A Close Look to Margin Complexity and Related Parameters

Michael Kallweit
Fakultät für Mathematik
Ruhr-Universität Bochum
D-44780 Bochum, Germany
michael.kallweit@rub.de

Hans Ulrich Simon
Fakultät für Mathematik
Ruhr-Universität Bochum
D-44780 Bochum, Germany
hans.simon@rub.de

Abstract

Concept classes can canonically be represented by sign-matrices, i.e., by matrices with entries 1 and -1. The question whether a sign-matrix (concept class) A can be learned by a machine that performs large margin classification is closely related to the "margin complexity" associated with A. We consider several variants of margin complexity, reveal how they are related to each other, and we reveal how they are related to other notions of learning-theoretic relevance like SQ-dimension, CSQ-dimension, and the Forster bound.

1 Introduction

Large margin classifiers implicitly use a feature map that transforms linearly inseparable data into feature vectors that can be linearly separated in feature space so as to achieve a (hopefully) large margin, which then leads to a small generalization error. Concept classes $\mathcal{C}$ over a domain $\mathcal{X}$ that can potentially be learned by large margin classifiers must therefore admit a linear arrangement consisting of hyperplanes and points (with the hyperplanes representing the concepts from $\mathcal{C}$ and the points representing the instances from $\mathcal{X}$) such that positive (resp. negative) examples appear as points lying in a positive (resp. negative) halfspace and having a certain "safety distance" to the corresponding separating hyperplane. In practice, a large "hard" margin cannot often be achieved so that softer notions of a margin come into play. Soft margins can be achieved by arrangements which occasionally put points close to the separating hyperplane (small margin) or, may be, even in the wrong half-space (negative margin). But one would still insist on something like a large "average margin". This will (roughly) be captured by our notion of average margin complexity.

In this paper, we deal with sign-matrices (which represent finite concept classes: every column is a Boolean function and the rows correspond to the instances), and we study various notions of (average) margin complexity, where "high (average) margin complexity" means that even the best arrangement achieves a small (average) margin only. Sign-matrices with high average margin complexity represent concept classes that cannot be successfully learned by large margin classifiers (thereby indicating the limitations of this approach). In a seminal paper, Forster (2002) presented a lower bound on the margin complexity (and on the dimension complexity which, however, is not considered in this paper) of a sign-matrix in terms of its spectral norm. Loosely speaking, the Forster bound measures the "amount of orthogonality" that is contained in A. It achieves its maximal value for Hadamard matrices. In this paper, we generalize the Forster bound by imposing probability distributions on the rows and the columns of A. In case of uniform distributions, the generalized bound collapses to the original-one. It is easy to construct matrices for which the original bound evaluates to a small number but, when the probability distributions are chosen properly, the generalized bound becomes large.

The SQ model of learning was introduced by Kearns (1998). It is an elegant abstraction from the PAC learning model of Valiant (1984). In this model, instead of having direct access to random examples (as in the PAC learning model) the learner obtains information about random examples via an oracle that provides estimates of various statistics about the unknown concept. Kearns showed that any learning algorithm that is successful in the SQ model can be converted, without much loss of efficiency, into a learning algorithm that is successful in the PAC learning model despite noise uniformly applied to the class labels of the examples. Furthermore, almost all concept classes known to be efficiently learnable in the PAC learning model can efficiently be learned in the SQ model too.

This is why the SQ model attracted a lot of attention in the Computational Learning Community. "Correlational Statistical Queries" are statistical queries of a special form and lead in the obvious way to the CSQ model of learning. As shown by Bshouty and Feldman (2002), the two models coincide in case of a fixed distribution, but, as shown by Feldman (2008), the SQ model is exponentially more powerful in the distribution-independent setting. Blum et al. (2003) have shown that the number of statistical queries needed for weak SQ-learning under a fixed distribution is polynomially related to the SQ-dimension (defined w.r.t. the same distribution). Feldman (2008) has defined the CSQ-dimension and has shown that it plays a similar role for distribution-independent weak learning in the CSQ model. In the same paper, he shows furthermore that CSQ-learnability is equivalent to evolvability (a framework introduced by Valiant (2009) and designed so as to catch the computational aspects of evolution).

In this paper, we will be concerned with the relations that hold between the various notions of margin complexity on one hand and parameters like SQ-dimension or CSQ-dimension on the other hand. The main results are as follows:[1]

- By means of semi-definite programming duality, we show in Section 3 that the optimal margin (the smallest distance between one of the points and one of the hyperplanes in a margin-optimal arrangement) coincides with the optimal average margin (the average distance between points and hyperplanes in an optimal arrangement) provided that the underlying distribution (according to which the average is taken) is chosen in a worst-case fashion. More formally:
$$\mathrm{mc}(A) = \max_Y \overline{\mathrm{mc}}_Y(A)$$
- In Section 3.1, we complement the well-known lower bound $\sqrt{mn}/\|A\|_2$ on the average margin complexity (w.r.t. uniform distributions on the rows and the columns of A) by the upper bound $mn/\|A\|_{tr}$. More formally:
$$\frac{\sqrt{mn}}{\|A\|_2} \le \overline{\mathrm{mc}}(A) \le \frac{mn}{\|A\|_{tr}}$$
- In Section 3.2, we identify two families of matrices whose average margin complexity (w.r.t. uniform distributions on the rows and the columns) is determined exactly: Hadamard matrices and matrices composed of all reflections of a single Boolean function.
- In Sections 4, 5, and 6, we determine relations between the various notions of margin complexity, the various versions of the Forster bound, the SQ-dimension and the CSQ-dimension. Here is a quick overview over our results:
  - Let p, q denote vectors assigning probabilities to the rows and the columns of a matrix A, respectively. We show that the SQ-dimension w.r.t. p of a sign-matrix A is polynomially related to the generalized Forster bound and also polynomially related to the average margin complexity of A according to
  $$\mathrm{SQdim}_p(A) < 2 \cdot \max_q \mathrm{FB}_{p,q}(A)^2 \le 2 \cdot \max_q \overline{\mathrm{mc}}_{p,q}(A)^2 < 2 \cdot \mathrm{SQdim}_p(A) \cdot (\mathrm{SQdim}_p(A) + 1)^2 \ .$$
  - We reveal the following polynomial relationship between the CSQ-dimension of a matrix $A \in \mathbb{R}^{m \times n}$ and the margin complexity of A:
  $$\mathrm{mc}(A) \le \mathrm{CSQdim}(A)^{1.5} \ \text{ and } \ \mathrm{CSQdim}(A) \le \lceil 32 \ln(4mn) \cdot \mathrm{mc}(A)^2 \rceil$$
  - We show that
  $$\mathrm{SQdim}(A^T) < 2 \cdot \mathrm{SQdim}(A) \cdot (\mathrm{SQdim}(A) + 1)^2 \ .$$
  This improves on $\mathrm{SQdim}(A^\top) \le 32 \cdot \mathrm{SQdim}(A)^4$, a result that had been shown before by Sherstov (2008).
  - We show that the generalized Forster bound is, up to a polynomial, not more effective than simply applying the classical bound to a properly chosen sub-matrix A'' of A. More formally:
  $$\max_{A''} \mathrm{FB}(A'') \le \max_{p,q} \mathrm{FB}_{p,q}(A) < 64 \cdot (1 + o(1)) \cdot \max_{A''} \mathrm{FB}(A'')^9$$

Although we are mainly interested in the study of sign-matrices, most of our notions and results deal with real-valued matrices because we do not want to impose unnecessary restrictions. A notable exception are the results in Section 5 which hold for sign-matrices only.

[1]The formal definitions needed for a precise understanding of the following statements are given in Section 2.

2 Definitions, Notations, and Facts

In this section, we provide the reader with the definitions and facts which will play a central role in the course of this paper.

