Combinatorics, Probability and Computing (2009) 18, 227–245. © 2009 Cambridge University Press
doi:10.1017/S0963548308009656 Printed in the United Kingdom

Learning Complexity vs Communication Complexity

NATI LINIAL[1†] and ADI SHRAIBMAN[2‡]

[1]School of Computer Science and Engineering, Hebrew University, Jerusalem, Israel
(e-mail: `nati@cs.huji.ac.il`)

[2]Department of Mathematics, Weizmann Institute of Science, Rehovot, Israel
(e-mail: `adi.shribman@gmail.com`)

Received 14 December 2006; revised 23 October 2008; first published online 15 January 2009

This paper has two main focal points. We first consider an important class of machine learning algorithms: large margin classifiers, such as Support Vector Machines. The notion of *margin complexity* quantifies the extent to which a given class of functions can be learned by large margin classifiers. We prove that up to a small multiplicative constant, margin complexity is equal to the inverse of discrepancy. This establishes a strong tie between seemingly very different notions from two distinct areas.

In the same way that matrix rigidity is related to rank, we introduce the notion of rigidity of margin complexity. We prove that sign matrices with small margin complexity rigidity are very rare. This leads to the question of proving lower bounds on the rigidity of margin complexity. Quite surprisingly, this question turns out to be closely related to basic open problems in communication complexity, *e.g.*, whether $PSPACE$ can be separated from the polynomial hierarchy in communication complexity.

Communication is a key ingredient in many types of learning. This explains the relations between the field of learning theory and that of communication complexity [6, 10, 16, 26]. The results of this paper constitute another link in this rich web of relations. These new results have already been applied toward the solution of several open problems in communication complexity [18, 20, 29].

1. Introduction

Several papers have investigated the relationships between parameters from learning theory and their counterparts from communication complexity. For example, unbounded error communication complexity can be characterized in terms of dimension complexity [26]. Also, as shown in [16], there is an equivalence between VC-dimension and one-way distributional complexity

[†] Supported by a grant from the Israel Science Foundation.

[‡] Most of the work on this paper was done while the author was a student at the Hebrew University.

with respect to product distributions. The present paper adds some new ingredients to the emerging relations between these two disciplines.

Large margin classifiers such as Support Vector Machines (a.k.a. SVM), occupy a central place in present-day machine learning in both theory and practice. Our initial motivation was to understand the strengths and weaknesses of large margin classifiers. (For general background on this subject see, *e.g.*, [8, 33]). This has led us quite naturally to investigate the margin complexity of sign matrices. We were pleased to discover that these parameters are also central to the field of communication complexity.

We first describe the learning-theoretic point of view, and define margin complexity, and then review the relevant background and explain our new results.

A *classification algorithm* receives as input a *sample* $(z_1, f(z_1)), \dots, (z_m, f(z_m))$ which is a sequence of points $\{z_i\}$ from a set $\mathcal{D}$ (the *domain*) and the corresponding evaluations of some unknown function $f : \mathcal{D} \to \{\pm 1\}$. The output of the algorithm is a function $h : \mathcal{D} \to \{\pm 1\}$, which should be close to f. Here we think of f as chosen by an adversary from a predefined class $\mathcal{F}$ (the so-called *concept class*). (In real-world applications, the choice of the class $\mathcal{F}$ represents our prior knowledge of the situation.)

Large margin classifiers take the following route to the solution of classification problems. The domain $\mathcal{D}$ is mapped into $\mathbb{R}^t$ (this map is usually called a *feature map*). If z_i is mapped to x_i for each i, our sample points are now $\{x_i\} \subset \mathbb{R}^t$. The algorithm then seeks a linear functional (*i.e.*, a vector) y that maximizes

$$m_f(\{x_i\}, y) = \min_i \frac{|\langle x_i, y\rangle|}{\|x_i\|_2 \|y\|_2}.$$

under the constraint that $\text{sign}(\langle x_j, y\rangle) = f(x_j)$, for all j. We denote this maximum by $m_f(\{x_i\})$.

Clearly, an acceptable linear functional y defines a hyperplane H that separates the points (above and below H) as dictated by the function f. What determines the performance of the classifier associated with y is the distances of the points x_i from H, *i.e.*, the *margin* $m_f(\{x_i\}, y)$. Thus, the margin captures the extent to which the family $\mathcal{F}$ can be described by the sign of a linear functional. We study the margin, in quest of those properties of a concept class that determine how well suited it is for such a description.

These considerations lead us to define the *margin of a class of functions*. But before we do that, some words about the feature map are in order. The theory and the practice of choosing a feature map is at present a subtle art. Making the proper choice of a feature map can have a major impact on the performance of the classification algorithm. Our intention here is to avoid this delicate issue and concentrate instead on the concept class *per se*. In order to bypass the dependence of our analysis on the choice of a feature map, we consider the *best-possible* choice. This explains the supremum in the definition of margin below:

$$m(\mathcal{F}) = \sup_{\{x_i\}} \inf_{f \in \mathcal{F}} m_f(\{x_i\}).$$

How should we model this set-up? For every set of m samples there is only a finite number, say n, of possible classifications by functions from the relevant concept class. Consequently, we can represent a concept class by an $m \times n$ sign matrix, each column of which represents a function

$f : [m] \to \{\pm 1\}$. It should be clear, then, that the margin of a sign matrix A is

$$m(A) = \sup \min_{i,j} \frac{|\langle x_i, y_j \rangle|}{\|x_i\|_2 \|y_j\|_2}, \tag{1.1}$$

where the supremum is over all $x_1, \dots, x_m, y_1, \dots, y_n \in \mathbb{R}^{m+n}$ such that $\operatorname{sign}(\langle x_i, y_j \rangle) = a_{ij}$, for all i,j. (It is not hard to show that there is no advantage in working in any higher-dimensional Euclidean space.) It is also convenient to define $\operatorname{mc}(A) = m(A)^{-1}$, the *margin complexity* of A.

We mention below some previous results and several simple observations on margin complexity. We begin with some very rough bounds.

Observation 1.1. *For every $m \times n$ sign matrix A,*

$$1 \leqslant \operatorname{mc}(A) \leqslant \min\{\sqrt{m}, \sqrt{n}\}.$$

The lower bound follows from Cauchy–Schwarz. For the upper bound, assume w.l.o.g. that $m \geqslant n$ and let x_i be the ith row of A and y_j the jth vector in the standard basis.

The first paper on margin complexity [6] mainly concerns the case of random matrices. Among other things they proved the following.

Theorem 1.2 (Ben-David, Eiron and Simon [6]). *Almost every*[1] *$n \times n$ sign matrix has margin complexity at least $\Omega(\sqrt{\frac{n}{\log n}})$.*

This theorem illustrates the general principle that random elements are complex. A main goal in that paper is to show that *VC-dimension* and margin complexity are very distinct measures of complexity, as follows.

