On the Smallest Possible Dimension and the Largest Possible Margin of Linear Arrangements Representing Given Concept Classes *

Jürgen Forster and Hans Ulrich Simon

Fakultät für Mathematik, Ruhr-Universität Bochum, D-44780 Bochum, Germany
{forster,simon}@lmi.ruhr-uni-bochum.de

Abstract. This paper discusses theoretical limitations of classification systems that are based on feature maps and use a separating hyper-plane in the feature space. In particular, we study the embeddability of a given concept class into a class of Euclidean half spaces of low dimension, or of arbitrarily large dimension but realizing a large margin. New bounds on the smallest possible dimension or on the largest possible margin are presented. In addition, we present new results on the rigidity of matrices and briefly mention applications in complexity and learning theory.

1 Introduction

During the last decade, there has been a lot of interest in maximal margin classifiers. Learning algorithms that calculate the hyper-plane with the largest margin on a sample and use this hyper-plane to classify new instances have a relatively solid theoretical foundation and have shown excellent empirical performance (e.g., see [4, 17, 5]). Typically, the instances are mapped (implicitly when a kernel function is used) to some (possibly high-dimensional) feature space before the hyper-plane with maximal margin is calculated.

Recently, several authors [3, 8] started to study the theoretical limitations of classification systems that build on embeddings in feature spaces and apply linear separation afterwards. From an abstract point of view, one may represent a finite concept class, consisting of n Boolean concepts over a domain of size m, as a binary matrix $M \in \{-1,+1\}^{m \times n}$, where $M_{i,j} = -1$ indicates that instance i is a negative example for concept j (analogously for positive examples). We may also view matrix M as our prior guess which sign patterns we believe are appropriate for a specific classification task. Anyway, if the classification task can be handled by a classification system of the type described above, there must be an appropriate feature mapping that allows to reproduce the sign patterns in M (at least approximately) by the hypothesis class of half spaces in the feature space. This leads us to the following central

* This work has been supported in part by the ESPRIT Working Group in Neural, Computational Learning II, NeuroCOLT2, No. 27150 and by the Deutsche Forschungsgemeinschaft Grant SI 498/4-1.

Definition 1. *A* linear arrangement *representing a matrix* $M \in \mathbb{R}^{m \times n}$ *is given by collections of vectors* $u_1, \ldots, u_m, v_1, \ldots, v_n \in \mathbb{R}^d$ *of Euclidean length* $\|u_i\|_2 = \|v_j\|_2 = 1$ *such that* $\operatorname{sign}\langle u_i, v_j\rangle = \operatorname{sign} M_{i,j}$ *for all* i, j. *Parameter* d *is called the* dimension *of the arrangement,* $\min_{i,j}\langle u_i, v_j\rangle$ *is the* minimal margin *and* $\frac{1}{mn}\sum_{i,j}\langle u_i, v_j\rangle$ *the* average margin *realized by the arrangement.*

In this definition, the signum function $\operatorname{sign} : \mathbb{R} \to \mathbb{R}$ is given by

$$\operatorname{sign}(x) = \begin{cases} +1 \ , & x > 0 \ , \\ 0 \ , & x = 0 \ , \\ -1 \ , & x < 0 \ . \end{cases}$$

A few comments are in place here:

- The definition of a linear arrangement representing M depends on M only through the *sign pattern* $(\operatorname{sign} M_{i,j})$ of M. One might ask why we did not simply focus on matrices with entries from $\{-1, 0, +1\}$. It will turn out, however, that some of our dimension- or margin-bounds depend on M *not* only through the sign pattern of M. For this reason, we preferred the more liberal definition.
- We may view vectors u_i as points in $\mathbb{R}^d$ and vectors v_j as normal vectors of the (positive) homogeneous half spaces $\{x \in \mathbb{R}^d : \langle v_j, x\rangle > 0\}$ (or vice versa). The restriction to *homogeneous* half spaces has purely technical reasons and is obviously not essential.
- If $M_{i,j} = +1$, then point u_i should be placed in the (positive) half space with normal vector v_j. The analogous remark is valid for $M_{i,j} = -1$ and negative half spaces. Since we restricted u_i, v_j to vectors of unit length, the absolute value of the scalar product $\langle u_i, v_j\rangle$ is the geometric distance of the point u_i from the boundary of the half space with normal vector v_j (and this distance cannot be trivially increased by scaling).

