Sign rank versus VC dimension*

Noga Alon† Shay Moran‡ Amir Yehudayoff§

October 2, 2016

Abstract

This work studies the maximum possible sign rank of $N \times N$ sign matrices with a given VC dimension d. For $d = 1$, this maximum is three. For $d = 2$, this maximum is $\tilde{\Theta}(N^{1/2})$. For $d > 2$, similar but slightly less accurate statements hold. The lower bounds improve over previous ones by Ben-David et al., and the upper bounds are novel.

The lower bounds are obtained by probabilistic constructions, using a theorem of Warren in real algebraic topology. The upper bounds are obtained using a result of Welzl about spanning trees with low stabbing number, and using the moment curve.

The upper bound technique is also used to: (i) provide estimates on the number of classes of a given VC dimension, and the number of maximum classes of a given VC dimension – answering a question of Frankl from '89, and (ii) design an efficient algorithm that provides an $O(N/\log(N))$ multiplicative approximation for the sign rank.

We also observe a general connection between sign rank and spectral gaps which is based on Forster's argument. Consider the $N \times N$ adjacency matrix of a Δ regular graph with a second eigenvalue of absolute value λ and $\Delta \leq N/2$. We show that the sign rank of the signed version of this matrix is at least Δ/λ. We use this connection to prove the existence of a maximum class $C \subseteq \{\pm 1\}^N$ with VC dimension 2 and sign rank $\tilde{\Theta}(N^{1/2})$. This answers a question of Ben-David et al. regarding the sign rank of large VC classes. We also describe limitations of this approach, in the spirit of the Alon-Boppana theorem.

We further describe connections to communication complexity, geometry, learning theory, and combinatorics.

*A preliminary version of this work was published in the proceeding of COLT'16.

†Sackler School of Mathematics and Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel, Microsoft Research, Herzliya, and School of Mathematics, Institute for Advanced Study, Princeton, NJ 08540. nogaa@tau.ac.il. Research supported in part by a USA-Israeli BSF grant, by an ISF grant, by the Israeli I-Core program and by the Oswald Veblen Fund.

‡Department of Computer Science, Technion-IIT, Microsoft Research, Herzliya, and Max Planck Institute for Informatics, Saarbrücken, Germany. shaymrn@cs.technion.ac.il.

§Department of Mathematics, Technion-IIT. Email: amir.yehudayoff@gmail.com. Horev fellow – supported by the Taub foundation. Research also supported by ISF and BSF.

Contents

1 Introduction

Boolean matrices (with $0, 1$ entries) and sign matrices (with ± 1 entries) naturally appear in many areas of research[1]. We use them e.g. to represent set systems and graphs in combinatorics, hypothesis classes in learning theory, and boolean functions in communication complexity.

This work further investigates the relation between two useful complexity measures on sign matrices.

Definition (Sign rank)**.** *For a real matrix M with no zero entries, let sign(M) denote the sign matrix such that $(sign(M))_{i,j} = sign(M_{i,j})$ for all i, j. The sign rank of a sign matrix S is defined as*

$$\textit{sign-rank}(S) = \min\{\textit{rank}(M) : \textit{sign}(M) = S\},$$

where the rank is over the real numbers.

The sign rank captures the minimum dimension of a real space in which the matrix can be embedded using half spaces through the origin [2] (see for example [48]).

Definition (Vapnik-Chervonenkis dimension)**.** *The VC dimension of a sign matrix S, denoted $VC(S)$, is defined as follows. A subset C of the columns of S is called shattered if each of the $2^{|C|}$ different patterns of ones and minus ones appears in some row in the restriction of S to the columns in C. The VC dimension of S is the maximum size of a shattered subset of columns.*

The VC dimension captures the size of the minimum ϵ-net for the underlying set system [38, 42].

The VC dimension and the sign rank appear in various areas of computer science and mathematics. One important example is learning theory, where the VC dimension captures the sample complexity of learning in the PAC model [19, 66], and the sign rank relates to the generalization guarantees of practical learning algorithms, such as support vector machines, large margin classifiers, and kernel classifiers [47, 32, 33, 34, 23, 67]. Loosely speaking, the VC dimension relates to learnability, while sign rank relates to learnability by linear classifiers. Another example is communication complexity, where the sign rank is equivalent to the unbounded error randomized communication complexity [55], and the VC dimension relates to one round distributional communication complexity under product distributions [43],

The main focus of this work is how large can the sign rank be for a given VC dimension. In learning theory, this question concerns the universality of linear classifiers. In communication complexity, this concerns the difference between randomized communication complexity with unbounded error and between communication complexity under product distribution with bounded error. Previous works have studied these differences from the communication complexity perspective [64, 63] and the learning theory perspective [15]. In this work we provide explicit matrices and stronger separations compared to those of [64, 63] and [15]. See the discussions in Section 1.2 and Section 2.4 for more details.

---

[1]There is a standard transformation of a boolean matrix B to the sign matrix $S = 2B - J$, where J is the all 1 matrix. The matrix S is called the signed version of B, and the matrix B is called the boolean version of S.

[2]That is, the columns correspond to points in $\mathbb{R}^k$ and the rows to half spaces through the origin (i.e. collections of all points $x \in \mathbb{R}^k$ so that $\langle x, v \rangle \geq 0$ for some fixed $v \in \mathbb{R}^k$).

1.1 Duality

We start by providing alternative descriptions of the VC dimension and sign rank, which demonstrate that these notions are dual to each other. The sign rank of a sign matrix S is the maximum number k such that

$$\forall\ M \text{ such that } \mathrm{sign}(M) = S\ \ \exists\ k \text{ columns } j_1, \ldots, j_k \\ \text{the columns } j_1, \ldots, j_k \text{ are linearly independent in } M$$

The *dual sign rank* of S is the maximum number k such that

$$\exists\ k \text{ columns } j_1, \ldots, j_k\ \ \forall\ M \text{ such that } \mathrm{sign}(M) = S \\ \text{the columns } j_1, \ldots, j_k \text{ are linearly independent in } M.$$

It turns out that the dual sign rank is almost equivalent to the VC dimension (the proof is in Section 3.1).

Proposition 1. *The dual sign rank of S is the VC dimension of the matrix $\begin{bmatrix} S \\ -S \end{bmatrix}$. As a corollary: $VC(S) \leq$ dual-sign-rank$(S) \leq 2VC(S) + 1$.*

As the dual sign rank is at most the sign rank, it follows that the VC dimension is at most the sign rank. This provides further motivation for studying the largest possible gap between sign rank and VC dimension; it is equivalent to the largest possible gap between the sign rank and the dual sign rank.

It is worth noting that there are some interesting classes of matrices for which these quantities are equal. One such example is the $2^n \times 2^n$ disjointness matrix $DISJ$, whose rows and columns are indexed by all subsets of $[n]$, and $DISJ_{x,y} = 1$ if and only if $|x \cap y| > 0$. For this matrix both the sign rank and the dual sign rank are exactly $n + 1$; indeed, its sign rank is at most $n + 1$, as witnessed by the matrix $\left(\sum_{i=1}^{n} v_i \cdot v_i^t\right) - \frac{1}{2}J$, where $J_{i,j} = 1$ for all i, j, and $v_i(x) = 1$ whenever $i \in x$. On the other hand, its dual sign is at least $n + 1$ as witnessed by the columns indexed by $\emptyset, \{1\}, \ldots, \{n\}$ that are shattered in $\begin{bmatrix} DISJ \\ -DISJ \end{bmatrix}$.

1.2 Sign rank versus VC dimension

The VC dimension is at most the sign rank. On the other hand, it is long known that the sign rank is not bounded from above by any function of the VC dimension. Alon, Haussler, and Welzl [7] provided examples of $N \times N$ matrices with VC dimension 2 for which the sign rank tends to infinity with N. [15] used ideas from [6] together with estimates concerning the Zarankiewicz problem to show that many matrices with constant VC dimension (at least 4) have high sign rank.

We further investigate the problem of determining or estimating the maximum possible sign rank of $N \times N$ matrices with VC dimension d. Denote this maximum by $f(N, d)$. We are mostly interested in fixed d and N tending to infinity.

We observe that there is a dichotomy between the behaviour of $f(N, d)$ when $d = 1$ and when $d > 1$. The value of $f(N, 1)$ is 3, but for $d > 1$, the value of $f(N, d)$ tends to infinity with N. We now discuss the behaviour of $f(N, d)$ in more detail, and describe our results.

We start with the case $d = 1$. The following theorem and claim imply that for all $N \geq 4$,

$$f(N, 1) = 3.$$

The following theorem which was proved by [7] shows that for $d = 1$, matrices with high sign rank do not exist. For completeness, we provide our simple and constructive proof in Section 3.2.1.

Theorem 2 ([7]). *If the VC dimension of a sign matrix M is one then its sign rank is at most 3.*

We also note that the bound 3 is tight (see Section 3.2.1 for a proof).

Claim 3. *For $N \geq 4$, the $N \times N$ signed identity matrix (i.e. the matrix with 1 on the diagonal and -1 off the diagonal) has VC dimension one and sign rank 3.*

Next, we consider the case $d > 1$, starting with lower bounds on $f(N, d)$. As mentioned above, two lower bounds were previously known: [7] showed that $f(N, 2) \geq \Omega(\log N)$. [15] showed that $f(N, d) \geq \omega(N^{1-\frac{2}{d}-\frac{1}{2^{d/2}}})$, for every fixed d, which provides a nontrivial result only for $d \geq 4$. We prove the following stronger lower bound.

Theorem 4. *The following lower bounds on $f(N, d)$ hold:*

1. $f(N, 2) \geq \Omega(N^{1/2}/\log N)$.
2. $f(N, 3) \geq \Omega(N^{8/15}/\log N)$.
3. $f(N, 4) \geq \Omega(N^{2/3}/\log N)$.
4. *For every fixed $d > 4$,*

$$f(N, d) \geq \Omega(N^{1-(d^2+5d+2)/(d^3+2d^2+3d)}/\log N).$$

To understand part 4 better, notice that

$$\frac{d^2 + 5d + 2}{d^3 + 2d^2 + 3d} = \frac{1}{d} + \frac{3d - 1}{d^3 + 2d^2 + 3d},$$

which is close to $1/d$ for large d. The proofs are described in Section 3.2, where we also discuss the tightness of our arguments.