Theorem 1.3 (Ben-David, Eiron and Simon [6]). *Let $d \geqslant 2$. Almost every matrix with VC-dimension at most $2d$ has margin complexity larger than*

$$\Omega\big(n^{\frac{1}{2} - \frac{1}{2d} - \frac{1}{2^{d+1}}}\big).$$

If $A : U \to V$ is a linear map between two normed spaces, we denote its operator norm by $\|A\|_{U \to V} = \max_{x:\|x\|_U = 1} \|Ax\|_V$, with the shorthand $\|\cdot\|_{p \to q}$ to denote $\|\cdot\|_{\ell_p \to \ell_q}$. A particularly useful instance of this norm is $\|A\|_{2 \to 2}$. It is well known that $\|A\|_{2\to 2}$ is the largest singular value of A. Moreover, this quantity can be computed efficiently. Forster [9] proved the following lower bound on margin complexity.

Claim 1.4 (Forster [9]). *For every $m \times n$ sign matrix A,*

$$\operatorname{mc}(A) \geqslant \frac{\sqrt{nm}}{\|A\|_{2\to 2}}.$$

[1] Here and below we adopt a common abuse of language and use the shorthand ‘almost every’ to mean ‘asymptotically almost every’.

This result has several nice consequences. For example, it implies that almost every $n \times n$ sign matrix has margin complexity $\Omega(\sqrt{n})$. Also, together with Observation 1.1 it yields that the margin complexity of an $n \times n$ Hadamard matrix is $\sqrt{n}$. Forster's proof is of interest too, and provides more insight than earlier proofs which were based on counting arguments.

Subsequent papers [10, 11, 12], following [9], improved Forster's bound in different ways. Connections were shown between margin complexity and other complexity measures. These papers also determine exactly the margin complexity of some specific families of matrices.

In [19] we noticed the relation between margin complexity and factorization norms. Let an operator $A : U \to V$ and a normed space W be given. In order to define the corresponding factorization of A, we should find how to express A as $A = XY$, where $Y : U \to W$ and $X : W \to V$, so as to minimize the product of X's and Y's operator norms. Of special interest is the case $U = \ell_1^n$, $V = \ell_\infty^m$ and $W = \ell_2$. We denote

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

It is well known (*e.g.*, [27]) that γ_2 is indeed a norm. Its dual is denoted, as usual by $\gamma_2^*(\cdot)$. One can easily check that $\|B\|_{1\to 2}$ is the largest ℓ_2 norm of a column of B, and $\|B\|_{2\to\infty}$ is the largest ℓ_2 norm of a row of B.

It is proved in [19] that, for every sign matrix A,

$$\mathrm{mc}(A) = \min_{B:\, b_{ij}a_{ij}\geqslant 1\ \forall i,j} \gamma_2(B). \tag{1.2}$$

This identity turns out to be very useful in the study of margin complexity. Some consequences drawn in [19] are as follows. For every sign matrix A,

- $\mathrm{mc}(A) = \max_{B:\mathrm{sign}(B)=A, \gamma_2^*(B)\leqslant 1}\langle A, B\rangle$,
- $\mathrm{mc}(A) \leqslant \gamma_2(A) \leqslant \sqrt{\mathrm{rank}(A)}$,
- let $RC(A)$ be the randomized or quantum communication complexity of A; then

$$2\log \mathrm{mc}(A) - \Theta(1) \leqslant RC(A) \leqslant O(\mathrm{mc}(A)^2).$$

In this paper, we derive another consequence of the relation between margin complexity and γ_2. *Discrepancy* is a combinatorial notion that comes up in many contexts: see, *e.g.*, [7, 22]. We prove here that margin and discrepancy are equivalent up to a constant factor for every sign matrix. Let A be an $m \times n$ sign matrix, and P a probability measure on its entries. We define

$$\mathrm{disc}_P(A) = \max_{S\subset[m], T\subset[n]} \left| \sum_{i\in S, j\in T} p_{ij}a_{ij} \right|.$$

The discrepancy of A is then defined by

$$\mathrm{disc}(A) = \min_P \mathrm{disc}_P(A).$$

Theorem 3.1. *For every sign matrix A,*

$$\mathrm{disc}(A) \leqslant m(A) \leqslant 8\ \mathrm{disc}(A).$$

Discrepancy is used to derive lower bounds on communication complexity in different models [5, 34]. Theorem 3.1 provides additional evidence for the role of margins in the field of communication complexity. (See also [6, 10, 20]). As described below, we find here additional new relations to communication complexity, specifically to questions about separation of communication complexity classes.

It is very natural to consider also classification algorithms that tolerate a certain probability of error but achieve larger margins. Namely, we are led to consider the following complexity measure,

$$\mathrm{mc}_r(A,l) = \min_{B:h(B,A)\leqslant l} \mathrm{mc}(B),$$

where $h(A,B)$ is the Hamming distance between the two matrices. We call this quantity *mc-rigidity*. The relation between this complexity measure and margin complexity is analogous to the relation between *rank rigidity* and rank. Rank rigidity (usually simply called rigidity) was first defined in [32] and has attracted considerable interest, *e.g.*, [15, 21, 30]. A major reason to study rank rigidity is that, as shown in [32], the construction of explicit examples of sign matrices with high rank rigidity would have very significant consequences in computational complexity.

It transpires that mc-rigidity behaves similarly. To begin, it does not seem easy to construct sign matrices with high mc-rigidity (where 'high' means close to the expected complexity of a random matrix). Furthermore, we are able to establish interesting relations between the construction of sign matrices of high mc-rigidity and the separation of complexity classes in communication complexity, as introduced and studied in [5, 21].

The mc-rigidity of random matrices is considered in [24, 25]. There it is shown that there is an absolute constant $1 > c > 0$ such that, for almost every $n \times n$ sign matrix,

$$\mathrm{mc}_r(A, cn^2) \geqslant \Omega(\sqrt{n}).$$

We give a bound on the number of sign matrices with small mc-rigidity that is much stronger than that of [24, 25]. Our proof is also significantly simpler.

Regarding explicit bounds, we prove the following lower bounds on mc-rigidity.

Theorem 5.1. *Every $m \times n$ sign matrix A satisfies*

$$\mathrm{mc}_r\left(A, \frac{mn}{8g}\right) \geqslant g,$$

provided that $g < \frac{mn}{2K_G\|A\|_{\infty\to 1}}$. (Here $K_G \leqslant 1.8$ is Grothendieck's constant: see Theorem 2.1.)

Theorem 5.2. *Every $n \times n$ sign matrix A with $\gamma_2(A) \geqslant \Omega(\sqrt{n})$ (this is a condition satisfied by almost every sign matrix) satisfies*

$$\mathrm{mc}_r(A, cn^2) \geqslant \Omega(\sqrt{\log n}),$$

for some constant $c > 0$.

In a 1986 paper [5], Babai, Frankl and Simon took a complexity-theoretic approach to communication complexity. They defined communication complexity classes analogous to computational complexity classes. For example, the polynomial hierarchy is defined as follows. We define the following classes of $2^m \times 2^m$ 0–1 matrices. We begin with Σ_0^{cc}, the set of combinatorial rectangles, and with $\Pi_0^{cc} = co\Sigma_0^{cc}$. From here we proceed to define

$$\Sigma_i^{cc} = \left\{ A | A = \bigvee_{j=1}^{2^{\text{polylog}(m)}} A_j, A_j \in \Pi_{i-1}^{cc} \right\},$$

$$\Pi_i^{cc} = \left\{ A | A = \bigwedge_{j=1}^{2^{\text{polylog}(m)}} A_j, A_j \in \Sigma_{i-1}^{cc} \right\}.$$

For more on communication complexity classes see [5, 17, 21].