It should become clear from these remarks that Definition 1 captures our original intention.

Like in the papers [3, 8], we are interested in linear arrangements with a "low" dimension or a "large" minimal (or average) margin, because in both cases one could invoke the usual VC-dimension machinery such as to guarantee small generalization errors. If no linear arrangement for M with low dimension or large margin exists, then this indicates that the underlying concept class will suffer from a large generalization error whenever we attempt to learn its concepts by a typical large margin classification system. It should also be mentioned that large lower bounds on the dimension imply small upper bounds on the minimal margin because a linear arrangement with a large minimal margin can be transformed into a linear arrangement of small dimension (and still large minimal margin) by random projection techniques from [14, 9, 1]. Finally, it should be mentioned that one can always find "trivial arrangements" of dimension $\min\{m, n\}$ with minimal margin $1/\min\{\sqrt{m}, \sqrt{n}\}$. However, this trivial embedding leads to poor bounds on the generalization error. The reader interested in a more detailed discussion is referred to [3, 8].

In [3], it was shown by means of counting arguments that almost all matrices of constant VC-dimension do not have linear arrangements being significantly better than the trivial arrangement. These results were complemented by results in [6, 8], where methods from functional analysis were used to show the existence or non-existence of nice linear arrangements for *concrete* matrices. In this paper, we complement these results further. In Section 3, we improve on an algebraic lower bound from [6] on the dimension of a linear arrangement of a matrix M in terms of its spectral norm. The new bound is given in terms of the singular values of M. In Section 4, we narrow the gap between the known lower and upper bounds on the smallest possible dimension of a linear arrangement representing Hadamard matrices. This is done by presenting (the first) non-trivial arrangements for these matrices. Section 5 discusses the issue of "matrix-rigidity": how many entries of a given matrix must be changed to lower its rank below a given threshold r? We introduce a new rigidity definition with applications in complexity and learning theory. Then, we present a new lower bound on the rigidity of matrices. Section 6 discusses new upper bounds on the largest possible margin realized by linear arrangements. The paper closes with Section 7, where some open problems are mentioned.

2 Preliminaries

We use the *singular value decomposition* of a matrix $A \in \mathbb{R}^{m \times n}$: Let r be the rank of A. Then there always exist matrices $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$ with orthonormal columns and nonnegative numbers $\sigma_1(A) \geq \cdots \geq \sigma_r(A) > 0$, called the *singular values* of A, such that $A = U \operatorname{diag}(\sigma_1(A), \ldots, \sigma_r(A)) V^\top$ (see [12]). We also write $\sigma_i(A) = 0$ for $i > r$.

The *spectral norm* $\|A\|$ of A is the largest singular value of A, i.e., $\|A\| = \sigma_1(A)$. The *Frobenius norm* is

$$\|A\|_{\mathrm{F}} = \left(\sum_{i,j} A_{i,j}^2 \right)^{1/2} .$$

It is well known (see, e.g., [10]) that

$$\|A\|_{\mathrm{F}}^2 = \sum_i \sigma_i(A)^2 .$$

It follows that $\|A\| \leq \|A\|_{\mathrm{F}}$.

The following two results play a central role in our proofs.

Theorem 1 (Fejer's Theorem, [12], Corollary 7.5.4). *A matrix $A \in \mathbb{R}^{m \times m}$ is positive semidefinite if and only if*

$$\sum_{i,\ell} A_{i,\ell} B_{i,\ell} \geq 0$$

for all positive semidefinite matrices $B \in \mathbb{R}^{m \times m}$.

Theorem 2 (Hoffman-Wielandt [11] and [10], Section 8.3). *Let A, B be matrices in $\mathbb{R}^{m\times n}$. Then*

$$\sum_i (\sigma_i(A) - \sigma_i(B))^2 \le \|A - B\|_{\mathrm{F}}^2 \ .$$

We will also make use of the following well known

Fact 1 There exists a d-dimensional linear arrangement representing a matrix $M \in \{-1,+1\}^{m\times n}$ if and only if there is a matrix $\widetilde{M} \in \mathbb{R}^{m\times n}$ of rank d with the same sign pattern as M.