What about upper bounds on $f(N, d)$? It is shown in [15] that for every matrix in a certain class of $N \times N$ matrices with constant VC dimension, the sign rank is at most $O(N^{1/2})$. The proof uses the connection between sign rank and communication complexity. However, there is no general upper bound for the sign rank of matrices of VC dimension d in [15], and the authors explicitly mention the absence of such a result.

Here we prove the following upper bounds, using a concrete embedding of matrices with low VC dimension in real space.

Theorem 5. *For every fixed $d \geq 2$,*

$$f(N,d) \leq O(N^{1-1/d}).$$

In particular, this determines $f(N,2)$ up to a logarithmic factor:

$$\Omega(N^{1/2}/\log N) \leq f(N,2) \leq O(N^{1/2}).$$

The above results imply existence of sign matrices with high sign rank. However, their proofs use counting arguments and hence do not provide a method of certifying high sign rank for explicit matrices. In the next section we show how one can derive a lower bound for the sign rank of many explicit matrices.

1.3 Sign rank and spectral gaps

Spectral properties of boolean matrices are known to be deeply related to their combinatorial structure. Perhaps the best example is Cheeger's inequality which relates spectral gaps to combinatorial expansion [27, 8, 9, 2, 39]. Here, we describe connections between spectral properties of boolean matrices and the sign rank of their signed versions.

Proving strong lower bounds on the sign rank of sign matrices turned out to be a difficult task. Alon, Frankl, and Rödl [6] were the first to prove that there are sign matrices with high sign rank, but they have not provided explicit examples. Later on, a breakthrough of [31] showed how to prove lower bounds on the sign rank of explicit matrices, proving, specifically, that Hadamard matrices have high sign rank. [56] proved that there is a function that is computed by a small depth three boolean circuit, but with high sign rank. It is worth mentioning that no explicit matrix whose sign rank is significantly larger than $N^{\frac{1}{2}}$ is known.

We focus on the case of regular matrices. A boolean matrix is Δ regular if every row and every column in it has exactly Δ ones, and a sign matrix is Δ regular if its boolean version is Δ regular.

An $N \times N$ real matrix M has N singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_N \geq 0$. The largest singular value of M is also called its spectral norm $\|M\| = \sigma_1 = \max\{\|Mx\| : \|x\| \leq 1\}$, where $\|x\|^2 = \langle x, x\rangle$ with the standard inner product. If the ratio $\sigma_2(M)/\|M\|$ is bounded away from one, or small, we say that M has a spectral gap.

We prove that if B has a spectral gap then the sign rank of S is high.

Theorem 6. *Let B be a Δ regular $N \times N$ boolean matrix with $\Delta \leq N/2$, and let S be its signed version. Then,*

$$\textit{sign-rank}(S) \geq \frac{\Delta}{\sigma_2(B)}.$$

In many cases a spectral gap for B implies that it has pseudorandom properties. This theorem is another manifestation of this phenomenon since random sign matrices have high sign rank (see [6]).

The theorem above provides a non trivial lower bound on the sign rank of S. There is a non trivial upper bound as well. The sign rank of a Δ regular sign matrix is at most $2\Delta + 1$. Here is a brief explanation of this upper bound (see [6] for a more detailed proof). Every row i in S

has at most 2Δ sign changes (i.e. columns j so that $S_{i,j} \neq S_{i,j+1}$). This implies that for every i, there is a real univariate polynomial G_i of degree at most 2Δ so that $G_i(j)S_{i,j} > 0$ for all $j \in [N] \subset \mathbb{R}$. To see how this corresponds to sign rank at most $2\Delta + 1$, recall that evaluating a polynomial G of degree 2Δ on a point $x \in \mathbb{R}$ corresponds to an inner product over $\mathbb{R}^{2\Delta+1}$ between the vector of coefficients of G, and the vector of powers of x.

Our proof of Theorem 6 and its limitations are discussed in detail in Section 3.3.

2 Applications

2.1 Learning theory

Universality of linear classifiers

Linear classifiers have been central in the study of machine learning since the introduction of the Perceptron algorithm in the 50's [58] and Support Vector Machines (SVM) in the 90's [21, 26]. The rising of kernel methods in the 90's [21, 62] enabled reducing many learning problems to the framework of halfspaces, making linear classifiers a central algorithmic tool.

These methods use the following two-step approach. First, embed the hypothesis class[3] in halfspaces of an Euclidean space (each point corresponds to a vector and for every hypothesis h, the vectors corresponding to $h^{-1}(1)$ and the vectors corresponding to $h^{-1}(-1)$ are separated by a hyperplane). Second, apply a learning algorithm for halfspaces.

If the embedding is to a low dimensional space then a good generalization rate is implied. For embeddings to large dimensional spaces, SVM theory offers an alternative parameter, namely the margin[4]. Indeed, a large margin also implies a good generalization rate. On the other hand, any embedding with a large margin can be projected to a low dimensional space using standard dimension reduction arguments [40, 12, 15].

Ben-David, Eiron, and Simon [15] utilized it to argue that "…any universal learning machine, which transforms data to a Euclidean space and then applies linear (or large margin) classification, cannot preserve good generalization bounds in general." Formally, they showed that: For any fixed $d > 1$, most hypothesis classes $C \subseteq \{\pm 1\}^N$ of VC dimension d have sign-rank of $N^{\Omega(1)}$. As discussed in Section 1.2, Theorem 4 quantitatively improves over their results.

In practice, linear classifiers are widely used in a variety of applications including handwriting recognition, image classification, medical science, bioinformatics, and more. The practical usefulness of linear classifiers and the argument of Ben-David, Eiron, and Simon manifest a gap between practice and theory that seems worth studying. We next discuss how Theorem 5, which provides a non-trivial upper bound on the sign rank, can be interpreted as a theoretical evidence which supports the practical usefulness of linear classifiers. Let $C \subseteq \{\pm 1\}^X$ be a hypothesis class, and let $\gamma > 0$. We say that C is *γ-weakly represented by halfspaces* if for every finite $Y \subseteq X$, the sign rank of $C|_Y$ is at most $O(|Y|^{1-\gamma})$. In other words, there exists an embedding of Y in $\mathbb{R}^k$ with $k = O(|Y|^{1-\gamma})$ such that each hypothesis in $C|_Y$ corresponds to

[3] In this context we use the more common term "hypothesis class" instead of "matrix."

[4] The margin of the embedding is the minimum over all hypotheses h of the distance between the convex hull of the vectors corresponding to $h^{-1}(1)$ and the convex hull of the vectors corresponding to $h^{-1}(-1)$

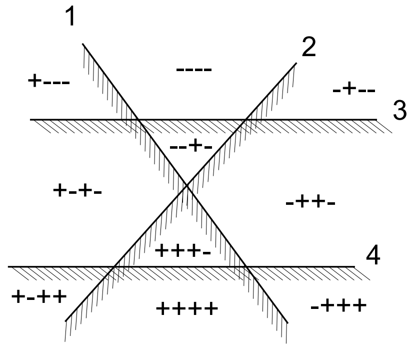

Figure 1: An arrangement of lines in the plane and the corresponding cells.

a halfspace in the embedding. Theorem 5 shows that any class C is γ-weakly represented by halfspaces where γ depends only on its VC dimension. Weak representations can be thought of as providing a compressed representation of $C|_Y$ using half-spaces in a dimension that is sublinear in $|Y|$. Such representations imply learnability; indeed, every γ-weakly represented class C is learnable, as the VC dimension of C is bounded from above by some function of of γ. While these quantitative relations between the VC dimension and γ may be rather loose, they show that in principle, any learnable class has a weak representation by halfspaces which certifies its learnability.

Maximum classes with large sign rank

Let $C \subseteq \{\pm 1\}^N$ be a class with VC dimension d. The class C is called maximum if it meets the Sauer-Shelah's bound [61] with equality[5]. That is, $|C| = \sum_{i=0}^{d} \binom{N}{i}$. Maximum classes were studied in different contexts such as machine learning, geometry, and combinatorics (e.g. [20, 30, 36, 13, 11, 45, 52, 59, 60]).

There are several known examples of maximum classes. A fairly simple one is the hamming ball of radius d, i.e., the class of all vectors with weight at most d. Another set of examples relates to the sign rank: Let H an arrangement of hyperplanes in $\mathbb{R}^d$. These hyperplanes cut $\mathbb{R}^d$ into cells; the connected components of $\mathbb{R}^d \setminus \left(\bigcup_{h \in H} h\right)$. Each cell c is associated with a sign vector $v_c \in \{\pm 1\}^H$ which describes the location of the cell relative to each of the hyperplanes. See Figure 2.1 for a planar arrangement. The sign rank of such a class is at most $d+1$. It is known (see e.g. [36]) that if the hyperplanes are in general position then the sign vectors of the cells form a maximum class of VC dimension d.

Gärtner and Welzl [36] gave a combinatorial characterization of maximum classes constructed using generic halfspaces. As an application of their characterization they note that hamming ball of radius d is a maximum class that can not be realized this way. By Lemma 18, however, the hamming ball of radius d has sign rank at most $2d+1$ (it is in fact exactly $2d+1$, since any set of $2d+1$ columns in its incidence matrix is antipodally shattered and therefore its

[5]Maximum classes are distinguished from maximal classes: A maximum class has the largest possible size among all classes of VC dimension d, and a maximal class is such that for every sign vector $v \notin C$, if v is added to C then the VC dimension is increased.

dual sign rank is at least $2d+1$). It is therefore natural to ask whether every maximum class has sign rank which depends only on d. A similar question was also asked by [15]. Theorem 8 in Section 2.2.1 gives a negative answer to this question, even when $d = 2$ (when $d = 1$, by Theorem 2 the sign rank is at most 3).