Vectors, Matrices, and Norms: The all-ones vector in a finite-dimensional Euclidean space is simply denoted $\vec{1}$. The vector with value 1 in component k and zeros elsewhere is denoted $\vec{e}_k$. The $(d \times d)$ identity matrix is denoted by I_d or simply by I. Whenever the notation hides the dimension, say d, of the underlying Euclidean space, then d will be clear from context. The *Hadamard product* of two matrices A, B yields the matrix $A \circ B = (a_{i,j} b_{i,j})$, i.e., the matrices are multiplied entrywise. With $\mathrm{diag}(p)$ we denote the diagonal matrix build up from a vector p (i.e. the components of p are on the main diagonal and the remaining is zero). The *trace norm* of $A \in \mathbb{R}^{m \times n}$, denoted $\|A\|_{tr}$, is defined as the sum of all singular values of A. Let $\|\cdot\|$ denote a vector norm. The notation $\|A\|$ is understood as the norm of the mn-dimensional vector that results by concatenating the n m-dimensional columns of A so as to form a single mn-dimensional vector. For example, the Euclidean norm applied to a matrix yields

$$\|A\|_2 = \sqrt{\sum_{i,j} A_{i,j}^2} \ ,$$

and this is sometimes called the *Frobenius norm* of A. The *operator norm* associated with $\|\cdot\|$ is given by

$$\|A\| = \max_{\|v\|=1} \|Av\| = \max_{\|v\| \le 1} \|Av\| \ .$$

For example, the operator norm associated with the Euclidean norm is given by

$$\|A\|_2 = \max_{\|v\|_2=1} \|Av\|_2 = \max_{\|v\|_2 \le 1} \|Av\|_2 \ ,$$

and this is sometimes called the *spectral norm* of A. We remind the reader to the following facts:

$$\|A\|_2 = \|A^\top\|_2 \ , \ \|AA^\top\|_2 = \|A^\top A\|_2 = \|A\|_2^2 \ , \ \|A\|_2 \le \|A\|_2$$
$$\|A\|_2 = \max_{\|u\|_2=\|v\|_2=1} u^\top Av = \max_{\|u\|_2, \|v\|_2 \le 1} u^\top Av \tag{1}$$

By viewing matrices as vectors, we may consider the inner product of two matrices. An inner product without further specification refers to the standard scalar product. For example, $\langle A, B\rangle = \sum_{i,j} A_{i,j} B_{i,j}$. The dual of a norm $\|\cdot\|$ is given by

$$\|u\|^* = \max_{\|v\| \le 1} \langle u, v\rangle \ .$$

For example, L_∞ is the dual of L_1, and the trace norm is the dual of the spectral norm. It is well known that $\|\cdot\|^{**} = \|\cdot\|$, i.e., twofold dualization gives the original norm. Furthermore, for two norms $\|\cdot\|_1$, $\|\cdot\|_2$ and every $c > 0$, we have

$$\|\cdot\|_1 \le c \cdot \|\cdot\|_2 \Leftrightarrow \|\cdot\|_1^* \ge \frac{\|\cdot\|_2^*}{c} \ .$$

SQ- and CSQ-Dimension: Let p be a probability vector, i.e., p has non-negative components that sum up to 1. Consider the inner product

$$\langle x, y\rangle_p := \sum_i p_i x_i y_i \ .$$

A collection of vectors $u_1, \ldots, u_d$ is said to be *almost p-orthogonal* if

$$\forall k \ne l \in \{1, \ldots, d\} : |\langle u_k, u_l\rangle_p| \le \frac{1}{d} \ .$$

The *SQ-dimension of a matrix $A \in \mathbb{R}^{m \times n}$ w.r.t. p* is given by

$$\mathrm{SQdim}_p(A) = \max\{d \in \{1, \ldots, n\} : \text{ there exist } d \text{ almost } p\text{-orthogonal column vectors in } A\} \ .$$

The *SQ-dimension of A* is given by

$$\mathrm{SQdim}(A) = \max_p \mathrm{SQdim}_p(A) \ .$$

A collection of (not necessarily different) vectors $h_1, \ldots, h_d \in [-1, 1]^m$ is said to be *universally correlated with $A \in \mathbb{R}^{m \times n}$* if, for every m-dimensional probability vector p and every $j \in \{1, \ldots, n\}$, there exists $k \in \{1, \ldots, d\}$ such that $|\langle h_k, A_j\rangle_p| \ge 1/d$. The *CSQ-dimension* of A is given by

$$\mathrm{CSQdim}(A) = \min\{d : \text{ there exists a collection of } d \text{ vectors that is universally correlated with } A\} \ .$$

Margin Complexity: A *d-dimensional (homogeneous linear) arrangement for a matrix* $A \in \mathbb{R}^{m\times n}$ is given by vectors

$$u_1,\ldots,u_n; v_1,\ldots,v_m \in \mathbb{R}^d$$

whose Euclidean norm is bounded by 1. With an arrangement $\mathcal{A} = (u_1,\ldots,u_n; v_1,\ldots,v_m)$ for matrix A, we associate the *margin parameters*

$$\gamma_{i,j}(A|\mathcal{A}) = \langle u_i, v_j\rangle \cdot A_{i,j} \ .$$

The *margin complexity of* A is given by

$$\mathrm{mc}(A) = \left(\max_{\mathcal{A}} \min_{i,j} \gamma_{i,j}(A|\mathcal{A})\right)^{-1} \ .$$

It is easy to see that, for every matrix A without zero-entries (the matrices we are mainly interested in), there is always an arrangement that makes all margin parameters strictly positive, which implies that the margin complexity of A is strictly positive and finite. As outlined already by Linial et al. (2007) and Lee and Shraibman (2009), the so-called γ_2-norm and its dual, γ_2^*, are related to margin complexity as follows. Let $r(M)$ denote the largest Euclidean norm of a row of the matrix M. With this notation γ_2 and γ_2^*, satisfy the following equations (which, for the purpose of this paper, may also serve as a definition of these norms):

$$\gamma_2(A) = \min_{A=XY^\top} r(X)r(Y) \ \ \text{and} \ \ \gamma_2^*(A) = \max_{\mathcal{A}} \sum_{i,j} \gamma_{i,j}(A|\mathcal{A}) \tag{2}$$

Thus, $\gamma_2^*(A)$ is basically the largest "total margin" that can be achieved by an arrangement for A.[2] Let $Y = (y_{i,j})$ be an $(m \times n)$-dimensional matrix with non-negative entries that sum up to 1. The *Y-average margin complexity of* A is given by

$$\overline{\mathrm{mc}}_Y(A) = (\gamma_2^*(Y \circ A))^{-1} \ .$$

In the special case where $y_{i,j} = p_i q_j$ for two probability vectors p, q (so that $Y = p\cdot q^\top$), we introduce the notations

$$\overline{\mathrm{mc}}_{p,q}(A) := \overline{\mathrm{mc}}_{pq^\top}(A) \ , \ \overline{\mathrm{mc}}_p(A) := \overline{\mathrm{mc}}_{p,\vec{1}/n}(A) \ , \ \overline{\mathrm{mc}}(A) := \overline{\mathrm{mc}}_{\vec{1}/m,\vec{1}/n}(A)$$

so that

$$\overline{\mathrm{mc}}(A) = \left(\frac{1}{mn}\cdot\gamma_2^*(A)\right)^{-1} = \frac{mn}{\gamma_2^*(A)} \ . \tag{3}$$

Note that vector $\vec{1}/n$ (resp. $\vec{1}/m$) makes the columns (resp. rows) of A uniformly distributed. Considering the "smallest p-average margin" leads to the following definition:

$$\overline{\mathrm{mc}}_{p,MIN}(A) = \left(\max_{\mathcal{A}} \min_{j} \sum_i p_i \gamma_{i,j}(A|\mathcal{A})\right)^{-1} \ .$$

The various margin complexities are obviously related as follows:

$$\begin{aligned} \overline{\mathrm{mc}}(A) \le \max_p \overline{\mathrm{mc}}_p(A) \quad &\le \quad \max_{p,q} \overline{\mathrm{mc}}_{p,q}(A) \le \max_Y \overline{\mathrm{mc}}_Y(A) \le \mathrm{mc}(A) \\ \max_q \overline{\mathrm{mc}}_{p,q}(A) \quad &\le \quad \overline{\mathrm{mc}}_{p,MIN}(A) \end{aligned}$$

We will argue later that $\max_Y \overline{\mathrm{mc}}_Y(A) = \mathrm{mc}(A)$ and $\max_q \overline{\mathrm{mc}}_{p,q}(A) = \overline{\mathrm{mc}}_{p,MIN}(A)$, but for the remaining inequalities the gap between the smaller and larger parameter can be exponentially large.