3 An Improved Lower Bound on the Dimension

The following lower bound on the dimension of a linear arrangement was proven by Forster:

Theorem 3 ([6]). *Let $u_1, \dots, u_m, v_1, \dots, v_n \in \mathbb{R}^d$ be a linear arrangement representing a matrix $M \in \{-1,+1\}^{m\times n}$. Then $d \ge \sqrt{mn}/\|M\|$.*

This theorem bounds d from by below in terms of the spectral norm $\|M\| = \sigma_1(M)$. Thus, the bound involves only the largest singular value of M. In this section (building on a powerful lemma from [6]), we derive an improved lower bound that makes use of the full spectrum of singular values. Our main result is as follows:

Theorem 4. *Let $M \in \mathbb{R}^{m\times n}$ be a matrix satisfying $|M_{i,j}| \ge 1$ for all i, j. Let r be the rank and $\sigma_1(M) \ge \dots \ge \sigma_r(M)$ the singular values of M. For $d \le r$ let $u_1, \dots, u_m, v_1, \dots, v_n \in \mathbb{R}^d$, $\|u_i\|_2 = \|v_j\|_2 = 1$, be a linear arrangement representing M. Then*

$$d \cdot \sum_{i=1}^{d} \sigma_i^2(M) \ge mn \ . \tag{1}$$

Proof. It was shown in [6] that any d-dimensional linear arrangement $u_1, \dots, u_m$, $v_1, \dots, v_n$ representing a matrix M can be "normalized" such as to satisfy

$$\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \langle u_i, v_j \rangle^2 \ge \frac{1}{d} \ . \tag{2}$$

We may therefore assume (2) without loss of generality. Consider the matrix $\widetilde{M} \in \mathbb{R}^{m\times n}$ with entries $\widetilde{M}_{i,j} = \langle u_i, v_j \rangle \in \mathbb{R}^{m\times n}$. From the assumptions of the theorem and (2), the following properties are evident:

Property 1 $\operatorname{sign} \widetilde{M}_{i,j} = \operatorname{sign} M_{i,j}$ and $|\widetilde{M}_{i,j}| \le 1 \le |M_{i,j}|$. This implies that $M_{i,j}\widetilde{M}_{i,j} \ge \widetilde{M}_{i,j}^2$.

Property 2 $\|\widetilde{M}\|_{\mathrm{F}}^2 \ge mn/d$.

Property 3 The rank of $\widetilde{M}$ is bounded by $d \leq r$. This implies that $\sigma_i(\widetilde{M}) = 0$ for $i > d$.

Because of Properties 1 and 2, $\|M - \widetilde{M}\|_{\mathrm{F}}^2$ can be upper-bounded as follows:

$$\begin{aligned} \|M - \widetilde{M}\|_{\mathrm{F}}^2 &= \sum_{i,j} M_{i,j}^2 - 2M_{i,j}\widetilde{M}_{i,j} + \widetilde{M}_{i,j}^2 \\ &\leq \sum_{i,j} M_{i,j}^2 - 2\widetilde{M}_{i,j}^2 + \widetilde{M}_{i,j}^2 \\ &= \|M\|_{\mathrm{F}}^2 - \|\widetilde{M}\|_{\mathrm{F}}^2 \\ &= \sum_{i=1}^{r} \sigma_i^2(M) - \|\widetilde{M}\|_{\mathrm{F}}^2 \\ &\leq \sum_{i=1}^{r} \sigma_i^2(M) - \frac{mn}{d} \ . \end{aligned}$$

Using Property 3 and the Hoffman-Wielandt inequality, $\|M - \widetilde{M}\|_{\mathrm{F}}^2$ can be lower-bounded as follows:

$$\|M - \widetilde{M}\|_{\mathrm{F}}^2 \geq \sum_{i=1}^{d} (\sigma_i(M) - \sigma_i(\widetilde{M}))^2 + \sum_{i=d+1}^{r} \sigma_i^2(M) \geq \sum_{i=d+1}^{r} \sigma_i^2(M) \ .$$