Some communication complexity classes were implicitly defined prior to [5]. In particular, it is possible to define and investigate the communication complexity analogues of important complexity classes such as P, NP, $coNP$, PH and AM. For example, it was shown in [1] that $P^{cc} = NP^{cc} \cap coNP^{cc}$.

It remains a major open question in this area whether the hierarchy can be separated. We approach this problem using results of Lokam [21] and Tarui [31]. (In our statement of Theorems 4.4 and 4.5 we adopt a common abuse of language and speak of individual matrices where we should refer to an infinite family of sign matrices of growing dimensions.)

Theorem 4.4. *Let A be an $n \times n$ sign matrix. If there exists a constant $c \geqslant 0$ such that, for every $c_1 \geqslant 0$,*

$$\text{mc}_r(A, n^2/2^{(\log\log n)^c}) \geqslant 2^{(\log\log n)^{c_1}},$$

then A is not in PH^{cc}.

Theorem 4.5. *Let A be an $n \times n$ sign matrix. If*

$$\text{mc}_r(A, n^2/2^{(\log\log n)^c}) \geqslant 2^{(\log\log n)^{c_1}}$$

for every $c, c_1 \geqslant 0$, then A is not in AM^{cc}.

As mentioned, questions about rigidity tend to be difficult, and mc-rigidity seems to follow this pattern as well. However, the following conjecture, if true, would shed some light on the mystery surrounding mc-rigidity:

Conjecture 6.1. *For every constant c_1 there are constants c_2, c_3 such that every $n \times n$ sign matrix A satisfying* $\text{mc}(A) \geqslant c_1\sqrt{n}$ *also satisfies*

$$\text{mc}_r(A, c_2 n^2) \geqslant c_3\sqrt{n}.$$

What the conjecture says is that every matrix with high margin complexity has high mc-rigidity as well. In particular, explicit examples are known for matrices of high margin complexity, *e.g.*, Hadamard matrices. It would follow that such matrices have high mc-rigidity too.

The rest of this paper is organized as follows. We start with relevant background and notation in Section 2. In Section 3 we prove the equivalence of discrepancy and margin. Section 4 contains the definition of mc-rigidity, the mc-rigidity of random matrices, and applications to the theory of communication complexity classes. In Section 5 we prove lower bounds on mc-rigidity, and discuss relations to rank rigidity. Open questions are discussed in Section 6.

2. Background and notation

Basic notation. Let A and B be two real matrices. We use the following notation.

- The inner product of A and B is denoted $\langle A, B\rangle = \sum_{ij} a_{ij}b_{ij}$.
- Matrix norms: $\|B\|_1 = \sum |b_{ij}|$ is B's ℓ_1 norm, $\|B\|_2 = \sqrt{\sum b_{ij}^2}$ is its ℓ_2 (Frobenius) norm, and $\|B\|_\infty = \max_{ij} |b_{ij}|$ is its ℓ_∞ norm.
- If A and B are sign matrices then $h(A,B) = \frac{1}{2}\|A-B\|_1$ denotes the Hamming distance between A and B.

Dimension complexity. The *dimension complexity* of a sign matrix A is defined as the smallest dimension $d = d(A)$ such that there exist sets of vectors $\{x_i\}, \{y_j\} \subset \mathbb{R}^d$ so that for all i, j there holds $a_{i,j} = \text{sign}(\langle x_i . y_j\rangle)$. Equivalently, it is not hard to see that

$$d(A) = \min_{B : A = \text{sign}(B)} \text{rank}(B).$$

(For more about this complexity measure see [6, 9, 10, 12, 19, 26].)

Discrepancy. Let A be a sign matrix, and let P be a probability measure on the entries of A. The P-discrepancy of A, denoted $\text{disc}_P(A)$, is defined as the maximum over all combinatorial rectangles R in A of $|P^+(R) - P^-(R)|$, where P^+ $[P^-]$ is the measure of the positive entries (negative entries).

The *discrepancy* of a sign matrix A, denoted $\text{disc}(A)$, is the minimum of $\text{disc}_P(A)$ over all probability measures P on the entries of A.

We make substantial use of *Grothendieck's inequality* (see, *e.g.*, [27, p. 64]), which we now recall.

Theorem 2.1 (Grothendieck's inequality). *There is a universal constant K_G such that, for every real matrix B and every $k \geqslant 1$,*

$$\max \sum b_{ij}\langle u_i, v_j\rangle \leqslant K_G \max \sum b_{ij}\epsilon_i\delta_j, \tag{2.1}$$

where the max are over the choice of $u_1, \dots, u_m, v_1, \dots, v_n$ as unit vectors in $\mathbb{R}^k$ and $\epsilon_1, \dots, \epsilon_m$, $\delta_1, \dots, \delta_n \in \{\pm 1\}$.

The constant K_G is called Grothendieck's constant. Its exact value is not known but it is proved that $1.5 \leqslant K_G \leqslant 1.8$.

As mentioned, we denote by γ_2^* the dual norm of γ_2, *i.e.*, for every real matrix B,

$$\gamma_2^*(B) = \max_{C : \gamma_2(C) \leqslant 1} \langle B, C\rangle.$$

We note that the norms γ_2^* and $\|\cdot\|_{\infty\to 1}$ are equivalent up to a small multiplicative factor, namely, for any real matrix,

$$\|B\|_{\infty\to 1} \leqslant \gamma_2^*(B) \leqslant K_G\|B\|_{\infty\to 1}. \tag{2.2}$$

The left inequality is easy, and the right inequality is a reformulation of Grothendieck's inequality. Both use the observation that the left-hand side of (2.1) equals $\gamma_2^*(B)$, and the max term on the right-hand side is $\|B\|_{\infty\to 1}$.

The norm dual to $\|\cdot\|_{\infty\to 1}$ is the *nuclear norm* from l_1 to l_∞. The nuclear norm of a real matrix B is defined as follows:

$$\nu(B) = \min\{\sum |w_i| \text{ such that } B \text{ can be expressed as}$$
$$\sum w_i x_i y_i^t = B \text{ for some choice of sign vectors } x_1, x_2, \ldots, y_1, y_2 \ldots\}.$$

See [13] for more details.

It is a simple consequence of the definition of duality and (2.2) that, for every real matrix B,

$$\gamma_2(B) \leqslant \nu(B) \leqslant K_G \cdot \gamma_2(B). \tag{2.3}$$

3. Margin and discrepancy are equivalent

Here we prove (recall that $\mathrm{mc}(A) = m(A)^{-1}$) the following.

Theorem 3.1. *For every sign matrix A,*

$$\mathrm{disc}(A) \leqslant m(A) \leqslant 8\ \mathrm{disc}(A).$$

We first define a variant of margin.