Some Variants of the Forster bound: It was shown by Forster and Simon (2006) that, for every $A \in \mathbb{R}^{m\times n}$,

$$\overline{\mathrm{mc}}(A) \ge \frac{\sqrt{mn}}{\|A\|_2} \ . \tag{4}$$

As for probability vectors p, q, we introduce the following notational convention: P and Q are defined as the diagonal matrices containing the components of p and q , respectively. That is $P := \mathrm{diag}(p)$ and $Q := \mathrm{diag}(q)$. But keep in mind that this convention is *not* applied to letters different from P

[2] Note that the arrangement that maximizes the total margin may have some negative individual margin parameters.

and Q. Let A be a real-valued matrix with m rows and n columns. Consider the following variant of the Forster bound:

$$\mathrm{FB}_{p,q}(A) = \frac{1}{\|P^{1/2}AQ^{1/2}\|_2}$$

For $q = \vec{1}/n$, we simply write $\mathrm{FB}_p(A)$ instead of $\mathrm{FB}_{p,\vec{1}/n}$. For this choice of q, $Q = \frac{1}{n}I_n$ and we obtain

$$\mathrm{FB}_p(A) = \frac{\sqrt{n}}{\|P^{1/2}A\|_2} \ .$$

Similarly, if $p = \vec{1}/m$, we simply write $\mathrm{FB}(A)$ instead of $\mathrm{FB}_p(A)$. For this choice of p, $P = \frac{1}{m}I_m$ and we obtain

$$\mathrm{FB}(A) = \frac{\sqrt{mn}}{\|A\|_2} \ ,$$

which is the "classical" Forster bound in (4). Here, and in what follows, we use the notation A' to indicate a sub-matrix of A that is formed by a subset of the columns of A. Let $n(A') \le n$ denote the number of columns in A'. Note that

$$\max_{A'} \mathrm{FB}_p(A') \le \max_q \mathrm{FB}_{p,q}(A)$$

because $\mathrm{FB}_{p,q}$ collapses to $\mathrm{FB}_p(A')$ when the components of q are either 0 or $1/n(A')$, and the non-zero components are in one-to-one correspondence to the columns of A that are used to build A'.

Semidefinite Programming (SDP): We write $A \succeq B$ iff $A - B$ is a symmetric positive semi-definite matrix. The following definitions and facts about semi-definite programming are taken from Alizadeh (1995). A *standard primal SDP* is an optimization problem of the following form:

$$\min_X \langle C, X\rangle \ \ \textbf{s.t.} \ \ \forall \rho = 1,\ldots,r : \langle A_\rho, X\rangle = b_\rho \ , X \succeq 0 \tag{5}$$

Here, the matrices C, A_i are assumed to be symmetric. As in Linear Programming there is a duality theory for SDPs. The variables for the dual are denoted $y_1, \ldots, y_r$ (one dual variable per equality-constraint in the primal). We say that the equality-constraints of the primal *induce* the matrix $\sum_{\rho=1}^r y_\rho A_\rho$. The dual of (5) looks as follows:

$$\max_y \langle b, y\rangle \ \ \textbf{s.t.} \ \ C - \sum_{\rho=1}^{r} y_\rho A_\rho \succeq 0$$

If the optimal values of the primal and dual are equal, we achieve "strong duality". Among the well-known sufficient conditions for strong duality is the following-one (where "SCQ" means "Slater's Constraint Qualification").

SCQ: There exists y such that $\sum_{\rho=1}^r y_\rho A_\rho$ is (strictly) positive definite.

Note that non-negativity constraints for individual variables w_i can be expressed within a constraint of the form $X \succeq 0$ because the matrix X could be of the form $\begin{bmatrix} X' & 0 \\ 0 & \mathrm{diag}(w_1,\ldots,w_s) \end{bmatrix}$. We may therefore liberalize our definition of a standard primal SDP and allow constraints of the form $w_i \ge 0$.

3 Margin Maximization and its Dual

The fact that the optimal margin can be computed in polynomial time using semi-definite programming (SDP) had been observed first by Linial et al. (2007). In this section, we make use of this observation and express several variants of margin optimization as instances of SDP. Throughout this section, A, M are $(m \times n)$-matrices, X is an $(m+n) \times (m+n)$-matrix containing variables of the primal SDP, index i ranges from 1 to m, index j ranges from 1 to n, and index k ranges from 1 to $m+n$.

We call into mind the fact that a semi-definite matrix X can be written in the form $X = W^\top \cdot W$ (e.g., Cholesky-decomposition). If X has $m+n$ rows and columns, respectively, then W has, say, d rows and $m+n$ columns. Let $W = [U \ V]$ be the decomposition of W with U containing the first m columns. Then,

$$W^\top \cdot W = [U \ V]^\top \cdot [U, V] = \begin{bmatrix} U^\top U & U^\top V \\ (U^\top V)^\top & V^\top V \end{bmatrix} \ .$$

Imposing constraints like $X_{i,i} = \langle U_i, U_i\rangle = 1$, $X_{m+j,m+j} = \langle V_j, V_j\rangle = 1$, we can view X as a representation of an arrangement $\mathcal{A}$ given by the columns of U and the columns of V. Note that $\gamma_{i,j}(A|\mathcal{A}) = A_{i,j}\langle U_i, V_j\rangle = A_{i,j}X_{i,m+j}$. This is why many variants of margin-maximization problems can be expressed as instances of SDP. The following results are applications of strong SDP-duality.

Theorem 1 *For every $A \in \mathbb{R}^{m\times n}$:* $\mathrm{mc}(A) = \max_Y \overline{\mathrm{mc}}_Y(A)$.

Proof: We will prove the equivalent statement

$$\max_{\mathcal{A}} \min_{i,j} \gamma_{i,j}(A|\mathcal{A}) = \min_Y \gamma_2^*(Y\circ A) \stackrel{(2)}{=} \min_Y \max_{\mathcal{A}} \sum_{i,j} \gamma_{i,j}(Y\circ A|\mathcal{A}) \ . \tag{6}$$

Finding an arrangement $\mathcal{A}$ for $M = Y\circ A$ that maximizes $\sum_{i,j}\gamma_{i,j}(M|\mathcal{A})$ can be expressed as a standard SDP-problem (with optimal value $\gamma_2^*(Y\circ A)$) as follows:

$$\min_X -\frac{1}{2}\cdot\sum_{i,j} M_{i,j}(X_{i,m+j}+X_{m+j,i}) \ \textbf{ s.t. } \ \forall k: X_{k,k}=1 \text{ and } X\succeq 0 \tag{7}$$

There are $m+n$ equality constraints, which leads to dual variables $y_1,\ldots,y_{m+n}$. The matrix induced by the equality-constraints equals $\mathrm{diag}(y_1,\ldots,y_{m+n})$. Obviously, condition SCQ is satisfied so that we have strong duality. The cost matrix of the primal is given by

$$C = \frac{1}{2}\cdot\begin{bmatrix} 0 & -M \\ -M^\top & 0\end{bmatrix}$$

Thus, the dual problem (with variables $-y_k/2$ substituted for y_k and $Y\circ A$ substituted for M) looks as follows:

$$\min_y \frac{1}{2}\cdot\sum_k y_k \ \textbf{ s.t. } \ \underbrace{\begin{bmatrix} \mathrm{diag}(y_1,\ldots,y_m) & -(Y\circ A) \\ -(Y\circ A)^\top & \mathrm{diag}(y_{m+1},\ldots,y_{m+n})\end{bmatrix}}_{=:S} \succeq 0 \tag{8}$$