We conclude that $\sum_{i=d+1}^{r} \sigma_i^2(M) \leq \sum_{i=1}^{r} \sigma_i^2(M) - mn/d$, which is equivalent to (1). •

Note that $d \cdot \sum_{i=1}^{d} \sigma_i^2(M) \geq mn$ implies

$$d^2\|M\|^2 = d^2\sigma_1^2(M) \geq d \cdot \sum_{i=1}^{d} \sigma_i^2(M) \geq mn \ ,$$

which, in turn, implies Forster's lower bound $d \geq \sqrt{mn}/\|M\|$. It follows that the new bound improves the old one. Note that both bounds coincide if and only if $\sigma_1(M) = \cdots = \sigma_d(M)$ for the smallest d that satisfies (1). If the first d singular values do not coincide, the new bound yields a strict improvement. The improvement is particularly strong if $\sigma_1(M)$ is relatively large compared to $\sigma_i(M)$ for $i \geq 2$. For instance, if an $(n \times n)$-matrix M with ± 1-entries and full rank n had the spectrum

$$\sigma_1^2(M) = n^{3/2} \text{ and } \sigma_2^2 = \cdots = \sigma_n^2(M) = \frac{n^2 - n^{3/2}}{n-1} < n \ ,$$

then Forster's lower bound from Theorem 3 would lead to $d \geq n^{1/4}$, whereas the new lower bound from Theorem 4 would lead to $d \geq n^{1/2}/2$ (as an easy evaluation shows).

4 Nontrivial Arrangements for Hadamard Matrices

Hadamard matrices H_n of size $2^n \times 2^n$ can be inductively defined as follows:

$$H_0 = (1)$$
$$H_{n+1} = \begin{pmatrix} H_n & H_n \\ H_n & -H_n \end{pmatrix}$$

Although the conjecture that each linear arrangement representing H_n requires $2^{\Omega(n)}$ Euclidean dimensions was long standing, it was not confirmed until recently [6]: since $\|H_n\| = 2^{n/2}$, an evaluation of Forster's lower bound from Theorem 3 yields that each linear arrangement representing H_n requires at least $2^{n/2}$ Euclidean dimensions. Clearly, 2^n Euclidean dimensions are enough because we may use the trivial embedding. To the best of our knowledge, no nontrivial linear arrangement for H_n is known so far. In this section, we present a non-trivial arrangement for H_n, thereby narrowing (but still not closing) the gap between $2^{n/2}$ and 2^n. For sake of simple exposition, we will restrict ourselves to even n.

Our main result of this section reads as follows:

Lemma 1. *For every even n, there is a matrix $\widetilde{H}_n \in \mathbb{R}^{2^n \times 2^n}$ of rank $3^{n/2}$ that has the same sign pattern as the Hadamard matrix H_n.*

Proof. Remember that the *Kronecker product* $A \otimes B$ of two matrices $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{p,q}$ is defined as

$$A \otimes B := \begin{pmatrix} A_{1,1}B & \cdots & A_{1,n}B \\ \vdots & & \vdots \\ A_{m,1}B & \cdots & A_{m,n}B \end{pmatrix} \in \mathbb{R}^{mp\times nq} \ .$$

The rank of the Kronecker product is $\operatorname{rank}(A \otimes B) = \operatorname{rank}(A) \cdot \operatorname{rank}(B)$. (See, e.g., [13], Theorem 4.2.15.) The Hadamard matrix H_n can be written as the n-th Kronecker power of the matrix H_1:

$$H_n = \begin{pmatrix} +1 & +1 \\ +1 & -1 \end{pmatrix}^{\otimes n} = \left(\begin{pmatrix} +1 & +1 \\ +1 & -1 \end{pmatrix} \otimes \begin{pmatrix} +1 & +1 \\ +1 & -1 \end{pmatrix} \right)^{\otimes \frac{n}{2}}$$
$$= \begin{pmatrix} +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & -1 & +1 \end{pmatrix}^{\otimes \frac{n}{2}} \ .$$

