11). *Let $A \in \{-1,1\}^{M\times N}$. Then*

$$\sqrt{\tfrac{1}{2}\,\mathrm{sq}(A)} < \mathrm{disc}^\times(A)^{-1} < (2\,\mathrm{sq}(A))^2.$$

We divide the proof in two parts: the upper bound on $\mathrm{disc}^\times(A)$ and the lower bound.

Lemma C.1 (Upper bound). *Let $A \in \{-1,1\}^{M\times N}$. Then* $\mathrm{disc}^\times(A) < \sqrt{2/\,\mathrm{sq}(A)}$.

Proof. Assume $\mathrm{sq}(A) = d$. 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 $\mathrm{disc}_{\mu\times U}(A) < \sqrt{2/d}$:

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

□

Lemma C.2 (Lower bound). *Let* $A \in \{-1,1\}^{M\times N}$. *Then* $\mathrm{disc}^{\times}(A) > 1/(2\,\mathrm{sq}(A))^2$.

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 $\mathrm{disc}_{\mu\times\lambda}(A)$ by constructing an efficient protocol for A with a suitable advantage.

Let $\mathrm{sq}(A) = d$. Then by Proposition 2.6, there are d rows $f_1, \ldots, f_d \in \{-1,1\}^N$ 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 $i = 1,\ldots,d$. This yields the following protocol for evaluating $A(x,y)$. Bob, who knows y, sends Alice the index i of the function f_i that is best correlated with the yth row of A. This costs $\lceil \log d \rceil$ bits. Alice, who knows x, announces $f_i(x)$ as the output of the protocol.

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

$$\mathrm{disc}_{\mu\times\lambda}(A) > \frac{1}{(d+1)\cdot 2^{1+\lceil\log d\rceil}} > \frac{1}{4d^2}.$$

□

Lemmas C.1 and C.2 immediately yield Theorem 7.1.