Finding an arrangement $\mathcal{A}$ for A that maximizes $\min_{i,j}\gamma_{i,j}(A|\mathcal{A})$ can be expressed as a standard SDP-problem (with slack variables $s_{i,j}$) as follows:

$$\min_{X,\mu,s} -\mu \ \textbf{ s.t. } \ \forall k: X_{k,k}=1\ ,\ \forall i,j: A_{i,j}(X_{i,m+j}+X_{m+j,i}) - s_{i,j} = 2\mu\ ,\ X\succeq 0\ ,\ s_{i,j}\geq 0\ ,\ \mu\geq 0$$

The $(m+n+mn+1)\times(m+n+mn+1)$-matrix of primal variables is then given by

$$\begin{bmatrix} X & 0 & 0 \\ 0 & \mathrm{diag}(s_{1,1},\ldots,s_{m,n}) & 0 \\ 0 & 0 & \mu\end{bmatrix} \ .$$

The dual variables are denoted y_k and $y_{i,j}$. Setting $Y = (y_{i,j})$, the matrix induced by the equality-constraints equals

$$\begin{bmatrix} \mathrm{diag}(y_1,\ldots,y_m) & Y\circ A & 0 & 0 \\ (Y\circ A)^\top & \mathrm{diag}(y_{m+1},\ldots,y_{m+n}) & 0 & 0 \\ 0 & 0 & -\mathrm{diag}(y_{1,1},\ldots,y_{m,n}) & 0 \\ 0 & 0 & 0 & -2\sum_{i,j} y_{i,j}\end{bmatrix} \ .$$

It is easy to see that condition SCQ is satisfied: one may assign value -1 to every variable $y_{i,j}$ and a sufficiently large value to every variable y_k so that all eigenvalues must be strictly positive according to the Geršgorin Disc Theorem. Thus, we have strong duality. The dual problem (with variables $-y_k/2$ substituted for y_k and $y_{i,j}/2$ substituted for $y_{i,j}$) looks as follows:[3]

$$\min_{Y,y} \frac{1}{2}\sum_k y_k \ \textbf{ s.t. } \ \sum_{i,j} y_{i,j} = 1\ ,\ y_{i,j}\geq 0\ ,\ \begin{bmatrix} \mathrm{diag}(y_1,\ldots,y_m) & -(Y\circ A) \\ -(Y\circ A)^\top & \mathrm{diag}(y_{m+1},\ldots,y_{m+n}\end{bmatrix} \succeq 0 \tag{9}$$

By strong duality, (9) equals $\max_{\mathcal{A}}\min_{i,j}\gamma_{i,j}(A|\mathcal{A})$, and (8) equals $\gamma_2^*(Y\circ A)$. A comparison of (9) with (8) shows that (6) holds. ■

[3] The constraint $C - \sum_\rho y_\rho A_\rho \succeq 0$ forces the $y_{i,j}$ to be non-negative and to satisfy $\sum_{i,j} y_{i,j} - 1 \geq 0$, but it is obvious that, for an optimal assignment to the variables $y_{i,j}$, their values will sum up to 1 exactly.

One can show that any arrangement of arbitrary dimension can be transformed (by virtue of Cholesky decomposition) into another arrangement of dimension at most $m+n+mn+1$ that achieves the same values for the respective margin parameters. Combined with a straightforward compactness and continuity argument this shows that there exists a maximizer $\mathcal{A}^*$ for $\min_{i,j} \gamma_{i,j}(A|\mathcal{A})$, and there exists a minimizer Y^* for $\gamma_2^*(Y \circ A)$. According to (6), both problems have the same optimal value, say γ^*, i.e.,

$$\gamma^* := \min_{i,j} \gamma_{i,j}(A|\mathcal{A}^*) = \gamma_2^*(Y^* \circ A) \ .$$

The following set $K(A)$ represents the "hard part" of the matrix $A \in \mathbb{R}^{m \times n}$ (thereby playing a similar role as support vectors in SVM optimization problems):

$$K(A|\mathcal{A}^*) := \{(i,j) :\ \gamma_{i,j}(A|\mathcal{A}) = \gamma^*\} \ \text{ and } K(A) := \bigcap_{\mathcal{A}^*} K(A|\mathcal{A}^*)$$

In the definition of $K(A)$, $\mathcal{A}^*$ ranges over all maximizers for $\min_{i,j} \gamma_{i,j}(A|\mathcal{A})$. We say that Y is *centered* on $K \subseteq \{1, \ldots, m\} \times \{1, \ldots, n\}$ if $y_{i,j} = 0$ for all $(i,j) \notin K$. With these notations, the following holds:

Corollary 2 *1. Every minimizer Y^* for $\gamma_2^*(Y \circ A)$ is centered on $K(A)$.*

2. $\max_{\mathcal{A}} \min_{i,j} \gamma_{i,j}(A|\mathcal{A}) = \max_{\mathcal{A}} \min_{(i,j) \in K(A)} \gamma_{i,j}(A|\mathcal{A})$.

Proof:

1. The claim is proved indirectly. Consider a matrix Y such that, for some $(i',j') \notin K(A)$, $y_{i',j'} > 0$. According to the definition of $K(A)$, there must exist a maximizer $\mathcal{A}^*$ for $\min_{i,j} \gamma_{i,j}(A|\mathcal{A})$ such that $(i',j') \notin K(A|\mathcal{A}^*)$. This implies that $\mathcal{A}^*$ achieves a Y-average margin strictly greater than γ^*. Thus, Y is not a minimizer for $\gamma_2^*(Y \circ A)$.

2. Let Y' range over all Y that are centered on $K(A)$. A straightforward modification of the proof of (6) shows that
$$\max_{\mathcal{A}} \min_{(i,j) \in K(A)} \gamma_{i,j}(A|\mathcal{A}) = \min_{Y'} \gamma_2^*(Y' \circ A) \tag{10}$$
Thus it suffices to show that
$$\min_{Y} \gamma_2^*(Y \circ A) = \min_{Y'} \gamma_2^*(Y' \circ A) \ .$$
But this is evident from the first part of the corollary.

■

The proof of the following result is similar to the proof of Theorem 1. It is found in Section A.

Theorem 3 *For every $A \in \mathbb{R}^{m \times n}$ and every probability vector p:* $\overline{\mathrm{mc}}_{p,MIN}(A) = \max_q \overline{\mathrm{mc}}_{p,q}(A)$.

3.1 Bounds on Average Margin Complexity

We make use of the inequalities

$$\|M\|_{tr} \le \gamma_2^*(M) \le \sqrt{mn} \cdot \|M\|_2 \ . \tag{11}$$

Because of (3), the second inequality is equivalent to (4), and it can also be found in (Linial et al., 2007). The first inequality is probably known as well but, since we are not aware of a proper reference, we will now provide the reader with a short proof for sake of completeness. Since the spectral norm is the dual of the trace norm, it suffices to show that

$$\gamma_2(M) \le \|M\|_2 \ .$$

We denote the rank of M by r, and we make use of the singular value decomposition

$$M = U \cdot \mathrm{diag}(\sigma_1, \ldots, \sigma_r) \cdot V^\top = \underbrace{(U \cdot \mathrm{diag}(\sqrt{\sigma_1}, \ldots, \sqrt{\sigma_r})}_{=:X} \cdot \underbrace{(V \cdot \mathrm{diag}(\sqrt{\sigma_1}, \ldots, \sqrt{\sigma_r})}_{=:Y}{}^\top \ . \tag{12}$$

Here, U is an $(m \times r)$-matrix whose columns $U_1, \ldots, U_r$ have unit norm and are pairwise orthogonal. Likewise, V is an $(n \times r)$-matrix whose columns $V_1, \ldots, V_r$ have unit norm and are pairwise orthogonal. $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0$ are the singular values of M. As a matter of fact, $\|M\|_2 = \sigma_1$. Now obviously

$$\gamma_2(M) \overset{(2)}{\le} r(X) \cdot r(Y) \le \sqrt{\sigma_1} \cdot \sqrt{\sigma_1} = \|M\|_2$$

which concludes the verification of (11). Because of (3), (11) is equivalent to

$$\frac{\sqrt{mn}}{\|M\|_2} \le \overline{\mathrm{mc}}(M) \le \frac{mn}{\|M\|_{tr}} \ . \tag{13}$$

3.2 Exact Determination of Average Margin Complexity

In general, the bounds (11) and (13) leave a gap. In this section, we consider families of matrices whose average margin complexity can be determined exactly: Hadamard matrices and matrices composed of all reflections of a single Boolean function.