Obviously the matrix

$$\widetilde{H}_n = \begin{pmatrix} +1 & +1 & +1 & +1 \\ +1 & -1 & +5 & -1 \\ +1 & +5 & -1 & -1 \\ +1 & -1 & -1 & +5 \end{pmatrix}^{\otimes \frac{n}{2}}$$

has the same sign pattern as H_n. The rank of $\widetilde{H}_n$ is

$$\mathrm{rank}(\widetilde{H}_n) = \left(\mathrm{rank} \begin{pmatrix} +1 & +1 & +1 & +1 \\ +1 & -1 & +5 & -1 \\ +1 & +5 & -1 & -1 \\ +1 & -1 & -1 & +5 \end{pmatrix} \right)^{\frac{n}{2}} = 3^{\frac{n}{2}} \ .$$

•

Combining Fact 1 with Lemma 1, we arrive at the following result:

Corollary 1. *For every even n, there exists a $3^{n/2}$-dimensional linear arrangement representing H_n.*

5 Rigidity

The rigidity of a matrix, $\mathcal{R}_M(r)$, is the number of entries of M that must be changed to reduce its rank to at most r. Let $H \in \{-1,+1\}^{N \times N}$ be a Hadamard matrix. Lokam [16] showed that $\mathcal{R}_H(r) \geq N^2/(r+1)^2$. Kashin and Razborov [15] improved this result to $\mathcal{R}_H(r) = \Omega(N^2/r)$ for $r \leq N/2$.

We consider a different definition of the rigidity of a matrix here. The new definition distinguishes between sign-preserving and sign-non-preserving changes of entries. Only sign-non-preserving changes are counted. More formally: $\widetilde{\mathcal{R}}_M(r)$ is the number of entries of M that must be exposed to a sign-non-preserving change to reduce the rank of M to at most r, where we allow arbitrarily many sign-preserving changes as well. Clearly, $\mathcal{R}_M(r) \geq \widetilde{\mathcal{R}}_M(r)$. Thus, lower bounds on $\widetilde{\mathcal{R}}_M(r)$ are harder to obtain than lower bounds on $\mathcal{R}_M(r)$. The main goal of this section is to demonstrate that quite strong lower bounds on $\widetilde{\mathcal{R}}_M(r)$ can be derived (in a surprisingly simple manner) from Theorem 3. For ease of exposition, we restrict ourselves to matrices M without zero-entries in what follows.

We start with a somewhat technical result, which, loosely speaking, tells that zero-substitutions do not help in reducing the rank given that sign-preserving changes are for free.

Lemma 2. *If the rank of $M \in \mathbb{R}^{m \times n}$ can be reduced to r by changing s entries of M in an arbitrary fashion and the remaining entries in a sign-preserving fashion, then this can also be achieved without changing one of the entries to zero.*

Proof. Remember our general assumption that M has no zero-entries. Let $\widetilde{M}$ be a matrix of rank at most r which results from M by changing s entries in an arbitrary fashion (possibly to zero) and the remaining entries in a sign-preserving fashion. It follows from Fact 1 that there exist vectors $u_1, \ldots, u_m, v_1, \ldots, v_n \in \mathbb{R}^r$ such that $\mathrm{sign}\, \widetilde{M}_{i,j} = \mathrm{sign}\, \langle u_i, v_j \rangle$ for all i, j. Consider indices i, j of an entry of M that was changed to $\widetilde{M}_{i,j} = 0$. It follows that $\langle u_i, v_j \rangle = 0$, i.e., point u_i lies on the hyper-plane with normal vector v_j. Reversely, each point u_i located on a hyper-plane with normal vector v_j corresponds to an entry of M that was changed to $\widetilde{M}_{i,j} = 0$. We can slightly shift all points being located on hyperplanes such as to avoid scalar products with value zero and such as to preserve the sign of all scalar products with a value different from zero. In this manner, we obtain an r-dimensional linear arrangement of a matrix $\hat{M}$ without zero entries that is still obtained from M by changing s entries in an arbitrary fashion and the remaining entries in a sign-preserving fashion. Applying Fact 1 again, we conclude that the rank of $\hat{M}$ is at most r. •

The next result is a simple application of Theorem 3.