Margin with sign vectors. Given an $m \times n$ sign matrix A, denote by $\Lambda = \Lambda(A)$ the set of all pairs of *sign matrices* X, Y such that the sign pattern of XY equals A, *i.e.*, $A = \mathrm{sign}(XY)$ and let

$$m_\nu(A) = \max_{(X,Y)\in\Lambda} \min_{i,j} \frac{|\langle x_i, y_j\rangle|}{\|x_i\|_2 \|y_j\|_2}. \tag{3.1}$$

Here x_i is the ith row of X, and y_j is the jth column of Y. The definition of m_ν is almost the same as that of margin (equation (1.1)), except that in defining m_ν we consider only pairs of *sign matrices* X, Y and not arbitrary matrices. It is therefore clear that $m_\nu(A) \leqslant m(A)$ for every sign matrix A. As we see next, the two parameters are equivalent up to a small multiplicative constant.

3.1. Proof of Theorem 3.1

First we prove that margin and m_ν are equivalent up to multiplication by the Grothendieck constant. Then we show that m_ν is equivalent to discrepancy up to a multiplicative factor of at most 4.

Lemma 3.2. *For every sign matrix A,*

$$m_\nu(A) \leqslant m(A) \leqslant K_G \cdot m_\nu(A),$$

where K_G is the Grothendieck constant.

Proof. The left inequality is an easy consequence of the definitions of m and m_ν, so we focus on the right one. Let $\mathcal{B}_\nu$ be the convex hull of rank-one sign matrices. The convex body $\mathcal{B}_\nu$ is the unit ball of the nuclear norm ν, which is dual to the operator norm from ℓ_∞ to ℓ_1. With this terminology we can express $m_\nu(A)$ as

$$m_\nu(A) = \max_{B \in \mathcal{B}_\nu} \min_{ij} \; a_{ij} b_{ij}. \tag{3.2}$$

It is not hard to check (using equation (1.2)) that $m(A)$ can be equivalently expressed as

$$m(A) = \max_{B \in \mathcal{B}_{\gamma_2}} \min_{ij} \; a_{ij} b_{ij},$$

where $\mathcal{B}_{\gamma_2}$ is the unit ball of the γ_2 norm.

Equation (2.3) can be restated as

$$\mathcal{B}_\nu \subset \mathcal{B}_{\gamma_2} \subset K_G \cdot \mathcal{B}_\nu.$$

Now let $B \in \mathcal{B}_{\gamma_2}$ be a real matrix satisfying $m(A) = \min_{ij} \; a_{ij} b_{ij}$. The matrix $K_G^{-1} B$ is in $\mathcal{B}_\nu$ and therefore

$$m_\nu(A) \geqslant K_G^{-1} \min_{ij} \; a_{ij} b_{ij} = K_G^{-1} m(A). \qquad \square$$

Remark. Grothendieck's inequality has an interesting consequence in the study of large margin classifiers. As mentioned above, such classifiers map the sample points into $\mathbb{R}^t$ and then seek an optimal linear classifier (a linear functional, *i.e.*, a real vector). Grothendieck's inequality implies that if we restrict ourselves to mapping the sample points only into $\{\pm 1\}^k$, then the resulting loss in margin is at worst a factor of K_G.

We return to prove the equivalence between m_ν and discrepancy. The following relation between discrepancy and the $\infty \to 1$ norm is fairly simple (*e.g.*, [4]):

$$\operatorname{disc}(A) \leqslant \min_P \|P \circ A\|_{\infty \to 1} \leqslant 4 \cdot \operatorname{disc}(A)$$

where $P \circ A$ denotes the Hadamard (entry-wise) product of the two matrices.

Lemma 3.3. *Let $\mathcal{P}$ denote the set of matrices whose elements are non-negative and sum up to* 1. *For every sign matrix A,*

$$m_\nu(A) = \min_{P \in \mathcal{P}} \|P \circ A\|_{\infty \to 1}. \tag{3.3}$$

Proof. We express m_ν as the optimum of some linear program and observe that the right-hand side of equation (3.3) is the optimum for the dual program. The statement then follows from LP duality.

Equation (3.2) allows us to express m_ν as the optimum of a linear program. The variables of this program correspond to a probability measure q on the vertices of the polytope $\mathcal{B}_\nu$, and an auxiliary variable δ is used to express $\min_{ij} \; a_{ij} b_{ij}$. The vertices of $\mathcal{B}_\nu$ are in 1:1 correspondence with all $m \times n$ sign matrices of rank one. We denote this collection of matrices by $\{X_i | i \in I\}$.

The linear program is:

$$\begin{array}{lc} & \text{maximize } \delta \\ \text{s.t.} & \\ & \sum_{i\in I} q_i(X_i \circ A) - \delta J \geqslant 0 \\ \forall\, i \in I & q_i \geqslant 0 \\ & \sum_i q_i = 1. \end{array}$$

Here J is the all-ones matrix. It is not hard to see that the dual of this linear program is:

$$\begin{array}{lc} & \text{minimize } \Delta \\ \text{s.t.} & \\ \forall\, i \in I & \langle P \circ A, X_i\rangle = \langle P, X_i \circ A\rangle \leqslant \Delta \\ \forall i, j & p_{ij} \geqslant 0 \\ & \sum_{i,j} p_{ij} = 1, \end{array}$$

where $P = (p_{ij})$. The optimum of the dual program is equal to the right-hand side of equation (3.3) by definition of $\|\cdot\|_{\infty\to 1}$. The statement of the lemma follows from LP duality. □

To conclude, we have proved the following.

Theorem 3.4. *The ratio between any two of the following four parameters is at most* 8 *for any sign matrix A,*

- $m(A) = \mathrm{mc}(A)^{-1}$,
- $m_v(A)$,
- $\mathrm{disc}(A)$,
- $\min_{P\in\mathcal{P}} \|P \circ A\|_{\infty\to 1}$, *where $\mathcal{P}$ is the set of matrices with non-negative entries that sum up to* 1.

4. Soft margin complexity, or mc-rigidity

As mentioned in the Introduction, some classification algorithms allow the classifier to make a few mistakes, and yield in return a better margin. Such algorithms are called *soft margin algorithms*. The complexity measure associated with these algorithms is what we call mc-rigidity. The mc-rigidity of a sign matrix A is defined as

$$\mathrm{mc}_r(A, l) = \min_{B : h(B,A)\leqslant l} \mathrm{mc}(B)$$

where $h(\cdot,\cdot)$ is the Hamming distance. We prove that low mc-rigidity is rare.

Theorem 4.1. *There is a constant $c > 0$ such that the number of $n \times n$ sign matrices A that satisfy $\mathrm{mc}_r(A, l) \leqslant \sqrt{\frac{l}{n}}$ is at most*

$$\left(\frac{n^2}{l}\right)^{c\cdot l\cdot\log\frac{n^2}{l}},$$

for every $0 < l \leqslant n^2/2$. *In particular, there exist* $\epsilon > 0$ *such that almost every* $n \times n$ *sign matrix* A *satisfies*

$$\mathrm{mc}_r(A, \epsilon n^2) > \sqrt{\epsilon n}.$$

The first part of the theorem is significantly better than previous bounds [6, 19, 24, 25]. Note that using Theorems 3.1 and 4.1 we get an upper bound on the number of sign matrices with small discrepancy. We do not know of a direct method to show that low-discrepancy matrices are so rare.