By definition, a *Hadamard matrix* H *of order* n is a sign-matrix that satisfies $H \cdot H^\top = n \cdot I$.

Corollary 4 *Let H be a Hadamard matrix of order n. Then, $\overline{\mathrm{mc}}(H) = \sqrt{n}$.*

Proof: A Hadamard matrix H of order n satisfies $\sigma_1(H) = \cdots = \sigma_n(H)$. Thus, $\|H\|_{tr} = n \cdot \|H\|_2$, which makes the upper bound in (13) collapse to the lower bound in (13). ■

Lemma 5 *If $M \in \mathbb{R}^{n\times n}$ and the matrices U, V in its singular value decomposition (12) have only entries from $\{\pm 1/\sqrt{n}\}$, then $\gamma_2^*(M) = n \cdot \|M\|_2$.*

Proof: According to (11), $\gamma_2^*(M) \le n \cdot \|M\|_2$. By duality of norms, the converse direction is equivalent to $\gamma_2(M) \le \|M\|_{tr}/n$. An inspection of (12) shows that, given our assumptions on the entries of U and V, $r(X) = r(Y) = \sqrt{\|M\|_{tr}/n}$. Thus, $\gamma_2(M) \le r(X)r(Y) = \|M\|_{tr}/n$, as required. ■

We briefly note that the assumptions of Lemma 5 can be weakened: it suffices to assume that the first columns of U and V, respectively, have entries from $\{\pm 1/\sqrt{n}\}$. The proof will then make use of strong SDP duality.

Corollary 6 *Let $f : \{-1,1\}^d \to \{-1,1\}$ be a Boolean function, and let $L_\infty(f)$ denote the largest Fourier-coefficient in terms of absolute value. Consider the $(2^d \times 2^d)$-matrix $F_{x,y} := f(x \circ y)$ where $x \circ y := (x_1y_1, \ldots, x_dy_d)$. Then: $\overline{\mathrm{mc}}(F) = 1/L_\infty(f)$.*

Proof: Let $\hat{F}$ be the matrix with the Fourier-coefficients of f on its main diagonal and zeros elsewhere. Let H denote the Silvester-type Hadamard matrix of order 2^d. It is well-known (e.g., Doliwa et al., 2008) that the spectral decomposition of $2^{-d}F$ has the form

$$2^{-d}F = 2^{-d/2}H \cdot \hat{F} \cdot 2^{-d/2}H \ .$$

This implies that $\|F\|_2 = 2^d \cdot L_\infty(f)$ and that $M = 2^{-d}F$ satisfies the assumptions of Lemma 5 so that $\gamma_2^*(F) = 2^d\|F\|_2 = 2^{2d}L_\infty(f)$. Thus, according to (3) with $n = m = 2^d$, $\overline{\mathrm{mc}}(F) = 1/L_\infty(f)$. ■

4 The Replication Trick

We have learned the replication trick from Sherstov (2008). He used it (together with the classical Forster bound on dimension complexity) to show that, for every sign-matrix A, $\mathrm{SQdim}(A) = \max_p \mathrm{SQdim}_p(A)$ is bounded from above by twice the square of the dimension complexity of A. Here, we will use the trick (together with (4)) for showing that, for every real-valued matrix A, $\overline{\mathrm{mc}}_{p,q}(A) \ge \mathrm{FB}_{p,q}(A)$.

Lemma 7 *Let $A \in \mathbb{R}^{m\times n}$. Let p be an m-dimensional probability vector with rational components r_i/R (so that $\sum_{i=1}^m r_i = R$). Similarly, let q be an n-dimensional probability vector with rational components s_j/S (so that $\sum_{j=1}^n s_j = S$). Let $A_{\vec{s}}$ be the matrix that results from A by duplicating the j-th column s_j-times. Let $A_{\vec{r},\vec{s}}$ denote the matrix that results from $A_{\vec{s}}$ by duplicating the i-th row r_i-times. With this notation, the following holds:*

$$\overline{\mathrm{mc}}_{p,q}(A) = \overline{\mathrm{mc}}(A_{\vec{r},\vec{s}}) \text{ and } \sqrt{RS} \cdot \|P^{1/2}AQ^{1/2}\|_2 = \|A_{\vec{r},\vec{s}}\|_2 \qquad (14)$$

Proof: We first show that $\overline{\mathrm{mc}}_{p,q}(A) \le \overline{\mathrm{mc}}(A_{\vec{r},\vec{s}})$. Any arrangement $\mathcal{A} = (u_1, \ldots, u_m; v_1, \ldots, v_n)$ for A induces an arrangement $\mathcal{A}'$ for $A_{\vec{r},\vec{s}}$ where the k-th duplicate of row i (resp. column j) is represented by the (same) vector u_i (resp. v_j). The average margin achieved by $\mathcal{A}'$ equals

$$\frac{1}{RS}\sum_i\sum_j r_is_j\langle u_i, v_j\rangle A_{i,j} = \sum_i\sum_j \frac{r_i}{R}\frac{s_j}{S}\langle u_i, v_j\rangle A_{i,j} \ .$$

But the right hand-side equals the (p,q)-average margin achieved by $\mathcal{A}$.
Now, we show that $\overline{\mathrm{mc}}(A_{\vec{r},\vec{s}}) \le \overline{\mathrm{mc}}_{p,q}(A)$. To this end, we start with an arrangement $\mathcal{A}'$ for $A_{\vec{r},\vec{s}}$

where the k-th duplicate of row i (resp. column j) is represented by $u_i(k)$ (resp. $v_j(k)$). The average margin achieved by $\mathcal{A}'$ equals

$$\begin{aligned}\frac{1}{RS}\sum_i\sum_j\sum_{k_i=1}^{r_i}\sum_{l_j=1}^{s_j}\langle u_i(k_i),v_j(l_j)\rangle A_{i,j} &= \frac{1}{RS}\sum_i\sum_j\left\langle\sum_{k_i=1}^{r_i}u_i(k_i),\sum_{l_j=1}^{s_j}v_j(l_j)\right\rangle A_{i,j}\\ &= \sum_i\sum_j\frac{r_i}{R}\frac{s_j}{S}\left\langle\frac{1}{r_i}\sum_{k_i=1}^{r_i}u_i(k_i),\frac{1}{s_j}\sum_{l_j=1}^{s_j}v_j(l_j)\right\rangle A_{i,j}\ .\end{aligned}$$

But the final term coincides with the (p,q)-average margin that is achieved for A by the vectors

$$u_i=\frac{1}{r_i}\sum_{k_i=1}^{r_i}u_i(k_i)\text{ and }v_j=\frac{1}{s_j}\sum_{l_j=1}^{s_j}v_j(l_j)\ .$$

Note that, by the triangle inequality, $\|u_i\|_2$ is bounded by 1 provided that $\|u_i(1)\|_2,\ldots,\|u_i(r_i)\|_2$ are bounded by 1. The analogous argument applies to v_j.
As for the second equation in (14), it suffices to show that $\|\sqrt{S}AQ^{1/2}\|_2=\|A_{\vec{s}}\|_2$. (We can then apply this equality with $P^{1/2}A$ substituted for A and proceed with a symmetry argument.) Our proof for $\|\sqrt{S}AQ^{1/2}\|_2=\|A_{\vec{s}}\|_2$ will make use of (1). We first show that $\|\sqrt{S}AQ^{1/2}\|_2\le\|A_{\vec{s}}\|_2$. Note that the entry (i,j) of matrix $\sqrt{S}AQ^{1/2}$ coincides with $\sqrt{s_j}A_{i,j}$. With any n-dimensional vector v, we associate the S-dimensional vector v' which is composed of sub-vectors $v'(1),\ldots,v'(n)$ of dimensions $s_1,\ldots,s_n$, respectively, such that $v'(j)=\frac{v_j}{\sqrt{s_j}}\cdot\vec{1}$. Note that $\|v'\|_2=\|v\|_2$. Furthermore note that

$$u^\top A_{\vec{s}}v'=\sum_i\sum_j s_ju_i\frac{v_j}{\sqrt{s_j}}A_{i,j}=\sum_i\sum_j u_iv_j(\sqrt{s_j}A_{i,j})=u^\top(\sqrt{S}AQ^{1/2})v\ .$$