Lemma 3. *Let $M \in \{-1,+1\}^{m\times n}$ and let $\widetilde{M}$ be a matrix without zero entries that is obtained from M by changing s entries to arbitrary new values except zero and by performing (possibly vacuous) sign-preserving changes to the remaining entries. Then*

$$\operatorname{rank}(\widetilde{M}) \geq \frac{\sqrt{mn}}{\|M\| + 2\sqrt{s}} \ . \tag{3}$$

Proof. A change from value $M_{i,j} \in \{-1,+1\}$ to a new value $\widetilde{M}_{i,j} \neq 0$ can be simulated by first performing a (possibly vacuous) sign-change from $M_{i,j}$ to sign $\widetilde{M}_{i,j}$ and then performing a (cost-free) sign-preserving change from sign $\widetilde{M}_{i,j}$ to $\widetilde{M}_{i,j}$. Thus, $\widetilde{M}$ has the same sign pattern as a matrix of the form $M' = M - S \in \{-1,+1\}^{m\times n}$ for some matrix S with at most s entries from $\{-2,+2\}$ (representing sign-changes) and with zero-entries everywhere else. Thus,

$$\|M'\| \leq \|M\| + \|S\| \leq \|M\| + \|S\|_{\mathrm{F}} \leq \|M\| + 2\sqrt{s} \ .$$

Applying Theorem 3, the lemma follows. •

Because of Lemma 2, the definition of $\widetilde{\mathcal{R}}_M()$ is not affected when we do not allow zero-substitutions. Solving (3) for s leads therefore immediately to

Corollary 2. $\widetilde{\mathcal{R}}_M(r) \geq (\sqrt{mn}/r - \|M\|)^2/4$.

Corollary 2 applied to an $(N \times N)$ Hadamard matrix H with $\|H\| = \sqrt{N}$ yields

Corollary 3. $\widetilde{\mathcal{R}}_H(r) \geq (N^2/r^2 - 2N^{3/2}/r + N)/4$.

This bound is not much worse than the afore-mentioned bounds of Lokam or Kashin and Razborov, although we allowed a much more powerful rank-reduction procedure.

We finally would like to mention an application of Lemma 3. Let $N = 2^n$. It was shown in [7] that the Boolean function induced by the $(N \times N)$-Hadamard matrix H_n (depending on $2n$ Boolean variables) cannot be computed by a 2-layer threshold circuit C_n with polynomially bounded weights in the hidden layer unless the size of C_n grows exponentially in n. Using Lemma 3, it can be shown that this lower bound remains valid when the circuit must only "approximate" H_n in the sense that $2^{(2-\varepsilon)n}$ of the 2^{2n} function values may be computed incorrectly. In a similar fashion, it follows that there is no $d(n)$-dimensional linear arrangement representing an approximation of H_n unless d grows exponentially in n.

6 Upper Bounds on the Margin

Throughout this section, $M \in \{-1,+1\}^{m \times n}$ and $u_1, \ldots, u_m, v_1, \ldots, v_n$ denotes a linear arrangement representing M. The following bounds on the minimal margin (valid for any choice of $u_1, \ldots, u_m, v_1, \ldots, v_n$) are already known:

Theorem 5 **([6]).** $\min_{i,j} |\langle u_i, v_j \rangle| \le \|M\|/\sqrt{m \cdot n}$.

Theorem 6 **([8]).** *Let $\widetilde{M} \in \mathbb{R}^{m \times n}$ be a matrix with the same sign pattern as M. Then the following holds:*

$$\min_{i,j} |\langle u_i, v_j \rangle| \le \frac{\sqrt{m} \cdot \|\widetilde{M}\|}{\sqrt{\sum_j \left(\sum_i |\widetilde{M}_{i,j}| \right)^2}} \quad . \tag{4}$$

In the special case $\widetilde{M} = M$, the bound from Theorem 6 coincides with the bound from Theorem 5. However, as demonstrated in [8], the bound in Theorem 6 can be much stronger than the bound in Theorem 5 if $\widetilde{M}$ is cleverly chosen.