Theorem 4.1 is reminiscent of bounds found in [3, 28] on the number of sign matrices of small dimensional complexity.

To prove Theorem 4.1 we use the following theorem by Warren (see [2] for a comprehensive discussion), and the lemma below it.

Theorem 4.2 (Warren 1968). *Let* $P_1, \ldots, P_m$ *be real polynomials in* $t \leqslant m$ *variables, of total degree* $\leqslant k$ *each. Let* $s(P_1, \ldots, P_m)$ *be the total number of sign patterns of the vectors* $(P_1(x), \ldots, P_m(x))$*, over* $x \in \mathbb{R}^t$. *Then*

$$s(P_1, \ldots, P_m) \leqslant \bigl(4ekm/t\bigr)^t.$$

In the next lemma we consider the relation between margin complexity and dimension complexity (see Section 2). This relation makes it possible to use Warren's theorem in the proof of Theorem 4.1.

Lemma 4.3. *Let* B *be an* $n \times n$ *sign matrix, and let* $0 < \rho < 1$. *There exists a matrix* $\tilde{B}$ *with Hamming distance* $h(B, \tilde{B}) < \rho n^2$*, such that*

$$d(\tilde{B}) \leqslant O(\log \rho^{-1} \cdot \mathrm{mc}(B)^2).$$

Proof. We use the following known fact (*e.g.*, [14, 23]). Let $x, y \in \mathbb{R}^n$ be two unit vectors with $|\langle x, y\rangle| \geqslant \epsilon$. Then

$$\Pr_L\bigl(\mathrm{sign}(\langle P_L(x), P_L(y)\rangle) \neq \mathrm{sign}(\langle x, y\rangle)\bigr) \leqslant 4e^{-k\epsilon^2/8},$$

where the probability is over k-dimensional subspaces L, and where $P_L : \mathbb{R}^n \to L$ is the projection onto L.

By definition of the margin complexity, there are two $n \times n$ matrices X and Y such that

- $B = \mathrm{sign}(XY)$,
- every entry in XY has absolute value $\geqslant 1$,
- $\|X\|_{2\to\infty} = \|Y\|_{1\to 2} = \sqrt{\mathrm{mc}(B)}$.

Let $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ denote the rows of X and columns of Y respectively. Take C such that $4e^{-C/8} \leqslant \rho$. Then, by the above fact, for $k = C\,\mathrm{mc}(B)^2$ there is a k-dimensional linear subspace L, such that projecting the points onto L preserves at least $(1-\rho)n^2$ signs of the n^2 inner products $\{\langle x_i, y_j\rangle\}$. □

To complete the proof of Theorem 4.1 let A be a sign matrix with $\mathrm{mc}_r(A,l) \leqslant \mu$. Namely, it is possible to flip at most l entries in A to obtain a sign matrix B with $\mathrm{mc}(B) \leqslant \mu$. Let $\rho = l/n^2$ and apply Lemma 4.3 to B. This yields a matrix E, such that the Hamming distance $h(\mathrm{sign}(E), B) \leqslant l$ and E has rank $O(\log \rho^{-1} \cdot \mu^2)$ (In the terminology of Lemma 4.3 $\tilde{B} = \mathrm{sign}(E)$.) To sum up, we change at most l entries in A to obtain B and then at most l more entries to obtain $\mathrm{sign}(E)$, a matrix of dimension complexity $O(\log \rho^{-1} \cdot \mu^2)$. Therefore $A = \mathrm{sign}(E + F_1 + F_2)$, where F_1, F_2 have support of size at most l each (corresponding to the entries where sign flips were made).

Now E can be expressed as $E = UV^t$ for some $n \times r$ matrices U and V with $r \leqslant c_1 \log \rho^{-1} \cdot \mu^2$ (c_1 is a constant). Let us fix one of the $\leqslant \binom{n^2}{l}^2$ choices for the supports of the matrices F_1, F_2 and consider the entries of U, V and the non-zero entries in F_1, F_2 as formal variables. Each entry in A is the sign of a polynomial of degree 2 in these variables. We apply Warren's theorem (Theorem 4.2) with these parameters to conclude that the number of $n \times n$ sign matrices A with $\mathrm{mc}_r(A,l) \leqslant \mu$ is at most

$$\binom{n^2}{l}^2 \cdot \left(8en^2/(2c_1 \log \rho^{-1} \cdot \mu^2 \cdot n + 2l)\right)^{2c_1 \log \rho^{-1} \cdot \mu^2 \cdot n + 2l}.$$

Recall that $\rho = l/n^2$ and substitute $\mu = \sqrt{\frac{l}{n}}$, to get

$$\binom{n^2}{l}^2 \cdot \left(8en^2 \Big/ \left(2c_1 \cdot l \cdot \log \frac{n^2}{l} + 2l\right)\right)^{2c_1 \cdot l \cdot \log \frac{n^2}{l} + 2l} = \left(\frac{n^2}{l}\right)^{O(l \cdot \log \frac{n^2}{l})}.$$

4.1. Communication complexity classes

Surprisingly, mc-rigidity is related to questions about separating communication complexity classes. It is not necessary to know the precise definitions of these classes in order to read what follows. (The interested reader can find the definitions, *e.g.*, in [21]). A major open problem from [5] is to separate the polynomial hierarchy. Lokam [21] has raised the question of explicitly constructing matrices that do not belong to AM^{cc}, the class of bounded round interactive proof systems. We tie these questions to mc-rigidity.

Theorem 4.4. *Let A be an $n \times n$ sign matrix. If there exists a constant $c \geqslant 0$ such that, for every $c_1 \geqslant 0$,*

$$\mathrm{mc}_r(A, n^2/2^{(\log\log n)^c}) \geqslant 2^{(\log\log n)^{c_1}},$$

then A is not in PH^{cc}.

Theorem 4.5. *Let A be an $n \times n$ sign matrix. If*

$$\mathrm{mc}_r(A, n^2/2^{(\log\log n)^c}) \geqslant 2^{(\log\log n)^{c_1}}$$

for every $c, c_1 \geqslant 0$, then A is not in AM^{cc}.

We now present the proofs of Theorems 4.4 and 4.5.

Proof of Theorem 4.4. The theorem is a consequence of the definition of mc_r and the following claim. For every $2^m \times 2^m$ sign matrix $A \in PH^{cc}$ and every constant $c \geqslant 0$ there is a constant $c_1 \geqslant 0$ and a matrix B such that:

(1) the entries of B are non-zero integers,
(2) $\gamma_2(B) \leqslant 2^{(\log\log n)^{c_1}}$,
(3) $h(A, \mathrm{sign}(B)) \leqslant n^2/2^{(\log\log n)^c}$.

The proof of this claim is based on a theorem of Tarui [31] (see also Lokam [21]).