Now, we show that $\|A_{\vec{s}}\|_2\le\|\sqrt{S}AQ^{1/2}\|_2$. To this end, we consider an m-dimensional vector u and an S-dimensional vector v'. We can think of v' as being composed of s_j-dimensional sub-vectors $v'(j)$ for $j=1,\ldots,n$. Then,

$$u^\top A_{\vec{s}}v'=\sum_i\sum_j\sum_{k_j=1}^{s_j}u_iv'(j)_{k_j}A_{i,j}=\sum_j\left(\sum_i u_iA_{i,j}\right)\left(\sum_{k_j=1}^{s_j}v'(j)_{k_j}\right)\ .$$

Setting $v_j:=\frac{1}{\sqrt{s_j}}\sum_{k_j=1}^{s_j}v'(j)_{k_j}$, the latter term equals

$$\sum_i\sum_j u_iv_j(\sqrt{s_j}A_{i,j})=u^\top(\sqrt{S}AQ^{1/2})v\ .$$

Note that

$$v_j^2=\frac{1}{s_j}\cdot\left\langle v'(j),\vec{1}\right\rangle^2\le\frac{1}{s_j}\cdot\langle v'(j),v'(j)\rangle\cdot\left\langle\vec{1},\vec{1}\right\rangle=\langle v'(j),v'(j)\rangle\ ,$$

which implies that $\|v\|_2\le\|v'\|_2$. ■

Corollary 8 *For all probability vectors p,q: $\overline{\mathrm{mc}}_{p,q}(A)\ge\mathrm{FB}_{p,q}(A)$.*

Proof: With the notation from Lemma 7, the following holds for all rational probability vectors p,q:

$$\overline{\mathrm{mc}}_{p,q}(A)=\overline{\mathrm{mc}}(A_{\vec{r},\vec{s}})\overset{(4)}{\ge}\mathrm{FB}(A_{\vec{r},\vec{s}})=\mathrm{FB}_{p,q}(A)\ .$$

In order to generalize this equality to arbitrary probability vectors (with possibly non-rational components), we can use that fact that $\mathbb{Q}$ is dense in $\mathbb{R}$ and apply an obvious continuity argument. ■

5 SQ-Dimension and Margin-Complexity

In this section we focus on sign matrices. We establish two more relations (see Lemmas 9 and 10), and then we put all pieces together and arrive at the inequalities in (15) and (16). As a by-product, we obtain two results, see (17) and (18), which might be of independent interest.

Lemma 9 *For every* $A \in \{-1,1\}^{m\times n}$*:* $\overline{\mathrm{mc}}_{p,MIN}(A) < \sqrt{\mathrm{SQdim}_p(A)} \cdot (\mathrm{SQdim}_p(A)+1)$.

Proof: Let $d := \mathrm{SQdim}_p(A)$. Select a subset $S = \{s(1),\ldots,s(d)\} \subseteq \{1,\ldots,n\}$ such that

$$\left(\forall k \neq l \in S : |\langle A_k, A_l\rangle_p| \leq \frac{1}{d}\right) \wedge \left(\forall j \in \{1,\ldots,n\}, \exists k(j) \in \{1,\ldots,d\} : |\left\langle A_j, A_{s(k(j))}\right\rangle_p| > \frac{1}{d+1}\right) .$$

Let $\sigma_j := \mathrm{sign}(\left\langle A_j, A_{s(k(j))}\right\rangle_p)$. We define a d-dimensional arrangement for A as follows:

$$\left(\forall i = 1,\ldots,m : u_i = \frac{1}{\sqrt{d}} \cdot (A_{i,s(1)},\ldots,A_{i,s(d)})\right) \wedge \left(\forall j = 1,\ldots,n : v_j = \sigma_j \cdot \vec{e}_{k(j)}\right)$$

It follows that $\langle u_i, v_j\rangle = \sigma_j \cdot A_{i,s(k(j))}/\sqrt{d}$, and our embedding exhibits the margin parameters

$$\gamma_{i,j} = \langle u_i, v_j\rangle \cdot A_{i,j} = \frac{\sigma_j \cdot A_{i,s(k(j))} \cdot A_{i,j}}{\sqrt{d}} .$$

Averaging w.r.t. to p yields

$$\sum_i p_i\gamma_{i,j} = \frac{1}{\sqrt{d}} \cdot |\left\langle A_j, A_{s(k(j))}\right\rangle_p| > \frac{1}{\sqrt{d}\cdot(d+1)} .$$

Since this holds for every choice j, we get $\overline{\mathrm{mc}}_{p,MIN}(A) < \sqrt{d}\cdot(d+1)$, as desired. ■

The proof of the following result builds on a proof by Sherstov (2008) for a quite similar result:

Lemma 10 *For every* $A \in \{-1,1\}^{m\times n}$*:* $\mathrm{SQdim}_p(A) < 2 \cdot \max_{A'} \mathrm{FB}_p(A')^2$.

Proof: Let $d = \mathrm{SQdim}_p(A)$, and let $S \subseteq \{1,\ldots,n\}$ be chosen as in the proof of Lemma 9. Let A' be the submatrix that is formed by the columns $s(1),\ldots,s(d)$ of A. It follows that

$$C := A'^{\top}PA' = (P^{1/2}A')^{\top}(P^{1/2}A') \in \mathbb{R}^{d\times d}$$

has ones on the main diagonal and entries of absolute value at most $1/d$ elsewhere. We apply an argument of Sherstov (2008) and conclude that

$$\|C\|_2 \leq \|C - I_d\|_2 + \|I_d\|_2 \leq \|C - I_d\|_2 + \|I_d\|_2 = \sqrt{\frac{d(d-1)}{d^2}} + 1 < 2 .$$

Note that

$$\|C\|_2 = \|P^{1/2}A'\|_2^2 .$$

The proof is now accomplished as follows:

$$\mathrm{FB}_p(A')^2 \geq \frac{d}{\|P^{1/2}A'\|_2^2} = \frac{d}{\|C\|_2} > \frac{d}{2}$$

■

The combination of Lemma 9, Lemma 10, and Corollary 8 demonstrates that the parameters $\mathrm{SQdim}_p(A)$, $\max_{A'} \mathrm{FB}_p(A')$, $\max_q \mathrm{FB}_{p,q}(A)$, and $\max_q \overline{\mathrm{mc}}_{p,q}(A) = \overline{\mathrm{mc}}_{p,MIN}(A)$ are related as follows:

$$\mathrm{SQdim}_p(A) < 2\max_{A'} \mathrm{FB}_p(A')^2 \leq 2\max_q \mathrm{FB}_{p,q}(A)^2 \leq 2\max_q \overline{\mathrm{mc}}_{p,q}(A)^2 < 2\mathrm{SQdim}_p(A)(\mathrm{SQdim}_p(A)+1)^2 \tag{15}$$

Applying the operation "$\max_p$" to (15), we get

$$\mathrm{SQdim}(A) < 2 \cdot \max_{p,q} \mathrm{FB}_{p,q}(A)^2 \leq 2 \cdot \max_{p,q} \overline{\mathrm{mc}}_{p,q}(A)^2 < 2 \cdot \mathrm{SQdim}(A) \cdot (\mathrm{SQdim}(A)+1)^2 . \tag{16}$$

Since $\max_{p,q} \overline{\mathrm{mc}}_{p,q}(A) = \max_{p,q} \overline{\mathrm{mc}}_{p,q}(A^\top)$ — an analogous remark is valid for $\max_{p,q} \mathrm{FB}_{p,q}$ — it follows from (16) that

$$\mathrm{SQdim}(A^\top) < 2 \cdot \mathrm{SQdim}(A) \cdot (\mathrm{SQdim}(A)+1)^2 \ . \tag{17}$$

This improves on a result by Sherstov (2008): he used a polynomial relation between $\mathrm{SQdim}(A)$ and the discrepancy of A with respect to product distributions for showing that $\mathrm{SQdim}(A^\top) \le 32 \cdot \mathrm{SQdim}(A)^4$.