In machine learning, maximum classes were studied extensively in the context of sample compression schemes. A partial list of works in this context includes [30, 45, 59, 60, 53, 29]. [30] is the first paper that designed sample compression schemes for maximum classes. Later, [45] improved it to an unlabeled sample compression scheme. [59] constructed an even simpler unlabeled sample compression scheme for maximum classes. Their scheme uses an approach suggested by [45] and their analysis resolved a conjecture from [45]. A crucial part in their work is establishing the existence of an embedding of any maximum class of VC dimension d in an arrangement of piecewise-linear hyperplanes in $\mathbb{R}^d$. Theorem 8 below shows that even for VC dimension 2, there are maximum classes $C \subseteq \{\pm 1\}^N$ of sign rank $\Omega(N^{1/2}/\log N)$. Thus, in order to make the piecewise-linear arrangement in $\mathbb{R}^2$ linear the dimension of the space must significantly grow to $\Omega(N^{1/2}/\log N)$.

2.2 Explicit examples

The spectral lower bound on sign rank gives many explicit examples of matrices with high sign rank, which come from known constructions of expander graphs and combinatorial designs. A rather simple such family of examples is finite projective geometries.

Let $d \geq 2$ and $n \geq 3$. Let P be the set of points in a d dimensional projective space of order n, and let H be the set of hyperplanes in the space. For $d = 2$, this is just a projective plane with points and lines. It is known (see, e.g., [17]) that

$$|P| = |H| = N_{n,d} := n^d + n^{d-1} + \ldots + n + 1 = \frac{n^{d+1}-1}{n-1}.$$

Let $A \in \{\pm 1\}^{P \times H}$ be the signed point-hyperplane incidence matrix:

$$A_{p,h} = \begin{cases} 1 & p \in h, \\ -1 & p \notin h. \end{cases}$$

Theorem 7. *The matrix A is $N \times N$ with $N = N_{n,d}$, its VC dimension is d, and its sign rank is larger than*

$$\frac{n^d - 1}{n^{\frac{d-1}{2}}(n-1)} \geq N^{\frac{1}{2} - \frac{1}{2d}}.$$

The theorem follows from known properties of projective spaces (see Section 3.4.1). A slightly weaker (but asymptotically equivalent) lower bound on the sign rank of A was given by [32].

The sign rank of A is at most $2N_{n,d-1} + 1 = O(N^{1-\frac{1}{d}})$, due to the observation in [6] mentioned after Theorem 6. To see this, note that A is $N_{n,d-1}$ regular as every point in the projective space is incident to $N_{n,d-1}$ hyperplanes, and every hyperplane contains $N_{n,d-1}$ points.

Other explicit examples come from spectral graph theory. Here is a brief description of matrices that are even more restricted than having VC dimension 2 but have high sign rank; no

3 columns in them have more than 6 distinct projections. An (N, Δ, λ)-graph is a Δ regular graph on N vertices so that the absolute value of every eigenvalue of the graph besides the top one is at most λ. There are several known constructions of (N, Δ, λ)-graphs for which $\lambda \leq O(\sqrt{\Delta})$, that do not contain short cycles. Any such graph with $\Delta \geq N^{\Omega(1)}$ provides an example with sign rank at least $N^{\Omega(1)}$, and if there is no cycle of length at most 6 then in the sign matrix we have at most 6 distinct projections on any set of 3 columns.

2.2.1 Maximum classes

Let P be the set of points in a projective plane of order n and let L be the set of lines in it. Let $N = N_{n,2} = |P| = |L|$. For each line $\ell \in L$, fix some linear order on the points in ℓ. A set $T \subset P$ is called an interval if $T \subseteq \ell$ for some line $\ell \in L$, and T forms an interval with respect to the order we fixed on ℓ.

Theorem 8. *The class R of all intervals is a maximum class of VC dimension 2. Moreover, there exists a choice of linear orders for the lines in L such that the resulting R has sign rank $\Omega(N^{1/2}/\log N)$.*

The proof of Theorem 8 is given in Section 3.4.1. The proof does not follow directly from Theorem 4 since it is not clear that the classes with VC dimension 2 and large sign rank which are guaranteed to exist by Theorem 4 can be extended to a maximum class.

2.3 Computing the sign rank

Linear Programming (LP) is one of the most famous and useful problems in the class P. As a decision problem, an LP problem concerns determining the satisfiability of a system

$$\ell_i(x) \geq 0,\ i = 1, \ldots, m$$

where each ℓ_i is an affine function defined over $\mathbb{R}^n$ (say with integer coefficients). A natural extension of LP is to consider the case in which each ℓ_i is a multivariate polynomial. Perhaps not surprisingly, this problem is much harder than LP. In fact, satisfiability of a system of polynomial inequalities is known to be a complete problem for the class $\exists\mathbb{R}$. The class $\exists\mathbb{R}$ is known to lie between PSPACE and NP (see [49] and references within).

Consider the problem of deciding whether the sign rank of a given $N \times N$ sign matrix is at most k. A simple reduction shows that to solve this problem it is enough to decide whether a system of real polynomial inequalities is satisfiable. Thus, this problem belongs to the class $\exists\mathbb{R}$. [14][6], and [18] showed that deciding if the sign rank is at most 3 is NP-hard, and that deciding if the sign rank is at most 2 is in P. Both [14], and [18] established the NP-hardness of deciding whether the sign-rank is at most 3 by a reduction from the problem of determining stretchacility of pseudo-line arrangements. This problem concerns whether a given combinatorial description of an arrangement of pseudo-lines can be realized ("stretched") by an arrangement of lines. [49], based on the works of [51], [65], and [57] showed that determining stretchability of pseudo-line

[6] Interestingly, their motivation for considering sign rank comes from image processing.

arrangements is in fact $\exists\mathbb{R}$-complete. Therefore, it follows[7] that determining whether the sign-rank is at most 3 is $\exists\mathbb{R}$-complete.

Another related work of [46] concerns the problem of computing the approximate rank of a sign matrix, for which they provide an approximation algorithm. They pose the problem of efficiently approximating the sign rank as an open problem.

Using an idea similar to the one in the proof of Theorem 5 we derive an approximation algorithm for the sign rank (see Section 3.4.2).

Theorem 9. *There exists a polynomial time algorithm that approximates the sign rank of a given N by N matrix up to a multiplicative factor of $c \cdot N/\log(N)$ where $c > 0$ is a universal constant.*

2.4 Communication complexity

We briefly explain the notions from communication complexity we use. For formal definitions, background and more details, see the textbook [44].

For a function f and a distribution μ on its inputs, define $D_\mu(f)$ as the minimum communication complexity of a protocol that correctly computes f with error $1/3$ over inputs from μ. Define $D^\times(f) = \max\{D_\mu(f) : \mu \text{ is a product distribution}\}$. Define the unbounded error communication complexity $U(f)$ of f as the minimum communication complexity of a randomized private-coin[8] protocol that correctly computes f with probability strictly larger than $1/2$ on every input.

Two works of [64, 63] showed that there are functions with small distributional communication complexity under product distributions, and large unbounded error communication complexity. In [64] the separation is as strong as possible but it is not for an explicit function, and the separation in [63] is not as strong but the underlying function is explicit.

The matrix A with $d = 2$ and $n \geq 3$ in our example from Section 2.2 corresponds to the following communication problem: Alice gets a point $p \in P$, Bob gets a line $\ell \in L$, and they wish to decide whether $p \in \ell$ or not. Let $f : P \times L \to \{0, 1\}$ be the corresponding function and let $m = \lceil \log_2(N) \rceil$. A trivial protocol would be that Alice sends Bob using m bits the name of her point, Bob checks whether it is incident to the line, and outputs accordingly.

Theorem 7 implies the following consequences. Even if we consider protocols that use randomness and are allowed to err with probability less than but arbitrarily close to $\frac{1}{2}$, then still one cannot do considerably better than the above trivial protocol. However, if the input $(p, \ell) \in P \times L$ is distributed according to a product distribution then there exists an $O(1)$ protocol that errs with probability at most $\frac{1}{3}$.

Corollary 10. *The unbounded error communication complexity of f is[9] $U(f) \geq \frac{m}{4} - O(1)$. The distributional communication complexity of f under product distributions is $D^\times(f) \leq O(1)$.*

[7] [49] considers a different type of combinatorial description than [14, 18], and therefore considered a different formulation of the stretchability problem. However, it is possible to transform between these descriptions in polynomial time.

[8] In the public-coin model, every boolean function has unbounded communication complexity at most two.

[9] By taking larger values of d, the constant $\frac{1}{4}$ may be increased to $\frac{1}{2} - \frac{1}{2d}$.

These two seemingly contradicting facts are a corollary of the high sign rank and the low VC dimension of A, using two known results. The upper bound on $D^{\times}(f)$ follows from the fact that VCdim$(A) = 2$, and the work of [43] which used the PAC learning algorithm to construct an efficient (one round) communication protocol for f under product distributions. The lower bound on $U(f)$ follows from that sign-rank$(A) \geq \Omega(N^{1/4})$, and the result of [55] that showed that unbounded error communication complexity is equivalent to the logarithm of the sign rank. See [64] for more details.

2.5 Counting VC classes

Let $c(N, d)$ denote the number of classes $C \subseteq \{\pm 1\}^N$ with VC dimension d. We give the following estimate of $c(N, d)$ for constant d and N large enough. The proof is given in Section 3.4.3.

Theorem 11. *For every $d > 0$, there is $N_0 = N_0(d)$ such that for all $N > N_0$:*

$$N^{\Omega(N^d/d^{d+1})} \leq c(N, d) \leq N^{O(N)^d}.$$

Let $m(N, d)$ denote the number of maximum classes $C \subseteq \{\pm 1\}^N$ of VC dimension d. The problem of estimating $m(N, d)$ was proposed by [35]. We provide the following estimate (see Section 3.4.3).

Theorem 12. *For every $d > 1$, there is $N_0 = N_0(d)$ such that for all $N > N_0$:*

$$N^{(1+o(1))\frac{1}{d+1}\binom{N}{d}} \leq m(N, d) \leq N^{(1+o(1))\sum_{i=1}^{d}\binom{N}{i}}.$$

The gap between our upper and lower bound is roughly a multiplicative factor of $d + 1$ in the exponent. In the previous bounds given by [35] the gap was a multiplicative factor of N in the exponent.