It should be clear how Boolean gates operate on 0–1 matrices. By definition of the polynomial hierarchy in communication complexity, every Boolean function f in Σ_k can be computed by an AC^0 circuit of polynomial size whose inputs are 0–1 matrices of size $2^m \times 2^m$ and rank 1. Namely,

$$f(x, y) = C(X_1, \ldots, X_s),$$

where C is an AC^0 circuit, $\{X_i\}_{i=1}^s$ are 0–1 rank-1 matrices, and $s \leqslant 2^{\mathrm{polylog}(m)}$.

Now, AC^0 circuits are well approximated by low degree polynomials, as proved by Tarui [31]. Let C be an AC^0 circuit of size $2^{\mathrm{polylog}(m)}$ acting on $2^m \times 2^m$ 0–1 matrices $\phi_1, \ldots, \phi_s$. Fix $0 < \delta = 2^{(\log m)^c}$ for some constant $c \geqslant 0$. Then there exists a polynomial $\Phi \in \mathbb{Z}[X_1, \ldots, X_s]$ such that:

(1) the sum of absolute values of the coefficients of Φ is at most $2^{\mathrm{polylog}(m)}$,
(2) the fraction of entries where the matrices $C(\phi_1, \ldots, \phi_s)$ and $\Phi(\phi_1, \ldots, \phi_s)$ differ is at most δ (here and below, when we evaluate $\Phi(\phi_1, \ldots, \phi_s)$, products are pointwise matrix products),
(3) where Φ and C differ, $\Phi(\phi_1, \ldots, \phi_s) \geqslant 2$.

Let us apply Tarui's theorem on the 0–1 version of A, and let $\Phi = \sum_{T\in\{0,1\}^s} a_T \Pi_{i\in T} X_i$ be the polynomial given by the theorem. Notice that $Y_T = \Pi_{i\in T} X_i$ has rank 1. Let

$$B = \left(\sum_{T\in\{0,1\}^s} a_T Y_T\right) - J.$$

Then,

(1) the entries of B are non-zero integers,
(2) $\gamma_2(B) \leqslant 1 + \sum_{T\in\{0,1\}^s} |a_T| \leqslant 2^{\mathrm{polylog}(m)}$,
(3) $h(A, \mathrm{sign}(B)) \leqslant \delta n^2$,

as claimed. □

Proof of Theorem 4.5. We first recall some background from [21]. A family $\mathcal{G} \subset 2^{[n]}$, is said to generate a family $\mathcal{F} \subset 2^{[n]}$ if every $F \in \mathcal{F}$ can be expressed as the union of sets from $\mathcal{G}$. We denote by $g(\mathcal{F})$ the smallest cardinality of a family $\mathcal{G}$ that generates $\mathcal{F}$. Each column in a 0–1 matrix Z is considered as the characteristic vector of a set and $\Phi(\mathrm{Z})$ is the family of all such sets. If A is an $n \times n$ sign matrix, we define $\mathcal{F}(\mathrm{A})$ as $\Phi(\bar{\mathrm{A}})$ where $\bar{A}$ is obtained by replacing each -1 entry in A by zero. We denote $g(\mathcal{F}(A))$ by $g(A)$. Finally there is the rigidity variant of $g(A)$:

$$g(A, l) = \min_{B : h(B,A)\leqslant l} g(B).$$

Lokam [21, Lemma 6.3] proved that if $g(A, n^2/2^{(\log\log n)^c}) \geqslant 2^{(\log\log n)^{\omega(1)}}$ for every $c > 0$ then $A \notin AM^{cc}$. We conclude the proof by showing that

$$g(A) \geqslant (\mathrm{mc}(A) - 1)/2$$

for every $n \times n$ sign matrix A.

Let $g = g(A)$ and let $\mathcal{G} = \{G_1, \ldots, G_g\}$ be a minimal family that generates $\mathcal{F}(A)$. Let X be the $n \times g$ 0–1 matrix whose ith column is the characteristic vector of G_i. Let Y denote a $g \times n$ 0–1 matrix that specifies how to express the columns of $\bar{A}$ by unions of sets in $\mathcal{G}$. Namely, if we choose to express the ith column in $\bar{A}$ as $\cup_{t \in T} G_t$, then the ith column in Y is the characteristic vector of T. Clearly XY is a non-negative matrix whose zero pattern is given by $\bar{A}$. Consequently, the matrix $B = XY - \frac{J}{2}$ satisfies

(1) $\mathrm{sign}(B) = A$,
(2) $|b_{ij}| \geqslant 1/2$, and
(3) $\gamma_2(B) \leqslant \gamma_2(XY) + \gamma_2(\frac{J}{2}) \leqslant g + 1/2$.

It follows that

$$\mathrm{mc}(A) \leqslant \gamma_2(2B) \leqslant 2g + 1,$$

as claimed. □

5. Lower bounds on mc-rigidity

To provide some perspective for our discussion of lower bounds on mc-rigidity, it is worthwhile to recall first some of the known results about rank rigidity. The best-known explicit lower bound for rank rigidity is for the $n \times n$ Sylvester–Hadamard matrix H_n [15], and has the following form. For every $r > 0$, at least $\Omega(\frac{n^2}{r})$ changes have to be made in H_n to reach a matrix with rank at most r. Our first lower bound has a similar flavour. For example, since $\|H_n\|_{\infty\to 1} = \Theta(n^{3/2})$ (*e.g.*, Lindsey's lemma), Theorem 5.1 below implies that at least $\Omega(\frac{n^2}{g})$ sign flips in H_n are required to reach a matrix with margin complexity $\leqslant g$. (This applies for all relevant values of g, since we only have to consider $g \leqslant O(\sqrt{n})$.)

Theorem 5.1. *Every $m \times n$ sign matrix A satisfies*

$$\mathrm{mc}_r\left(A, \frac{mn}{8g}\right) \geqslant g,$$

provided that $g < \frac{mn}{2K_G\|A\|_{\infty\to 1}}$.

We conjecture that there is an absolute constant $\epsilon_0 > 0$ such that, for every sign matrix A with $\mathrm{mc}(A) \geqslant \Omega(\sqrt{n})$, at least $\Omega(n^2)$ sign flips are needed in A to reach a sign matrix with margin complexity $\leqslant \epsilon_0 \cdot \mathrm{mc}(A)$. Theorem 5.1 yields this conclusion only when $\epsilon_0 \leqslant O(\frac{1}{\sqrt{n}})$. The next theorem offers a slight improvement and yields a similar conclusion already for $\epsilon_0 \leqslant O(\sqrt{\frac{\log n}{n}})$. (Recall that $\mathrm{mc}(A) \leqslant \gamma_2(A)$ for every sign matrix A. Thus $\mathrm{mc}(A) \geqslant \Omega(\sqrt{n})$ entails the assumption of Theorem 5.2.)

Theorem 5.2. *Every $n \times n$ sign matrix A with $\gamma_2(A) \geqslant \Omega(\sqrt{n})$ satisfies*

$$\mathrm{mc}_r(A, \delta n^2) \geqslant \Omega(\sqrt{\log n}),$$

for some $\delta > 0$.

The proofs of Theorems 5.1 and 5.2 use some information about the Lipschitz constants of two of our complexity measures.