Recall that, by convention, A' ranges over all sub-matrices of A which can be composed by (complete) columns of A. Let A'' range over all sub-matrices of A. We claim that

$$\max_{A''} \mathrm{FB}(A'') \le \max_{p,q} \mathrm{FB}_{p,q}(A) < 64 \cdot (1+o(1)) \cdot \max_{A''} \mathrm{FB}(A'')^9 \ . \tag{18}$$

The first inequality is obvious. The last inequality is obtained by applying (15) twice, the first time on A and the second time on the transpose of A'.

6 CSQ-Dimension and Margin Complexity

Feldman (2008) has shown the following result. If a concept class $\mathcal{C}$ over domain $\mathcal{X}$ has CSQ-dimension d, then there exists a family W consisting of d Boolean base functions such every function in $\mathcal{C}$ can be written as the majority of $O(\log(|\mathcal{X}|)d^2)$ functions properly chosen from W. Viewing a sign-matrix A as a concept class, it is not hard to infer from that result an upper bound on $\mathrm{mc}(A)$ in terms of $\mathrm{CSQdim}(A)$. However, a direct derivation of such an upper bound (as in the proof of the following lemma) leads to a tighter relationship:

Lemma 11 *For every $A \in \mathbb{R}^{m\times n}$: $\mathrm{mc}(A) \le \mathrm{CSQdim}(A)^{1.5}$.*

Proof: Let $d := \mathrm{CSQdim}(A)$, and let $h_1, \ldots, h_d \in \mathbb{R}^m$ be universally correlated with A. According to Lemma 1, there exists a matrix $Y = (y_{i,j})$ such that $\mathrm{mc}(A) = \overline{\mathrm{mc}}_Y(A)$. We will present a d-dimensional arrangement $u_1, \ldots, u_m; v_1, \ldots, v_n$ whose Y-average margin equals $1/d^{1.5}$ (which proves the lemma). To this end, we set

$$u_i := \frac{1}{\sqrt{d}} \cdot (h_{i,1}, \ldots, h_{i,d})$$

where $h_{i,k}$ denotes the i-th component of vector h_k. Note that $\|u_i\|_2 \le 1$. Furthermore, let $Y_j := y_{1,j} + \cdots + y_{m,j}$, and let the m-dimensional probability vector p_j be given by $(p_j)_i := y_{i,j}/Y_j$. Because $h_1, \ldots, h_d$ is universally correlated with A, the following holds. For every j, there exists $k(j) \in \{1, \ldots, d\}$ such that $|\langle h_{k(j)}, A_j \rangle_{p_j}| \ge 1/d$. Now we set $\sigma_j := \mathrm{sign}(\langle h_{k(j)}, A_j \rangle_{p_j})$, $v_j := \sigma_j \cdot \vec{e}_{k(j)}$, and we bound the Y-average margin from below as follows:

$$\sum_i \sum_j y_{i,j} \langle u_i, v_j \rangle A_{i,j} = \frac{1}{\sqrt{d}} \cdot \sum_j Y_j \sigma_j \sum_i \frac{y_{i,j}}{Y_j} h_{i,k(j)} A_{i,j} = \frac{1}{\sqrt{d}} \cdot \sum_j Y_j |\langle h_{k(j)}, A_j \rangle_{p_j}| \ge \frac{1}{d^{1.5}}$$

■

As for the converse direction, we get the following result:

Lemma 12 *For every $A \in \mathbb{R}^{m\times n}$ and $\ell(m,n) := 32\ln(4mn)$:*

$$\mathrm{CSQdim}(A) \le \lceil \ell(m,n) \cdot \mathrm{mc}(A)^2 \rceil$$

Proof: Consider an arrangement $\mathcal{A}$ that maximizes $\gamma := \min_{i,j} \gamma_{i,j}(A|\mathcal{A})$ so that $\mathrm{mc}(A) = 1/\gamma$. It is well-known[4] that $\mathcal{A}$ can be transformed into another arrangement $\mathcal{A}' = (u_1, \ldots, u_m; v_1, \ldots, v_n)$ that is d-dimensional for $d := \lceil \ell(m,n)/\gamma^2 \rceil$ and still satisfies $\min_{i,j} \gamma_{i,j}(A|\mathcal{A}') \ge \gamma/2$. For every $k \in \{1, \ldots, d\}$, let $u_{i,k}$ denote the i-th component of u_k. We will show that $h_1, \ldots, h_d$ given by

$$h_k = (u_{1,k}, \ldots, u_{m,k})$$

is universally correlated with A. To this end, let p be an arbitrary but fixed m-dimensional probability vector, and let $v'_j = \|v_j\|_1^{-1} \cdot v_j$ so that

$$\|v'_j\|_1 = \sum_{k=1}^{d} |v'_{j,k}| = 1 \ . \tag{19}$$

[4]This is a typical application of random projections (see Johnson and Lindenstrauss, 1984, Arriaga and Vempala, 1999). E.g., apply Corollary 19 in the paper by Ben-David et al. (2002).

Note that $\|v_j\|_1 \le \sqrt{d}$ since $\mathcal{A}'$ is a d-dimensional arrangement. It follows that

$$\min_{i,j} \langle u_i, v'_j \rangle A_{i,j} \ge \frac{\gamma}{2\sqrt{d}} \ ,$$

and the following holds for every $j \in \{1, \ldots, n\}$:

$$\begin{aligned} \frac{\gamma}{2\sqrt{d}} &\le \sum_i p_i \langle u_i, v'_j \rangle A_{i,j} = \sum_i p_i A_{i,j} \sum_{k=1}^{d} |v'_{j,k}| \operatorname{sign}(v_{j,k}) u_{i,k} \\ &= \sum_{k=1}^{d} |v'_{j,k}| \sum_i p_i \operatorname{sign}(v_{j,k}) u_{i,k} A_{i,j} = \sum_{k=1}^{d} |v'_{j,k}| \langle \operatorname{sign}(v_{j,k}) h_k, A_j \rangle_p \end{aligned}$$

The latter sum is a convex combination of inner products because of (19), and, as the above calculation shows, the inner products achieve a value of at least $\gamma/(2\sqrt{d})$ on the average. By the pigeon-hole principle, there exists $k(j) \in \{1, \ldots, d\}$ such that

$$\operatorname{sign}(v_{j,k(j)}) \cdot \langle h_{k(j)}, A_j \rangle_p \ge \frac{\gamma}{2\sqrt{d}} \ .$$

It is easily checked that $\gamma/(2\sqrt{d}) \ge 1/d$ (by solving this inequality for d and comparing with the above definition of d). It follows that, as announced above, $h_1, \ldots, h_d$ is universally correlated with A. ■

Conclusions: Looking back, we have seen a hierarchy of margin optimization problems, and the dual versions of these problems nicely reflect why the optimal values become smaller when we go up in the hierarchy. In the dual setting, we are always faced with a problem of maximizing the total margin of a matrix of the form $Y \circ A$ (which is the Y-average margin of A). The crucial issues are the structure of the the matrix Y and whether its choice is under control of "nature" or under control of an "intelligent adversary":

(a) The easiest problem from the perspective of the margin-maximizer results when Y is of the form $pq^\top$ for fixed and "benign" p, q. Here "benign" means that the distribution p on the rows of A (= instances of the domain) and the distribution q on the columns of A (= possible target concepts) are resulting from a learning application (and not from settings within a worst-case analysis). In this situation the goal of the margin-maximizer roughly corresponds to achieving a reasonably large "soft margin" on the average.