2.6 Counting graphs

Here we describe an application of our method for proving Theorem 5 to counting graphs with a given forbidden substructure.

Let $G = (V, E)$ be a graph (not necessarily bipartite). The universal graph $U(d)$ is defined as the bipartite graph with two color classes A and $B = 2^A$ where $|A| = d$, and the edges are defined as $\{a, b\}$ iff $a \in b$. The graph G is called $U(d)$-free if for all two disjoint sets of vertices $A, B \subset V$ so that $|A| = d$ and $|B| = 2^d$, the bipartite graph consisting of all edges of G between A and B is not isomorphic to $U(d)$. In Theorem 24 of [5], which improves Theorem 2 there, it is proved that for $d \geq 2$, the number of $U(d + 1)$-free graphs on N vertices is at most

$$2^{O(N^{2-1/d}(\log N)^{d+2})}.$$

The proof in [5] is quite involved, consisting of several technical and complicated steps. Our methods give a different, quick proof of an improved estimate, replacing the $(\log N)^{d+2}$ term by a single $\log N$ term.

Theorem 13. *For every fixed $d \geq 1$, the number of $U(d+1)$-free graphs on N vertices is at most $2^{O(N^{2-1/d}\log N)}$.*

The proof of the theorem is given in Section 3.4.4.

2.7 Geometry

Differences and similarities between finite geometries and real geometry are well known. An example of a related problem is finding the minimum dimension of Euclidean space in which we can embed a given finite plane (i.e. a collection of points and lines satisfying certain axioms). By embed we mean that there are two one-to-one maps e_P, e_L so that $e_P(p) \in e_L(\ell)$ iff $p \in \ell$ for all $p \in P, \ell \in L$. The Sylvester-Gallai theorem shows, for example, that Fano's plane cannot be embedded in any finite dimensional real space if points are mapped to points and lines to lines.

How about a less restrictive meaning of embedding? One option is to allow embedding using half spaces, that is, an embedding in which points are mapped to points but lines are mapped to half spaces. Such embedding is always possible if the dimension is high enough: Every plane with point set P and line set L can be embedded in $\mathbb{R}^P$ by choosing $e_P(p)$ as the p'th unit vector, and $e_L(\ell)$ as the half space with positive projection on the vector with 1 on points in ℓ and -1 on points outside ℓ. The minimum dimension for which such an embedding exists is captured by the sign rank of the underlying incidence matrix; namely it is either the sign rank or the sign rank minus one.

Corollary 14. *A finite projective plane of order $n \geq 3$ cannot be embedded in $\mathbb{R}^k$ using half spaces, unless $k > N^{1/4} - 1$ with $N = n^2 + n + 1$.*

Roughly speaking, the corollary says that there are no efficient ways to embed finite planes in real space using half spaces.

3 Proofs

3.1 Duality

Here we discuss the connection between VC dimension and dual sign rank.

We start with an equivalent definition of dual sign rank, that is based on the following notion. We say that a set of columns C is *antipodally shattered* in a sign matrix S if for each $v \in \{\pm 1\}^C$, either v or $-v$ appear as a row in the restriction of S to the columns in C. Equivalently, C is antipodally shattered if it is shattered in the matrix $\begin{bmatrix} S \\ -S \end{bmatrix}$.

Claim 15. *The set of columns C is antipodally shattered in S if and only if in every matrix M with sign$(M) = S$ the columns in C are linearly independent.*

Proof. First, assume C is such that there exists some M with $\text{sign}(M) = S$ in which the columns in C are linearly dependent. For a column $j \in C$, denote by $M(j)$ the j'th column in M. Let $\{\alpha_j : j \in C\}$ be a set of real numbers so that $\sum_{j \in C} \alpha_j M(j) = 0$ and not all α_j's are zero. Consider the vector $v \in \{\pm 1\}^C$ such that $v_j = 1$ if $\alpha_j \geq 0$ and $v_j = -1$ if $\alpha_j < 0$. The restriction of S to C does not contain v nor $-v$ as a row, which certifies that C is not antipodally shattered by S.

Second, let C be a set of columns which is not antipodally shattered in S. Let $v \in \{\pm 1\}^C$ be such that both $v, -v$ do not appear as a row in the restriction of S to C. Consider the subspace $U = \{u \in \mathbb{R}^C : \sum_{j \in C} u_j v_j = 0\}$. For each sign vector $s \in \{\pm 1\}^C$ so that $s \neq \pm v$, the space U contains some vector u_s such that $\text{sign}(u_s) = s$. Let M be so that $\text{sign}(M) = S$ and in addition for each row in S that has pattern $s \in \{\pm\}^C$ in S restricted to C, the corresponding row in M restricted to C is $u_s \in U$. All rows in M restricted to C are in U, and therefore the set $\{M(j) : j \in C\}$ is linearly dependent. □

Now, we prove Proposition 1:

Proposition (Restatement of Proposition 1)**.** *The dual sign rank of S is the VC dimension of the matrix* $\begin{bmatrix} S \\ -S \end{bmatrix}$*. As a corollary:* $VC(S) \leq$ *dual-sign-rank*$(S) \leq 2VC(S) + 1$.

Proof. That the dual sign rank is the VC dimension of $\begin{bmatrix} S \\ -S \end{bmatrix}$, is an immediate corollary of Claim 15. Next we show that $VC(S) \leq \text{dual-sign-rank}(S) \leq 2VC(S) + 1$. The left inequality: The VC dimension of S is at most the maximum size of a set of columns that is antipodally shattered in S, which by the above claim equals the dual sign rank of S. The right inequality: Let C be a largest set of columns that is antipodally shattered in S. By the claim above, the dual sign rank of S is $|C|$. Let $A \subseteq C$ such that $|A| = \lfloor |C|/2 \rfloor$. If A is shattered in S then we are done. Otherwise, there exists some $v \in \{\pm 1\}^A$ that does not appear in S restricted to A. Since C is antipodally shattered by S, this implies that S contains all patterns in $\{\pm 1\}^C$ whose restriction to A is $-v$. In particular, S shatters $C \setminus A$ which is of size at least $\lfloor |C|/2 \rfloor$. □

3.2 Sign rank versus VC dimension

In this section we study the maximum possible sign rank of $N \times N$ matrices with VC dimension d, presenting the proofs of Proposition 1 and Theorems 5 and 4. We also show that the arguments supply a new, short proof and an improved estimate for a problem in asymptotic enumeration of graphs studied by [5].

3.2.1 VC dimension one

Our goal in this section is to show that sign matrices with VC dimension one have sign rank at most 3, and that 3 is tight. Before reading this section, it may be a nice exercise to prove that the sign rank of the $N \times N$ signed identity matrix is exactly three (for $N \geq 4$).

Let us start by recalling a geometric interpretation of sign rank. Let M by an $R \times C$ sign matrix. A d-dimensional embedding of M using half spaces consists of two maps e_R, e_C so that for every row $r \in [R]$ and column $c \in [C]$, we have that $e_R(r) \in \mathbb{R}^d$, $e_C(c)$ is a half space in $\mathbb{R}^d$, and $M_{r,c} = 1$ iff $e_R(r) \in e_C(c)$. The important property for us is that if M has a d-dimensional embedding using half spaces then its sign rank is at most $d + 1$. The $+1$ comes from the fact that the hyperplanes defining the half spaces do not necessarily pass through the origin.

Our goal in this section is to embed M with VC dimension one in the plane using half spaces. The embedding is constructive and uses the following known claim (see, e.g., Theorem 11 in [28]).

Claim 16 ([28])**.** *Let M be an $R \times C$ sign matrix with VC dimension one so that no row appears twice in it, and every column c is shattered (i.e. the two values ± 1 appear in it). Then, there is a column $c_0 \in [C]$ and a row $r_0 \in [R]$ so that $M_{r_0,c_0} \neq M_{r,c_0}$ for all $r \neq r_0$ in $[R]$.*

Proof. For every column c, denote by ones_c the number of rows $r \in [R]$ so that $M_{r,c} = 1$, and let $m_c = \min\{\mathit{ones}_c, R - \mathit{ones}_c\}$. Assume without loss of generality that $m_1 \leq m_c$ for all c, and that $m_1 = \mathit{ones}_1$. Since all columns are shattered, $m_1 \geq 1$. To prove the claim, it suffices to show that $m_1 \leq 1$.

Assume towards a contradiction that $m_1 \geq 2$. For $b \in \{1, -1\}$, denote by $M^{(b)}$ the submatrix of M consisting of all rows r so that $M_{r,1} = b$. The matrix $M^{(1)}$ has at least two rows. Since all rows are different, there is a column $c \neq 1$ so that two rows in $M^{(1)}$ differ in c. Specifically, column c is shattered in $M^{(1)}$. Since $\text{VCdim}(M) = 1$, it follows that c is not shattered in $M^{(-1)}$, which means that the value in column c is the same for all rows of the matrix $M^{(-1)}$. Therefore, $m_c < m_1$, which is a contradiction. □

The embedding we construct has an extra structure which allows the induction to go through: The rows are mapped to points on the unit circle (i.e. set of points $x \in \mathbb{R}^2$ so that $\|x\| = 1$).

Lemma 17. *Let M be an $R \times C$ sign matrix of VC dimension one so that no row appears twice in it. Then, M can be embedded in $\mathbb{R}^2$ using half spaces, where each row is mapped to a point on the unit circle.*

The lemma immediately implies Threorem 2 due to the connection to sign rank discussed above.

Proof. The proof follows by induction on C. If $C = 1$, the claim trivially holds.

The inductive step: If there is a column that is not shattered, then we can remove it, apply induction, and then add a half space that either contains or does not contain all points, as necessary. So, we can assume all columns are shattered. By Claim 16, we can assume without loss of generality that $M_{1,1} = 1$ but $M_{r,1} = -1$ for all $r \neq 1$.

Denote by r_0 the row of M so that $M_{r_0,c} = M_{1,c}$ for all $c \neq 1$, if such a row exists. Let M' be the matrix obtained from M by deleting the first column, and row r_0 if it exists, so that no row in M' appears twice. By induction, there is an appropriate embedding of M' in $\mathbb{R}^2$.