Lemma 5.3. *The Hamming distance of two sign matrices A, B is at least*

$$h(A,B) \geqslant \frac{1}{2}\big(\|B\|_{\infty\to 1} - \|A\|_{\infty\to 1}\big).$$

Proof. Let x and y be two sign vectors satisfying $\sum b_{ij}x_iy_j = \|B\|_{\infty\to 1}$. If M is the Hamming distance between A and B, then

$$\begin{aligned}\|A\|_{\infty\to 1} &\geqslant \sum a_{ij}x_iy_j = \sum b_{ij}x_iy_j + \sum(a_{ij} - b_{ij})x_iy_j \\ &\geqslant \sum b_{ij}x_iy_j - \sum |a_{ij} - b_{ij}| = \|B\|_{\infty\to 1} - 2M\end{aligned}$$

□

We next need a similar result for γ_2.

Lemma 5.4. *For every pair of sign matrices A and B,*

$$h(A,B) \geqslant \left(\frac{|\gamma_2(A) - \gamma_2(B)|}{4}\right)^4.$$

In the proof of Lemma 5.4 we need a bound on the γ_2 of sparse $(-1,0,1)$-matrices given by the following lemma.

Lemma 5.5. *Let A be a $(-1,0,1)$-matrix with N non-zero entries. Then $\gamma_2(A) \leqslant 2N^{1/4}$.*

Proof. We find matrices B and C such that $A = B + C$ and

$$\gamma_2(B), \gamma_2(C) \leqslant N^{1/4},$$

since γ_2 is a norm, and $\gamma_2(A) \leqslant 2N^{1/4}$.

Let I be the set of rows of A with more than $N^{1/2}$ non-zero entries, we define the matrices B and C by:

$$b_{ij} = \begin{cases} a_{ij} & \text{if } i \in I, \\ 0 & \text{otherwise,} \end{cases}$$

$$c_{ij} = \begin{cases} a_{ij} & \text{if } i \notin I, \\ 0 & \text{otherwise,} \end{cases}$$

The matrix B has at most $N^{1/2}$ non-zero rows, and each row in C has at most $N^{1/2}$ non-zero entries. Thus, by considering the trivial factorizations ($IX = XI = X$), we conclude that $\gamma_2(B), \gamma_2(C) \leqslant N^{1/4}$. Obviously $A = B + C$, which concludes the proof. □

Proof of Lemma 5.4. Let A and B be two sign matrices. The matrix $\frac{1}{2}(A-B)$ is a $(-1,0,1)$-matrix with $h(A,B)$ non-zero entries. Thus, by Lemma 5.5,

$$\gamma_2(A-B) \leqslant 4h^{1/4}(A,B).$$

Since γ_2 is a norm,

$$\gamma_2(A-B) \geqslant |\gamma_2(A)-\gamma_2(B)|.$$

The claim follows by combining the above two inequalities. □

We can now complete the proof of Theorems 5.1 and 5.2.

Proof of Theorem 5.1. It is proved in [19] that, for every $m \times n$ sign matrix Z,

$$\|Z\|_{\infty\to 1} \geqslant \frac{mn}{K_G \cdot \mathrm{mc}(Z)}.$$

We apply this to a matrix B with $\mathrm{mc}(B)=g$ and conclude that $\|B\|_{\infty\to 1} \geqslant \frac{mn}{gK_G}$. On the other hand, by assumption, $\|A\|_{\infty\to 1} \leqslant \frac{mn}{2gK_G}$, so by Lemma 5.3, $h(A,B) \geqslant \frac{mn}{4gK_G} \geqslant \frac{mn}{8g}$. □

Proof of Theorem 5.2. Let A be an $n \times n$ sign matrix with $\gamma_2(A) \geqslant \epsilon\sqrt{n}$, for some constant ϵ. By Lemma 5.4 there is a constant $\delta > 0$ such that every sign matrix B with $h(A,B) \leqslant \delta n^2$ satisfies $\gamma_2(B) \geqslant \frac{\epsilon}{2}\sqrt{n}$. As observed in the Discussion Section in [20], every $n \times n$ sign matrix B with $\gamma_2(B) \geqslant \Omega(\sqrt{n})$ also satisfies $\mathrm{mc}(A) \geqslant \Omega(\sqrt{\log n})$. It follows that $\mathrm{mc}_r(A, cn^2) \geqslant \Omega(\sqrt{\log n})$. □

5.1. Relations with rank rigidity

We next discuss the relation between mc-rigidity and rank rigidity. First we prove the following lower bound on rank rigidity, which compares favourably with the best known bounds (see, *e.g.*, [15]). This lower bound is related to mc-rigidity in that it is proved by the same method used to prove Theorem 5.1. We then prove lower bounds in terms of mc-rigidity on a variant of rank rigidity.

Claim 5.6. *Let A be an $n \times n$ sign matrix and let $r < \frac{n^2}{2K_G\|A\|_{\infty\to 1}}$. In order to turn A into a matrix of rank $\leqslant r$ by changing entries, at least $\Omega(\frac{n^2}{r})$ entries in A must be reversed.*

Proof. Let $\tilde{B}$ be a matrix of rank r obtained by changing entries in A, and let $B = \mathrm{sign}(\tilde{B})$ be its sign matrix. Then

$$r \geqslant d(B) \geqslant \frac{n^2}{K_G\|B\|_{\infty\to 1}}.$$

The first inequality follows from the definition of dimension complexity and the latter from a general bound proved in [19]. It follows that

$$\|B\|_{\infty\to 1} \geqslant \frac{n^2}{rK_G}.$$

By assumption, $\|A\|_{\infty\to 1} \leqslant \frac{n^2}{2rK_G}$, and so Lemma 5.3 implies that the sign matrices A and B differ in at least $\Omega(\frac{n^2}{r})$ places. □

We turn to discuss rank rigidity when only bounded changes are allowed.

Definition ([21]). For A a sign matrix, and $\theta \geqslant 0$, define

$$R_A^+(r,\theta) = \min_B\{h(A,B) : \operatorname{rank}(B) \leqslant r,\ \forall_{i,j}\ 1 \leqslant |b_{ij}| \leqslant \theta\}.$$

Claim 5.7. *For every sign matrix A,*

$$R_A^+(\mathrm{mc}_r^2(A,l)/\theta^2,\theta) \geqslant l.$$

Proof. Let A be a sign matrix, and B a real matrix with $\theta \geqslant |b_{ij}| \geqslant 1$ for all i, j, and $\operatorname{rank}(B) \leqslant \mathrm{mc}_r^2(A,l)/\theta^2$. Denote the sign matrix of B by $\tilde{A}$. Then $h(A,\tilde{A}) \leqslant h(A,B)$. Also, it holds that

$$\mathrm{mc}(\tilde{A}) \leqslant \gamma_2(B) \leqslant \|B\|_\infty\sqrt{\operatorname{rank}(B)} \leqslant \mathrm{mc}_r(A,l).$$

The first inequality follows from equation (1.2). The second inequality follows since $\gamma_2(Z) \leqslant \sqrt{\operatorname{rank}(Z)}$ for every sign matrix Z. (This inequality is well known in Banach space theory. See, *e.g.*, [19] for a proof.) We conclude that $h(A,B) \geqslant h(A,\tilde{A}) \geqslant l$. Since this is true for every matrix B satisfying the assumptions, $R_A^+(\mathrm{mc}_r^2(A,l)/\theta^2,\theta) \geqslant l$. □

6. Discussion and open problems

It remains a major open problem to derive lower bounds on mc-rigidity. In particular, the following conjecture seems interesting and challenging.