(b) The problem becomes harder when q (Case 1) or both of p, q (Case 2) are under control of an adversary so that $\max_q \overline{\mathrm{mc}}_{p,q}(A) = \overline{\mathrm{mc}}_{p,MIN}(A)$ in Case 1 and $\max_{p,q} \overline{\mathrm{mc}}_{p,q}(A) = \max_p \overline{\mathrm{mc}}_{p,MIN}(A)$ in Case 2 would be the appropriate complexity measures. Maximizing over all choices of q means choosing the target concept in a worst-case fashion. Maximizing over all choices of p (as in Case 2) means that the domain distribution is chosen in a worst-case fashion although it is still fixed (because the chosen arrangement may depend on p). Because of the polynomial relation between average margin complexity and the SQ-dimension, Case 1 corresponds to weak learning in the SQ model under a fixed distribution. A similar remark is valid for Case 2 but here we have to cope with the hardest fixed distribution.

(c) The hardest problem results when an adversary controls Y, and Y is an arbitrary matrix with non-negative entries summing up to 1 (as opposed to a matrix of the form $pq^\top$, which is the special case where Y has rank 1). Now $\max_Y \overline{\mathrm{mc}}_Y(A) = \mathrm{mc}(A)$ is the appropriate complexity measure, and the learning goal is to achieve a reasonably large hard margin for every possible target concept. Because of the polynomial relation between $\mathrm{mc}(A)$ and the CSQ-dimension, the learning goal can be achieved iff the concept class is distribution-independently weakly learnable in the CSQ model.

Feldman (2008) has shown that there exist classes (e.g., Boolean decision lists) which are distribution independently (weakly or strongly)[5] learnable in the SQ model but not (not even weakly) in the CSQ model. This also shows that $\max_{p,q} \overline{\mathrm{mc}}_{p,q}(A)$ and $\max_Y \overline{\mathrm{mc}}_Y(A)$ are not polynomially related. (There is even an exponential gap.) Thus imposing the rank 1 constraint on Y makes much of a difference.

[5] Because of the Boosting-result by Aslam and Decatur (1998) weak learners can be transformed into strong learners in this model without much loss of efficiency.

Open Problems: The level of distribution-independent SQ-learning is located somewhere between (b) and (c). It would be interesting to find a combinatorial parameter (or another variant of margin optimization?) that characterizes this level. A parameter of this kind must be lower-bounded by the SQ-dimension and upper-bounded by the CSQ-dimension. It would furthermore be interesting to find a concept class that separates distribution-independent SQ-learning from SQ-learning w.r.t. the hardest fixed distribution.
The correspondence between maximization of the average margin and typical soft-margin optimization problems would be more convincing if we replaced $\gamma_{i,j}(A|\mathcal{A})$ by $\min\{\gamma_{i,j}(A|\mathcal{A}), \gamma\}$ for some $\gamma > 0$ so that few extremely large margin parameters cannot provide compensation for many small or negative margin parameters.[6] It would be interesting to know whether results similar to the ones in this paper can be shown for this "average clipped margin".

References

Farid Alizadeh. Interior point methods to semidefinite programming with applications to combinatorial optimization. *SIAM Journal on Optimization*, 5(13), 1995.

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.

Javed A. Aslam and Scott E. Decatur. General bounds on statistical query learning and PAC learning with noise via hypothesis boosting. *Information and Computation*, 141(2):85–118, 1998.

Shai Ben-David, Nadav Eiron, and Hans U. Simon. Limitations of learning via embeddings in euclidean half-spaces. *Journal of Machine Learning Research*, 3:441–461, 2002.

Avrim Blum, Adam Kalai, and Hal Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the Association on Computing Machinery*, 50(4):506–519, 2003.

Nader H. Bshouty and Vitaly Feldman. On using extended statistical queries to avoid membership queries. *Journal of Machine Learning Research*, 2:359–395, 2002.

Thorsten Doliwa, Michael Kallweit, and Hans U. Simon. Dimension and margin bounds for reflection-invariant kernels. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 157–167, 2008.

Vitaly Feldman. Evolvability from learning algorithms. In *Proceedings of the 2008 ACM International Symposium on Theory of Computing*, pages 619–628, 2008.

Jürgen Forster. A linear lower bound on the unbounded error communication complexity. *Journal of Computer and System Sciences*, 65(4):612–625, 2002.

Jürgen Forster and Hans U. Simon. On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes. *Theoretical Computer Science*, 350(1):40–48, 2006.

W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mapping into Hilbert spaces. *Contemp. Math.*, 26:189–206, 1984.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the Association on Computing Machinery*, 45(6):983–1006, 1998.

Troy Lee and Adi Shraibman. Lower bounds in communication complexity. *Foundations and Trends in Theoretical Computer Science*, 3(4):263–399, 2009.

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

Alexander Sherstov. Halfspace matrices. *Computational Complexity*, 17(2):149–178, 2008.

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

Leslie G. Valiant. Evolvability. *Journal of the Association on Computing Machinery*, 56(1):3:1–3:21, 2009.

[6] Furthermore note the connection to using hinge-loss in soft-margin optimization problems.

A Proof of Theorem 3

We will prove the equivalent statement

$$\max_{\mathcal{A}} \min_j \sum_i p_i \gamma_{i,j}(A|\mathcal{A}) = \min_q \gamma_2^*(pq^\top \circ A) \ . \tag{20}$$

We know from the proof of Theorem 1 that $\gamma_2^*(pq^\top \circ A)$ coincides with (8) provided that $Y = pq^\top$. Let us now discuss the left hand-side of (20). Setting $M := P \cdot A$, we obtain $\min_j \sum_i p_i \gamma_{i,j}(A|\mathcal{A}) = \min_j \sum_i \gamma_{i,j}(M|\mathcal{A})$.[7] Finding an arrangement $\mathcal{A}$ for M that maximizes $\min_j \sum_i \gamma_{i,j}(M|\mathcal{A})$ can be expressed as a standard SDP-problem (with slack variables s_j) as follows:

$$\min_{X,\mu,s} -\mu \ \ \textbf{s.t.} \ \ \forall k : X_{k,k} = 1 \ , \ \forall j : \sum_i M_{i,j}(X_{i,m+j} + X_{m+j,i}) - s_j = 2\mu \ , \ X \succeq 0 \ , \ \mu \geq 0 \ , \ s_j \geq 0$$

The $(m+2n+1) \times (m+2n+1)$-matrix of primal variables is then given by

$$\begin{bmatrix} X & 0 & 0 \\ 0 & \mathrm{diag}(s_1, \ldots, s_n) & 0 \\ 0 & 0 & \mu \end{bmatrix} \ .$$

The dual variables are denoted y_k and q_j. The matrix induced by the equality-constraints equals

$$\begin{bmatrix} \mathrm{diag}(y_1, \ldots, y_m) & M \cdot Q & 0 & 0 \\ (M \cdot Q)^\top & \mathrm{diag}(y_{m+1}, \ldots, y_{m+n}) & 0 & 0 \\ 0 & 0 & -Q & 0 \\ 0 & 0 & 0 & -2(q_1 + \cdots + q_n) \end{bmatrix} \ .$$

It is easy to see that condition SCQ is satisfied. Thus, we have strong duality. The dual problem (with variables $-y_k/2$ substituted for y_k, $q_j/2$ substituted for q_j, and $P \cdot A$ substituted for M) looks as follows:

$$\min_{q,y} \frac{1}{2} \sum_k y_k \ \ \textbf{s.t.} \ \ \sum_j q_j = 1 \ , \ q_j \geq 0 \ , \ \begin{bmatrix} \mathrm{diag}(y_1, \ldots, y_m) & -(P \cdot A \cdot Q) \\ -(P \cdot A \cdot Q)^\top & \mathrm{diag}(y_{m+1}, \ldots, y_{m+n}) \end{bmatrix} \succeq 0 \tag{21}$$

By strong duality, $\max_{\mathcal{A}} \min_j \sum_i p_i \gamma_{i,j}(A|\mathcal{A})$ equals (21). As discussed above, $\gamma_2^*(pq^\top \circ A)$ equals (8) provided that $Y = pq^\top$ so that $Y \circ A = pq^\top \circ A = P \cdot A \cdot Q$. A comparison of (21) and (8) shows that (20) holds.

[7] We remind the reader to the convention $P = \mathrm{diag}(p)$ and $Q = \mathrm{diag}(q)$.