The following is illustrated in Figure 2. Let $x \in \mathbb{R}^2$ be the point on the unit circle to which the first row in M' was mapped to (this row corresponds to the first row of M as well). The half spaces in the embedding of M' are defined by lines, which mark the borders of the half spaces. The unit circle intersects these lines in finitely many points. Let y, z be the two closest points to x among all these intersection points. Let y' be the point on the circle in the middle between x, y, and let z' be the point on the circle in the middle between x, z. Add to the configuration one more half space which is defined by the line passing through y', z'. If in addition row r_0 exists, then map r_0 to the point x_0 on the circle which is right in the middle between y, y'.

This is the construction. Its correctness follows by induction, by the choice of the last added half space which separates x from all other points, and since if x_0 exists it belongs to the same cell as x in the embedding of M'. □

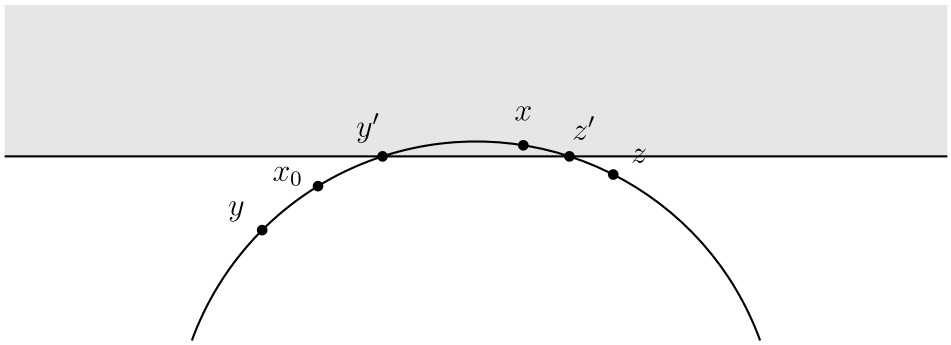

Figure 2: An example of a neighbourhood of x. All other points in embedding of M' are to left of y and right of z on the circle. The half space defined by the line through y', z' is coloured light gray.

We conclude the section by showing that the bound 3 above cannot be improved.

Proof of Claim 3. One may deduce the claim from Forster's argument, but we provide a more elementary argument. It suffices to consider the case $N = 4$. Consider an arrangement of four half planes in $\mathbb{R}^2$. These four half planes partition $\mathbb{R}^2$ to eight cones with different sign signatures, as illustrated in Figure 2. Let M be the 8×4 sign matrix whose rows are these sign signatures. The rows of M form a distance preserving cycle (i.e. the distance along cycle is hamming distance) of length eight in the discrete cube of dimension four[10].

Finally, the signed identity matrix is not a submatrix of M. To see this, note that the four rows of the signed identity matrix have pairwise hamming distance two, but there are no such four points (not even three points) on this cycle of length eight.

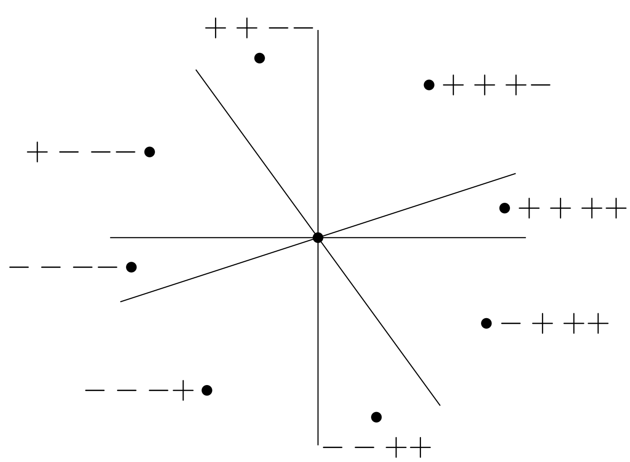

Figure 3: Four lines defining four half planes, and the corresponding eight sign signatures.

[10]The graph with vertex set $\{\pm 1\}^4$ where every two vectors of hamming distance one are connected by an edge.

□

3.2.2 The upper bound

In this subsection we prove Theorem 5. The proof is short, but requires several ingredients. The first one has been mentioned already, and appears in [6]. For a sign matrix S, let $SC(S)$ denote the maximum number of sign changes (SC) along a column of S. Define $SC^*(S) = \min SC(M)$ where the minimum is taken over all matrices M obtained from S by a permutation of the rows.

Lemma 18 ([6])**.** *For any sign matrix S, sign-rank$(S) \leq SC^*(S) + 1$.*

Of course we can replace here rows by columns, but for our purpose the above version will do. The second result we need is a theorem of [69] (see also [24]). As observed, for example, in [50], plugging in its proof a result of [37] improves it by a logarithmic factor, yielding the result we describe next. For a function g mapping positive integers to positive integers, we say that a sign matrix S satisfies a primal shatter function g if for any integer t and any set I of m columns of S, the number of distinct projections of the rows of S on I is at most $g(t)$. The result of Welzl (after its optimization following [37]) can be stated as follows[11].

Lemma 19 ([69], see also [24, 50])**.** *Let S be a sign matrix with N rows that satisfies the primal shatter function $g(t) = ct^d$ for some constants $c \geq 0$ and $d > 1$. Then $SC^*(S) \leq O(N^{1-1/d})$.*

Proof of Theorem 5. Let S be an $N \times N$ sign matrix of VC dimension $d > 1$. By Sauer's lemma [61], it satisfies the primal shatter function $g(t) = t^d$. Hence, by Lemma 19, $SC^*(S) \leq O(N^{1-1/d})$. Therefore, by Lemma 18, sign-rank$(S) \leq O(N^{1-1/d})$. □

On the tightness of the argument. The proof of Theorem 5 works, with essentially no change, for a larger class of sign matrices than the ones with VC dimension d. Indeed, the proof shows that the sign rank of any $N \times N$ matrix with primal shatter function at most ct^d for some fixed c and $d > 1$ is at most $O(N^{1-1/d})$. In this statement the estimate is sharp for all integers d, up to a logarithmic factor. This follows from the construction in [10], which supplies $N \times N$ boolean matrices so that the number of 1 entries in them is at least $\Omega(N^{2-1/d})$, and they contain no d by $D = (d-1)! + 1$ submatrices of 1's. These matrices satisfy the primal shatter function $g(t) = D\binom{t}{d} + \sum_{i=0}^{d-1}\binom{t}{i}$ (with room to spare). Indeed, if we have more than that many distinct projections on a set of t columns, we can omit all projections of weight at most $d-1$. Each additional projection contains 1's in at least one set of size d, and the same d-set cannot be covered more than D times. Plugging this matrix in the counting argument that gives a lower bound for the sign rank using Lemma 21 proven below supplies an $\Omega(N^{1-1/d}/\log N)$ lower bound for the sign rank of many $N \times N$ matrices with primal shatter function $O(t^d)$.

We have seen in Lemma 18 that sign rank is at most of order SC^*. Moreover, for a fixed r, many of the $N \times N$ sign matrices with sign rank at most r also have SC^* at most r: Indeed, a

[11] The statement in [69] and the subsequent papers is formulated in terms of somewhat different notions, but it is not difficult to check that it is equivalent to the statement below.

simple counting argument shows that the number of $N \times N$ sign matrices M with $SC(M) < r$ is

$$\left(2 \cdot \sum_{i=0}^{r-1} \binom{N-1}{i}\right)^N = 2^{\Omega(rN \log N)},$$

so, the set of $N \times N$ sign matrices with $SC^*(M) < r$ is a subset of size $2^{\Omega(rN \log N)}$ of all $N \times N$ sign matrices with sign rank at most r.

How many $N \times N$ matrices of sign rank at most r are there? by Lemma 21 proved in the next section, this number is at most $2^{O(rN \log N)}$. So, the set of matrices with $SC^* < r$ is a rather large subset of the set of matrices with sign rank at most r.

It is reasonable, therefore, to wonder whether an inequality in the other direction holds. Namely, whether all matrices of sign rank r have SC^* order of r. We now describe an example which shows that this is far from being true, and also demonstrates the tightness of Lemma 19. Namely, for every constant $d > 1$, there are $N \times N$ matrices S, which satisfy the primal shatter function $g(t) = ct^d$ for a constant c, and on the other hand $SC^*(S) \geq \Omega(N^{1-1/d})$. Consider the grid of points $P = [n]^d$ as a subset of $\mathbb{R}^d$. Denote by $e_1, \ldots, e_d$ the standard unit vectors in $\mathbb{R}^d$. For $i \in [n-1]$ and $j \in [d]$, define the hyperplane $h_{i,j} = \{x : \langle x, e_j \rangle > i + (1/2)\}$. Denote by H the set of these $d(n-1)$ axis parallel hyperplanes. Let S be the $P \times H$ sign matrix defined by P and H. That is, $S_{p,h} = 1$ iff $p \in h$. First, the matrix S satisfies the primal shatter function ct^d, since every family of t hyperplanes partition $\mathbb{R}^d$ to at most ct^d cells. Second, we show that

$$SC^*(S) \geq \frac{n^d - 1}{d(n-1)} \geq \frac{|P|^{1-1/d}}{d}.$$

Indeed, fix some order on the rows of S, that is, order the points $P = \{p_1, \ldots, p_N\}$ with $N = |P|$. The key point is that one of the hyperplanes $h_0 \in H$ is so that the number of $i \in [N-1]$ for which $S_{p_i,h_0} \neq S_{p_{i+1},h_0}$ is at least $(n^d - 1)/(d(n-1))$: For each i there is at least one hyperplane h that separates p_i and p_{i+1}, that is, for which $S_{p_i,h} \neq S_{p_{i+1},h}$. The number of such pairs of points is $n^d - 1$, and the number of hyperplanes is just $d(n-1)$.

3.2.3 The lower bound

In this subsection we prove Theorem 4. Our approach follows the one of [6], which is based on known bounds for the number of sign patterns of real polynomials. A similar approach has been subsequently used by [15] to derive lower bounds for $f(N, d)$ for $d \geq 4$, but here we do it in a slightly more sophisticated way and get better bounds.