It had been observed by Ben-David [2] that the proof of Theorem 5 in [6] implicitly shows a stronger result: the upper bound from this theorem also applies to the average margin. See Corollary 5 below. Although the essential arguments for this observation can already be found in [6], it will be much more convenient to have a proof that addresses the concept of an average margin in a more explicit manner. To this end, we will prove a useful intermediate result (Lemma 4 below) in this section that serves two purposes. First, it leads to a simple proof of Corollary 5. Second, it leads to a bound that is similar to the bound from Theorem 6, but has a somewhat simpler form. See Corollary 4 below.

Lemma 4. *For any matrix $\widetilde{M} \in \mathbb{R}^{m \times n}$, the following holds:*

$$\sum_{i,j} \widetilde{M}_{i,j} \langle u_i, v_j \rangle \le \sqrt{m \cdot n} \|\widetilde{M}\| \quad .$$

Proof. For every j we have that

$$\sum_i \widetilde{M}_{i,j} \langle u_i, v_j \rangle = \left\langle \sum_i \widetilde{M}_{i,j} u_i, v_j \right\rangle \overset{\|v_j\|_2 = 1}{\le} \|\sum_i \widetilde{M}_{i,j} u_i\|_2 \ , \tag{5}$$

where we used the Cauchy-Schwartz inequality. The lemma can now be derived as follows:

$$\begin{aligned}
\frac{1}{n}\left(\sum_{i,j}\widetilde{M}_{i,j}\langle u_i,v_j\rangle\right)^2 &\le \sum_j\left(\sum_i \widetilde{M}_{i,j}\langle u_i,v_j\rangle\right)^2 \\
&\stackrel{(5)}{\le} \sum_j\left\langle \sum_i \widetilde{M}_{i,j}u_i, \sum_\ell \widetilde{M}_{\ell,j}u_\ell\right\rangle \\
&= \sum_{i,\ell}\left(\sum_j \widetilde{M}_{i,j}\widetilde{M}_{\ell,j}\right)\langle u_i,u_\ell\rangle \\
&= \sum_{i,\ell}\left(\widetilde{M}\widetilde{M}^\top\right)_{i,\ell}\langle u_i,u_\ell\rangle \\
&\stackrel{(*)}{\le} \sum_{i,\ell}\left(\|\widetilde{M}\|^2\,\mathrm{I}_m\right)_{i,\ell}\langle u_i,u_\ell\rangle \\
&= \|\widetilde{M}\|^2\sum_i\|u_i\|_2^2 = m\cdot\|\widetilde{M}\|^2
\end{aligned}$$

The first inequality is obtained when the Cauchy-Schwartz inequality is applied to the vectors $a, b \in \mathbb{R}^n$ given by $a_j = \sum_i \widetilde{M}_{i,j}\langle u_i, v_j\rangle$ and $b_j = 1$. Inequality $(*)$ holds because the matrices $A := \|\widetilde{M}\|^2\,\mathrm{I}_m - \widetilde{M}\widetilde{M}^\top$ and $B := (\langle u_i, u_\ell\rangle)_{i,\ell}$ are positive semidefinite. Therefore $\sum_{i,\ell} A_{i,\ell}B_{i,\ell} \ge 0$ because of Fejer's Theorem 1. •

Corollary 4. *If $\widetilde{M}$ has the same sign pattern as M, then*

$$\min_{i,j}|\langle u_i,v_j\rangle| \le \frac{\sqrt{m\cdot n}\,\|\widetilde{M}\|}{\sum_{i,j}|\widetilde{M}_{i,j}|}\ . \tag{6}$$

Proof. Note that $\langle u_i, v_j\rangle$ has the same sign as $\widetilde{M}_{i,j}$. This implies that $\widetilde{M}_{i,j}\langle u_i, v_j\rangle = |\widetilde{M}_{i,j}|\cdot|\langle u_i, v_j\rangle|$. The corollary now follows from Lemma 4. •

Corollary 5. $\frac{1}{m\cdot n}\sum_{i,j}|\langle u_i, v_j\rangle| \le \frac{\|M\|}{\sqrt{m\cdot n}}$.