Conjecture 6.1. *For every constant c_1 there are constants c_2, c_3 such that every $n \times n$ sign matrix A satisfying* $\mathrm{mc}(A) \geqslant c_1\sqrt{n}$ *also satisfies*

$$\mathrm{mc}_r(A, c_2n^2) \geqslant c_3\sqrt{n}.$$

This conjecture says that every matrix with high margin complexity has a high mc-rigidity as well. This is helpful since we do have general techniques for proving lower bounds on margin complexity, *e.g.*, [9, 19]. In particular, an $n \times n$ Hadamard matrix has margin complexity $\sqrt{n}$ ([9]). Thus, Conjecture 6.1 combined with Theorems 4.4 and 4.5 implies that $PH^{cc} \neq PSPACE^{cc}$ and $AM^{cc} \neq IP^{cc}$, since Sylvester–Hadamard matrices are in $IP^{cc} \cap PSPACE^{cc}$. The relation between margin complexity and discrepancy (Theorem 3.1) adds another interesting angle to these statements.

References

[1] Aho, A. V., Ullman, J. D. and Yannakakis, M. (1983) On notions of information transfer in VLSI circuits. In *Proc. 15th ACM STOC*, pp. 133–139.

[2] Alon, N. (1995) Tools from higher algebra.In *Handbook of Combinatorics*, Vol. 1, North-Holland, pp. 1749–1783.

[3] Alon, N., Frankl, P. and Rödl, V. (1985) Geometrical realizations of set systems and probabilistic communication complexity. In *Proc. 26th Symposium on Foundations of Computer Science*, IEEE Computer Society Press, pp. 277–280.

[4] Alon, N. and Naor, A. (2004) Approximating the cut-norm via Grothendieck's inequality. In *Proc. 36th ACM STOC*, pp. 72–80.

[5] Babai, L., Frankl, P. and Simon, J. (1986) Complexity classes in communication complexity. In *Proc. 27th IEEE Symposium on Foundations of Computer Science*, pp. 337–347.

[6] Ben-David, S., Eiron, N. and Simon, H. U. (2002) Limitations of learning via embeddings in Euclidean half spaces. *J. Machine Learning Research* **3** 441–461.

[7] Chazelle, B. (2000) *The Discrepancy Method: Randomness and Complexity*, Cambridge University Press.

[8] Cristianini, N. and Shawe-Taylor, J. (1999) *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press. New York.

[9] Forster, J. (2001) A linear lower bound on the unbounded error probabilistic communication complexity. In *SCT: Annual Conference on Structure in Complexity Theory*, IEEE Computer Society Press, pp. 100–106.

[10] Forster, J., Krause, M., Lokam, S. V., Mubarakzjanov, R., Schmitt, N. and Simon, H. U. (2001) Relations between communication complexity, linear arrangements, and computational complexity. In *Proc. 21st Conference on Foundations of Software Technology and Theoretical Computer Science*, pp. 171–182.

[11] Forster, J., Schmitt, N. and Simon, H. U. (2001) Estimating the optimal margins of embeddings in Euclidean half spaces. In *Proc. 14th Annual Conference on Computational Learning Theory* (COLT 2001) *and 5th European Conference on Computational Learning Theory* (EuroCOLT 2001), Vol. 2111 of *Lecture Notes in Computer Science*, Springer, Berlin, pp. 402–415.

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

[13] Jameson, G. J. O. (1987) *Summing and Nuclear Norms in Banach Space Theory*, London Mathematical Society Student Texts, Cambridge University Press.

[14] Johnson, W. B. and Lindenstrauss, J. (1984) Extensions of Lipshitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability*, AMS, Providence, RI, pp. 189–206.

[15] Kashin, B. and Razborov, A. (1998) Improved lower bounds on the rigidity of Hadamard matrices. *Mathematical Notes* **63** 471–475.

[16] Kremer, I., Nisan, N. and Ron, D. (1995) On randomized one-round communication complexity. In *Proc. 35th IEEE FOCS*, pp. 596–605.

[17] Kushilevitz, E. and Nisan, N. (1997) *Communication Complexity*, Cambridge University Press.

[18] Lee, T., Shraibman, A. and Špalek, R. (2008) A direct product theorem for discrepancy. In *Annual IEEE Conference on Computational Complexity*, pp. 71–80.

[19] Linial, N., Mendelson, S., Schechtman, G. and Shraibman, A. (2007) Complexity measures of sign matrices. *Combinatorica* **27** 439–463.

[20] Linial, N. and Shraibman, A. (2007) Lower bounds in communication complexity based on factorization norms. In *Proc. 39th ACM STOC*, pp. 699–708.

[21] Lokam, S. V. (1995) Spectral methods for matrix rigidity with applications to size-depth tradeoffs and communication complexity. In *IEEE Symposium on Foundations of Computer Science*, pp. 6–15.

[22] Matoušek, J. (1999) *Geometric Discrepancy: An Illustrated Guide*, Vol. 18 of *Algorithms and Combinatorics*, Springer.

[23] Matoušek, J. (2002) *Lectures on Discrete Geometry*, Vol. 212 of *Graduate Texts in Mathematics*, Springer.

[24] Mendelson, S. (2005) Embeddings with a Lipschitz function. *Random Struct. Alg.* **27** 25–45.

[25] Mendelson, S. (2005) On the limitations of embedding methods. In *Proc. 18th Annual Conference on Learning Theory* (COLT05), Vol. 3559 of *Lecture Notes in Computer Science*, Springer, pp. 353–365.

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

[27] Pisier, G. (1986) *Factorization of Linear Operators and Geometry of Banach Spaces*, Vol. 60 of *CBMS Regional Conference Series in Mathematics*, Published for the Conference Board of the Mathematical Sciences, Washington, DC.

[28] Pudlák, P. and Rödl, V. (1994) Some combinatorial–algebraic problems from complexity theory. *Discrete Math.* **136** 253–279.

[29] Sherstov, A. A. (2008) Communication complexity under product and nonproduct distributions. In *Annual IEEE Conference on Computational Complexity*, pp. 64–70.

[30] Shokrollahi, M. A., Spielman, D. A. and Stemann, V. (1997) A remark on matrix rigidity. *Inform. Process. Lett.* **64** 283–285.

[31] Tarui, J. (1993) Randomized polynomials, threshold circuits and polynomial hierarchy. *Theoret. Comput. Sci.* **113** 167–183.

[32] Valiant, L. G. (1977) Graph-theoretic arguments in low level complexity. In *Proc. 6th MFCS*, Vol. 53 of *Lecture Notes in Computer Science*, Springer, pp. 162–176.

[33] Vapnik, V. N. (1999) *The Nature of Statistical Learning Theory*, Springer, New York.

[34] Yao, A. (1983) Lower bounds by probabilistic arguments. In *Proc. 15th ACM STOC*, pp. 420–428.