Although we can use the estimate in [6] for the number of sign matrices with a given sign rank, we prefer to describe the argument by directly applying a result of [68], described next.

Let $P = (P_1, P_2, \ldots, P_m)$ be a list of m real polynomials, each in ℓ variables. Define the semi-variety

$$V = V(P) = \{x \in \mathbb{R}^\ell : P_i(x) \neq 0 \text{ for all } 1 \leq i \leq m\}.$$

For $x \in V$, the sign pattern of P at x is the vector

$$(sign(P_1(x)), sign(P_2(x)), \ldots, sign(P_m(x))) \in \{-1, 1\}^m.$$

Let $s(P)$ be the total number of sign patterns of P as x ranges over all of V. This number is bounded from above by the number of connected components of V.

Theorem 20 ([68]). *Let $P = (P_1, P_2, \ldots, P_m)$ be a list of real polynomials, each in ℓ variables and of degree at most k. If $m \geq \ell$ then the number of connected components of $V(P)$ (and hence also $s(P)$) is at most $(4ekm/\ell)^\ell$.*

An $N \times N$ matrix M is of rank at most r iff it can be written as a product $M = M_1 \cdot M_2$ of an $N \times r$ matrix M_1 by an $r \times N$ matrix M_2. Therefore, each entry of M is a quadratic polynomial in the $2Nr$ variables describing the entries of M_1 and M_2. We thus deduce the following from Warren's Theorem stated above. A similar argument has been used by [16].

Lemma 21. *Let $r \leq N/2$. Then, the number of $N \times N$ sign matrices of sign rank at most r does not exceed $(O(N/r))^{2Nr} \leq 2^{O(rN \log N)}$.*

For a fixed r, this bound for the logarithm of the above quantity is tight up to a constant factor: As argued in Subsection 3.2.2, there are at least some $2^{\Omega(rN \log N)}$ matrices of sign rank r.

In order to derive the statement of Theorem 4 from the last lemma it suffices to show that the number of $N \times N$ sign matrices of VC dimension d is sufficiently large. We proceed to do so. It is more convenient to discuss boolean matrices in what follows (instead of their signed versions).

Proof of Theorem 4. There are 4 parts as follows.
1. The case $d = 2$: Consider the $N \times N$ incidence matrix A of the projective plane with N points and N lines, considered in the previous sections. The number of 1 entries in A is $(1 + o(1))N^{3/2}$, and it does not contain $J_{2\times 2}$ (the 2×2 all 1 matrix) as a submatrix, since there is only one line passing through any two given points. Therefore, any matrix obtained from it by replacing ones by zeros has VC dimension at most 2, since every matrix of VC dimension 3 must contain $J_{2\times 2}$ as a submatrix. This gives us $2^{(1+o(1))N^{3/2}}$ distinct $N \times N$ sign matrices of VC dimension at most 2. Lemma 21 therefore establishes the assertion of Theorem 4, part 1.
2. The case $d = 3$: Call a 5×4 boolean matrix heavy if its rows are the all 1 row and the 4 rows with Hamming weight 3. Call a 5×4 boolean matrix heavy-dominating if there is a heavy matrix which is smaller or equal to it in every entry.

We claim that there is a boolean $N \times N$ matrix B so that the number of 1 entries in it is at least $\Omega(N^{23/15})$, and it does not contain any heavy-dominating 5×4 submatrix. Given such a matrix B, any matrix obtained from B by replacing some of the ones by zeros have VC dimension at most 3. This implies part 2 of Theorem 4, using Lemma 21 as before.

The existence of B is proved by a probabilistic argument. Let C be a random boolean matrix in which each entry, randomly and independently, is 1 with probability $p = \frac{1}{2N^{7/15}}$. Let X be the random variable counting the number of 1 entries of C minus twice the number of 5×4 heavy-dominant submatrices C contains. By linearity of expectation,

$$\mathbb{E}(X) \geq N^2 p - 2N^{4+5} p^{1\cdot 4+4\cdot 3} = \Omega(N^{23/15}).$$

Fix a matrix C for which the value of X is at least its expectation. Replace at most two 1 entries by 0 in each heavy-dominant 5×4 submatrix in C to get the required matrix B.

3. The case $d = 4$: The basic idea is as before, but here there is an explicit construction that beats the probabilistic one. Indeed, [22] constructed an $N \times N$ boolean matrix B so that the number of 1 entries in B is at least $\Omega(N^{5/3})$ and it does not contain $J_{3\times 3}$ as a submatrix (see also [10] for another construction). No set of 5 rows in every matrix obtained from this one by replacing 1's by 0's can be shattered, implying the desired result as before.

4. The case $d > 4$: The proof here is similar to the one in part 2. We prove by a probabilistic argument that there is an $N \times N$ boolean matrix B so that the number of 1 entries in it is at least

$$\Omega(N^{2-(d^2+5d+2)/(d^3+2d^2+3d)})$$

and it contains no heavy-dominant submatrix. Here, heavy-dominant means a $1+(d+1)+\binom{d+1}{2}$ by $d+1$ matrix that is bigger or equal in each entry than the matrix whose rows are all the distinct vectors of length $d+1$ and Hamming weight at least $d-1$. Any matrix obtained by replacing 1's by 0's in B cannot have VC dimension exceeding d. The result follows, again, from Lemma 21.

We start as before with a random matrix C in which each entry, randomly and independently, is chosen to be 1 with probability

$$p = \frac{1}{2} \cdot N^{\frac{2-1-(d+1)-\binom{d+1}{2}-(d+1)}{1\cdot(d+1)+(d+1)\cdot d+\binom{d+1}{2}\cdot(d-1)-1}} = \frac{1}{2N^{(d^2+5d+2)/(d^3+2d^2+3d)}}.$$

Let X be the random variable counting the number of 1 entries of C minus three times the number of heavy-dominant submatrices C contains. As before, $\mathbb{E}(X) \geq \Omega(N^2 p)$, and by deleting some of the 1's in C we get B. □

3.3 Sign rank and spectral gaps

The lower bound on the sign rank uses Forster's argument [31], who showed how to relate sign rank to spectral norm. He proved that if S is an $N \times N$ sign matrix then

$$\text{sign-rank}(S) \geq \frac{N}{\|S\|}.$$

We would like to apply Forster's theorem to the matrix S in our explicit examples. The spectral norm of S, however, is too large to be useful: If S is $\Delta \leq N/3$ regular and x is the all 1 vector then $Sx = (2\Delta - N)x$ and so $\|S\| \geq N/3$. Applying Forster's theorem to S yields that its sign rank is $\Omega(1)$, which is not informative.

Our solution is based on the observation that Forster's argument actually proves a stronger statement. His proof works as long as the entries of the matrix are not too close to zero, as was already noticed in [32]. We therefore use a variant of the spectral norm of a sign matrix S which we call star norm and denote by[12]

$$\|S\|^* = \min\{\|M\| : M_{i,j}S_{i,j} \geq 1 \text{ for all } i, j\}.$$

[12]The minimizer belongs to a closed subset of the bounded set $\{M : \|M\| \leq \|S\|\}$.

Three comments seem in place. (i) We do not think of the star norm as a norm. (ii) It is always at most the spectral norm, $\|S\|^* \leq \|S\|$. (iii) Every M in the above minimum satisfies sign-rank(M) = sign-rank(S).

Theorem 22 ([32])**.** *Let S be an $N \times N$ sign matrix. Then,*

$$\textit{sign-rank}(S) \geq \frac{N}{\|S\|^*}.$$

For completeness, in Section 3.3.2 we provide a short proof of this theorem (which uses the main lemma from [31] as a black box). To get any improvement using this theorem, we must have $\|S\|^* \ll \|S\|$. It is not a priori obvious that there is a matrix S for which this holds. The following lemma shows that spectral gaps yield such examples.

Theorem 23. *Let S be a Δ regular $N \times N$ sign matrix with $\Delta \leq N/2$, and B its boolean version. Then,*

$$\|S\|^* \leq \frac{N \cdot \sigma_2(B)}{\Delta}.$$

In other words, every regular sign matrix whose boolean version has a spectral gap has a small star norm. Theorem 22 and Theorem 23 immediately imply Theorem 6. In Section 2.2, we provided concrete examples of matrices with a spectral gap, that have applications in communication complexity, learning theory and geometry.

Proof of Theorem 23. Define the matrix

$$M = \frac{N}{\Delta} B - J.$$

Observe that since $N \geq 2\Delta$ it follows that $M_{i,j} S_{i,j} \geq 1$ for all i, j. So,

$$\|S\|^* \leq \|M\|.$$

Since B is regular, the all 1 vector y is a right singular vector of B with singular value Δ. Specifically, $My = 0$. For every x, write $x = x_1 + x_2$ where x_1 is the projection of x on y and x_2 is orthogonal to y. Thus,

$$\langle Mx, Mx \rangle = \langle Mx_2, Mx_2 \rangle = \frac{N^2}{\Delta^2} \langle Bx_2, Bx_2 \rangle.$$

Note that $\|B\| \leq \Delta$ (and hence $\|B\| = \Delta$). Indeed, since B is regular, there are Δ permutation matrices $B^{(1)}, \ldots, B^{(\Delta)}$ so that B is their sum. The spectral norm of each $B^{(i)}$ is one. The desired bound follows by the triangle inequality.

Finally, since x_2 is orthogonal to y,

$$\|Bx_2\| \leq \sigma_2(B) \cdot \|x_2\| \leq \sigma_2(B) \cdot \|x\|.$$

So,

$$\|M\| \leq \frac{N \cdot \sigma_2(B)}{\Delta}.$$

□

3.3.1 Limitations

It is interesting to understand whether the approach above can give a better lower bound on sign rank. There are two parts to the argument: Forster's argument, and the upper bound on $\|S\|^*$. We can try to separately improve each of the two parts.

Any improvement over Forster's argument would be very interesting, but as mentioned there is no significant improvement over it even without the restriction induced by VC dimension, so we do not discuss it further.