Proof. Note that $\langle u_i, v_j\rangle$ has the same sign as $M_{i,j} \in \{-1, 1\}$. This implies that $M_{i,j}\langle u_i, v_j\rangle = |\langle u_i, v_j\rangle|$. The corollary now follows from Lemma 4 •

Note that (6) is weaker than (4) since it can alternatively be derived from (4) and the Cauchy-Schwarz inequality. However, the bound seems more natural in the simpler form (6). For instance, (6) does not change if we replace $\widetilde{M}$ by its transpose $\widetilde{M}^\top$. Furthermore, it exhibits a comparably simple dependence on $\widetilde{M}$.

7 Open Problems

We would like to close the gap between the lower and the upper bound on the dimension of linear arrangements representing Hadamard matrices. It would be nice to find natural concept classes that exhibit the typical spectrum of singular values that makes our new lower bound on the dimension (depending on *all* singular values) superior to the old lower bound (depending on the *largest* singular value only). The hardest open problem might be to characterize the smallest possible dimension or the largest possible margin of a linear arrangement for an arbitrary given matrix M by general bounds that are (at least approximately) tight. The issue of linear arrangements that *approximately* represent a given matrix was briefly considered in Section 5. It certainly deserves the right of further investigation.

Acknowledgments Thanks to Shai Ben-David for telling us his observation that Forster's paper [6] implicitly shows an upper bound on the *average* margin. Thanks to Dietrich Braess for pointing us to simplifications based on the clever usage of Fejer's Theorem.

References

1. Rosa I. Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *Proceedings of the 40'th Annual Symposium on the Foundations of Computer Science*, pages 616–623, 1999.
2. Shai Ben-David. Personal Communication.
3. Shai Ben-David, Nadav Eiron, and Hans Ulrich Simon. Limitations of learning via embeddings in Euclidean half spaces. In *Proceedings of the 14th Annual Workshop on Computational Learning Theory*, pages 385–401. Springer Verlag, 2001.
4. Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.
5. Nello Christianini and John Shawe-Taylor. *An Introduction to Support Vector Machines.* Cambridge University Press, 2000.
6. Jürgen Forster. A linear bound on the unbounded error probabilistic communication complexity. In *Proceedings of the 16th Annual Conference on Computational Complexity*, pages 100–106, 2001.
7. Jürgen Forster, Matthias Krause, Satyanarayana V. Lokam, Rustam Mubarakzjanov, Niels Schmitt, and Hans U. Simon. Relations between communication complexity, linear arrangements, and computational complexity. In *Proceedings of the 21'st Annual Conference on the Foundations of Software Technology and Theoretical Computer Science*, pages 171–182, 2001.
8. Jürgen Forster, Niels Schmitt, and Hans Ulrich Simon. Estimating the optimal margins of embeddings in Euclidean half spaces. In *Proceedings of the 14th Annual Workshop on Computational Learning Theory*, pages 402–415. Springer Verlag, 2001.
9. P. Frankl and H. Maehara. The Johnson-Lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory (B)*, 44:355–362, 1988.

10. Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* The John Hopkins University Press, 1991.
11. A. J. Hoffman and H. W. Wielandt. The variation of the spectrum of a normal matrix. *Duke Mathematical Journal*, 20:37–39, 1953.
12. Roger A. Horn and Charles R.Johnson. *Matrix Analysis.* Cambridge University Press, 1985.
13. Roger A. Horn and Charles R.Johnson. *Topics in Matrix Analysis.* Cambridge University Press, 1991.
14. W. B. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert spaces. *Contemp. Math.*, 26:189–206, 1984.
15. B. S. Kashin and A. A. Razborov. Improved lower bounds on the rigidity of Hadamard matrices. *Mathematical Notes*, 63(4):471–475, 1998.
16. Satyanarayana V. Lokam. Spectral methods for matrix rigidity with applications to size-depth tradeoffs and communication complexity. In *Proceedings of the 36th Symposium on Foundations of Computer Science*, pages 6–15, 1995.
17. Vladimir Vapnik. *Statistical Learning Theory.* Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, 1998.