To improve the second part, we would like to find examples with the biggest spectral gap possible. The Alon-Boppana theorem [54] optimally describes limitations on spectral gaps. The second eigenvalue σ of a Δ regular graph is not too small,

$$\sigma \geq 2\sqrt{\Delta - 1} - o(1),$$

where the $o(1)$ term vanishes when N tends to infinity (a similar statement holds when the diameter is large [54]). Specifically, the best lower bound on sign rank this approach can yield is roughly $\sqrt{\Delta}/2$, at least when $\Delta \leq N^{o(1)}$.

But what about general lower bounds on $\|S\|^*$? It is well known that any $N \times N$ sign matrix S satisfies $\|S\| \geq \sqrt{N}$. We prove a generalization of this statement.

Lemma 24. *Let S be an $N \times N$ sign matrix. For $i \in [N]$, let γ_i be the minimum between the number of 1's and the number of -1's in the i'th row. Let $\gamma = \gamma(S) = \max\{\gamma_i : i \in [N]\}$. Then,*

$$\|S\|^* \geq \frac{N - \gamma}{\sqrt{\gamma} + 1}.$$

This lemma provides limitations on the bound from Theorem 23. Indeed, $\gamma(S) \leq \frac{N}{2}$ and $\frac{N-\gamma}{\sqrt{\gamma}+1}$ is a monotone decreasing function of γ, which implies $\|S\|^* \geq \Omega(\sqrt{N})$. Interestingly, Lemma 24 and Theorem 23 provide a quantitively weaker but a more general statement than the Alon-Boppana theorem: If B is a Δ regular $N \times N$ boolean matrix with $\Delta \leq N/2$, then

$$\frac{N \cdot \sigma_2(B)}{\Delta} \geq \frac{N - \Delta}{\sqrt{\Delta} + 1} \;\; \Rightarrow \;\; \sigma_2(B) \geq \left(1 - \frac{\Delta}{N}\right)\left(\sqrt{\Delta} - 1\right).$$

This bound is off by roughly a factor of two when the diameter of the graph is large. When the diameter is small, like in the case of the projective plane which we discuss in more detail below, this bound is actually almost tight: The second largest singular value of the boolean point-line incidence matrix of a projective plane of order n is $\sqrt{n}$ while this matrix is $n + 1$ regular (c.f., e.g., [3]).

It is perhaps worth noting that in fact here there is a simple argument that gives a slightly stronger result for boolean regular matrices. The sum of squares of the singular values of B is the trace of $B^t B$, which is $N\Delta$. As the spectral norm is Δ, the sum of squares of the other singular values is $N\Delta - \Delta^2 = \Delta(N-\Delta)$, implying that

$$\sigma_2(B) \geq \sqrt{\frac{\Delta(N-\Delta)}{N-1}},$$

which is (slightly) larger than the bound above.

Proof of Lemma 24. Let M be a matrix so that $\|M\| = \|S\|^*$ and $M_{i,j}S_{i,j} \geq 1$ for all i,j. Assume without loss of generality[13] that γ_i is the number of -1's in the i'th row of S. If $\gamma = 0$, then S has only positive entries which implies $\|M\| \geq N$ as claimed. So, we may assume $\gamma \geq 1$. Let t be the largest real so that

$$t^2 = \frac{(N-\gamma-t)^2}{\gamma}. \tag{1}$$

That is, if $\gamma = 1$ then $t = \frac{N-\gamma}{2}$ and if $\gamma > 1$ then

$$t = \frac{-(N-\gamma) + \sqrt{(N-\gamma)^2 + (\gamma-1)(N-\gamma)^2}}{\gamma - 1}.$$

In both cases,

$$t = \frac{N-\gamma}{\sqrt{\gamma}+1}.$$

We shall prove that

$$\|M\| \geq t.$$

There are two cases to consider. One is that for all $i \in [N]$ we have $\sum_j M_{i,j} \geq t$. In this case, if x is the all 1 vector then

$$\|M\| \geq \frac{\|Mx\|}{\|x\|} \geq t.$$

The second case is that there is $i \in [N]$ so that $\sum_j M_{i,j} < t$. Assume without loss of generality that $i = 1$. Denote by C the subset of the columns j so that $M_{1,j} < 0$. Thus,

$$\begin{aligned}\sum_{j \in C} |M_{1,j}| &> \sum_{j \notin C} M_{1,j} - t \\ &\geq |[N] \setminus C| - t && (|M_{i,j}| \geq 1 \text{ for all } i,j) \\ &\geq N - \gamma - t. && (|C| \leq \gamma)\end{aligned}$$

Convexity of $x \mapsto x^2$ implies that

$$\left(\sum_{j \in C} |M_{1,j}|\right)^2 \leq |C| \sum_{j \in C} M_{1,j}^2,$$

[13] Multiplying a row by -1 does not affect $\|S\|^*$.

so by (1)

$$\sum_j M_{1,j}^2 \geq \frac{(N-\gamma-t)^2}{\gamma} = t^2.$$

In this case, if x is the vector with 1 in the first entry and 0 in all other entries then

$$\|(M)^T x\| = \sqrt{\sum_j M_{1,j}^2} \geq t = t\|x\|.$$

Since $\|(M)^T\| = \|M\|$, it follows that $\|M\| \geq t$. $\square$

3.3.2 Forster's theorem

Here we provide a proof of Forster's theorem, that is based on the following key lemma, which he proved.

Lemma 25 ([31]). *Let $X \subset \mathbb{R}^k$ be a finite set in general position, i.e., every k vectors in it are linearly independent. Then, there exists an invertible matrix B so that*

$$\sum_{x \in X} \frac{1}{\|Bx\|^2} Bx \otimes Bx = \frac{|X|}{k} I,$$

where I is the identity matrix, and $Bx \otimes Bx$ is the rank one matrix with (i,j) entry $(Bx)_i (Bx)_j$.

The lemma shows that every X in general position can be linearly mapped to BX that is, in some sense, equidistributed. In a nutshell, the proof of the lemma is by finding $B_1, B_2, \ldots$ so that each B_i makes $B_{i-1}X$ closer to being equidistributed, and finally using that the underlying object is compact, so that this process reaches its goal.

Proof of Theorem 22. Let M be a matrix so that $\|M\| = \|S\|^*$ and $M_{i,j} S_{i,j} \geq 1$ for all i, j. Clearly, sign-rank(S) = sign-rank(M). Let X, Y be two subsets of size N of unit vectors in $\mathbb{R}^k$ with k = sign-rank(M) so that $\langle x, y \rangle M_{x,y} > 0$ for all x, y. Lemma 25 says that we can assume

$$\sum_{x \in X} x \otimes x = \frac{N}{k} I; \tag{2}$$

If necessary replace X by BX and Y by $(B^T)^{-1} Y$, and then normalize (the assumption required in the lemma that X is in general position may be obtained by a slight perturbation of its vectors).

The proof continues by bounding $D = \sum_{x \in X, y \in Y} M_{x,y} \langle x, y \rangle$ in two different ways.
First, bound D from above: Observe that for every two vectors u, v, Cauchy-Schwartz inequality implies

$$\langle Mu, v \rangle \leq \|Mu\| \|v\| \leq \|M\| \|u\| \|v\|. \tag{3}$$

Thus,

$$
\begin{aligned}
D &= \sum_{i=1}^{k}\sum_{x\in X}\sum_{y\in Y} M_{x,y}x_i y_i \\
&\le \sum_{i=1}^{k} \|M\| \sqrt{\sum_{x\in X} x_i^2}\sqrt{\sum_{y\in Y} y_i^2} && \text{((3))} \\
&\le \|M\| \sqrt{\sum_{i=1}^{k}\sum_{x\in X} x_i^2}\sqrt{\sum_{i=1}^{k}\sum_{y\in Y} y_i^2} = \|M\|N. && \text{(Cauchy-Schwartz)}
\end{aligned}
$$

Second, bound D from below: Since $|M_{x,y}| \ge 1$ and $|\langle x, y\rangle| \le 1$ for all x, y, using (2),

$$
D = \sum_{x\in X}\sum_{y\in Y} M_{x,y}\langle x, y\rangle \ge \sum_{x\in X}\sum_{y\in Y} (\langle x, y\rangle)^2 = \sum_{y\in Y}\sum_{x\in X}\langle y, (x\otimes x)y\rangle = \frac{N}{k}\sum_{y\in Y}\langle y, y\rangle = \frac{N^2}{k}.
$$

□

3.4 Applications

3.4.1 Explicit examples

Here we prove Theorem 7 and Theorem 8.

Proof of Theorem 7. It is well known that the VC dimension of A is d, but we provide a brief explanation. The VC dimension is at least d by considering any set of d independent points (i.e. so that no strict subset of it spans it). The VC dimension is at most d since every set of $d + 1$ points is dependent in a d dimensional space.

The lower bound on the sign rank follows immediately from Theorem 6, and the following known bound on the spectral gap of these matrices.

Lemma 26. *If B is the boolean version of A then*

$$
\frac{\sigma_2(B)}{\Delta} = \frac{n^{\frac{d-1}{2}}(n-1)}{n^d - 1} \le N_{n,d}^{-\frac{1}{2}+\frac{1}{2d}}.
$$

The proof is so short that we include it here.

Proof. We use the following two known properties (see, e.g., [17]) of projective spaces. Both the number of distinct hyperplanes through a point and the number of distinct points on a hyperplane are $N_{n,d-1}$. The number of hyperplanes through two distinct points is $N_{n,d-2}$.

The first property implies that A is $\Delta = N_{n,d-1}$ regular. These properties also imply

$$
BB^T = (N_{n,d-1} - N_{n,d-2})\, I + N_{n,d-2}J = n^{d-1}I + N_{n,d-2}J,
$$

where J is the all 1 matrix. Therefore, all singular values except the maximum one are $n^{\frac{d-1}{2}}$. □

□

Proof of Theorem 8. We first show that R is indeed a maximum class of VC dimension 2. The VC dimension of R is 2: It is at least 2 because R contains the set of lines whose VC dimension is 2. It is at most 2 because no three points p_1, p_2, p_3 are shattered. Indeed if they all belong to a line ℓ then without loss of generality according to the order of ℓ we have $p_1 < p_2 < p_3$ which implies that the pattern 101 is missing. Otherwise, they are not co-linear and the pattern 111 is missing.

To see that R is a maximum class, note that there are exactly $N + 1$ intervals of size at most one (one empty interval and N singletons). For each line $\ell \in L$, the number of intervals of size at least two which are subsets of ℓ is exactly $\binom{|\ell|}{2} = \binom{n+1}{2}$. Since every two distinct lines intersect in exactly one point, it follows that each interval of size at least two is a subset of exactly one line. It follows that the number of intervals is

$$1 + N + N \cdot \binom{n+1}{2} = 1 + N + \binom{N}{2}.$$

Thus, R is indeed a maximum class of VC dimension 2.

Next we show that there exists a choice of a linear order for each line such that the resulting R has sign rank $\Omega(N^{\frac{1}{2}}/\log N)$. By the proof of Theorem 4, case $d = 2$, there is a choice of a subset for each line such that the resulting N subsets form a class of sign rank $\Omega(N^{\frac{1}{2}}/\log N)$. We can therefore pick the linear orders in such a way that each of these N subsets forms an interval, and the resulting maximum class (of all possible intervals with respect to these orders) has sign rank at least as large as $\Omega(N^{\frac{1}{2}}/\log N)$. □

3.4.2 Computing the sign rank

In this section we describe an efficient algorithm that approximates the sign rank (Theorem 9).

The algorithm uses the following notion. Let V be a set. A pair $\{v, u\} \subseteq V$ is *crossed* by a vector $c \in \{\pm 1\}^V$ if $c(v) \neq c(u)$. We also say that the vector c is crossed by the pair $\{u, v\}$. Let T be a tree with vertex set $V = [N]$ and edge set E. Let S be a $V \times [N]$ sign matrix. The *stabbing number* of T in S is the largest number of edges in T that are crossed by the same column of S. For example, if T is a path then T defines a linear order (permutation) on V and the stabbing number is the largest number of sign changes among all columns with respect to this order.

Welzl [69] gave an efficient algorithm for computing a path T with a low stabbing number for matrices S with VC dimension d. The analysis of the algorithm can be improved by a logarithmic factor using a result of [37].

Theorem 27 ([69, 37])**.** *There exists a polynomial time algorithm such that given a $V \times [N]$ sign matrix S with $|V| = N$, outputs a path on V with stabbing number at most $200N^{1-1/d}$ where $d = VC(S)$.*

For completeness, and since to the best of our knowledge no explicit proof of this theorem appears in print, we provide a description and analysis of the algorithm. We assume without loss of generality that the rows of S are pairwise distinct.

We start by handling the case[14] $d = 1$. In this case, we directly output a tree that is a path (i.e., a linear order on V). If $d = 1$, then Claim 16 implies that there is a column with at most 2 sign changes with respect to any order on V. The algorithm first finds by recursion a path T for the matrix obtained from S by removing this column, and outputs the same path T for the matrix S as well. By induction, the resulting path has stabbing number at most 2 (when there is a single column the stabbing number can be made 1).

For $d > 1$, the algorithm constructs a sequence of N forests $F_0, F_1, \ldots, F_{N-1}$ over the same vertex set V. The forest F_i has exactly i edges, and is defined by greedily adding an edge e_i to F_{i-1}. As we prove below, the tree F_{N-1} has a stabbing number at most $100N^{1-1/d}$. The tree F_{N-1} is transformed to a path T as follows. Let $v_1, v_2, \ldots, v_{2N-1}$ be an eulerian path in the graph obtained by doubling every edge in F_{N-1}. This path traverses each edge of F_{N-1} exactly twice. Let S' be the matrix with $2N - 1$ rows and N columns obtained from S be putting row v_i in S as row i, for $i \in [2N - 1]$. The number of sign changes in each column in S' is at most $2 \cdot 100N^{1-1/d}$. Finally, let T be the path obtained from the eulerian path by leaving a single copy of each row of S. Since deleting rows from S' cannot increase the number of sign changes, the path T is as stated.

The edge e_i is chosen as follows. The algorithm maintains a probability distribution p_i on $[N]$. The weight $w_i(e)$ of the pair $e = \{v, u\}$ is the probability mass of the columns e crosses, that is, $w_i(e) = p_i(\{j \in [N] : S_{u,j} \neq S_{v,j}\})$. The algorithm chooses e_i as an edge with minimum w_i-weight among all edges that are not in F_{i-1} and do not close a cycle in F_{i-1}.

The distributions $p_1, \ldots, p_N$ are chosen iteratively as follows. The first distribution p_1 is the uniform distribution on $[N]$. The distribution p_{i+1} is obtained from p_i by doubling the relative mass of each column that is crossed by e_i. That is, let $x_i = w_i(e_i)$, and for every column j that is crossed by e_i define $p_{i+1}(j) = \frac{2p_i(j)}{1+x_i}$, and for every other column j define $p_{i+1}(j) = \frac{p_i(j)}{1+x_i}$.

This algorithm clearly produces a tree on V, and the running time is indeed polynomial in N. It remains to prove correctness. We claim that each column is crossed by at most $O(N^{1-1/d})$ edges in T. To see this, let j be a column in S, and let k be the number of edges crossing j. It follows that

$$p_N(j) = \frac{1}{N} \cdot 2^k \cdot \frac{1}{(1 + x_1)(1 + x_2) \ldots (1 + x_{N-1})}.$$

To upper bound k, we use the following claim.

Claim 28. *For every i we have $x_i \leq 4e^2(N - i)^{-1/d}$.*

The claim completes the proof of Theorem 27: Since $p_N(j) \leq 1$ and $d > 1$,

$$\begin{aligned} k &\leq \log N + \log(1 + x_1) + \ldots + \log(1 + x_{N-1}) \\ &\leq \log(N) + 2(\ln(1 + x_1) + ... + \ln(1 + x_{N-1})) && (\forall x : \ \log(x) \leq 2\ln(x)) \\ &\leq \log(N) + 2(x_1 + ... + x_{N-1}) \\ &\leq \log N + 8e^2 N^{1-1/d} \leq 100N^{1-1/d}. \end{aligned}$$

The claim follows from the following theorem of Haussler.

[14] This analysis also provides an alternative proof for Lemma 17.

Theorem 29 ([37]). *Let p be a probability distribution on $[N]$, and let $\epsilon > 0$. Let $S \in \{\pm 1\}^{V \times [N]}$ be a sign matrix of VC dimension d so that the p-distance between every two distinct rows u, v is large:*

$$p(\{j \in [N] : S_{v,j} \neq S_{u,j}\}) \geq \epsilon.$$

Then, the number of distinct rows in S is at most

$$e(d+1)\left(2e/\epsilon\right)^d \leq \left(4e^2/\epsilon\right)^d.$$

Proof of Claim 28. Haussler's theorem states that if the number of distinct rows is M, then there must be two distinct rows of p_i-distance at most $4e^2 M^{-1/d}$. There are $N-i$ connected components in F_i. Pick $N-i$ rows, one from each component. Therefore, there are two of these rows whose distance is at most $4e^2 M^{-1/d} = 4e^2(N-i)^{-1/d}$. Now, observe that the w_i-weight of the pair $\{u, v\}$ equals the p_i-distance between u, v. Since e_i is chosen to have minimum weight, $x_i \leq 4e^2(N-i)^{-1/d}$ □

We now describe the approximation algorithm. Let S be an $N \times N$ sign matrix of VC dimension d. Run Welzl's algorithm on S, and get a permutation of the rows of S that yield a low stabbing number. Let s be the maximum number of sign changes among all columns of S with respect to this permutation. Output $s+1$ as the approximation to the sign rank of S.

We now analyze the approximation ratio. By Lemma 18 the sign rank of S is at most $s+1$. Therefore, the approximation factor $\frac{s+1}{\text{sign-rank}(S)}$ is at least 1. On the other hand, Proposition 1 implies that $d \leq$ sign-rank(S). Thus, by the guarantee of Welzl's algorithm,

$$\frac{s+1}{\text{sign-rank}(S)} \leq O\left(\frac{N^{1-1/d}}{\text{sign-rank}(S)}\right) \leq O\left(\frac{N^{1-1/d}}{d}\right).$$

This factor is maximized for $d = \Theta(\log N)$ and is therefore at most $O(N/\log N)$.

3.4.3 Counting VC classes

Here we prove Theorems 11 and 12. It is convenient for both to set

$$f = \sum_{i=0}^{d} \binom{N}{i}.$$

Proof of Theorem 11. We start with the upper bound. Enumerate the members of each such class C as follows. Start with the (lexicographically) first member $c \in C$, call it c_1. Assuming $c_1, c_2, \ldots, c_i$ have already been chosen, let c_{i+1} be the member c among the remaining vectors in C whose hamming distance from the set $\{c_1, \ldots, c_i\}$ is minimum (in case of equalities we take the first one lexicographically). This gives an enumeration $c_1, \ldots, c_m$ of the members of C, and $m \leq f$.

We now upper bound the number of possible families. There are at most 2^N ways to choose c_1. If the distance of c_{i+1} from the previous sets is $h = h_{i+1}$, then we can determine c_{i+1} by giving the index $j \leq i$ so that the distance between c_{i+1} and c_j is h, and by giving the symmetric difference of c_{i+1} and c_j. There are less than $m \leq f$ ways to choose the index, and at most

$\binom{N}{h} < (eN/h)^h$ options for the symmetric difference. The crucial point is that by Theorem 29 the number of i for which $h_i \geq D$ is less than $e(d+1)(2eN/D)^d$. Hence the number of i for which h_i is between 2^ℓ and $2^{\ell+1}$ is at most $e(d+1)(2eN/2^\ell)^d$. This upper bounds $c(N,d)$ by at most

$$2^N m^f \prod_\ell \left((eN/2^{\ell+1})^{2^{\ell+1}}\right)^{e(d+1)(2eN/2^\ell)^d} \leq 2^N f^f N^{(O(N))^d} = N^{(O(N))^d}.$$

We now present a lower bound on the number of (maximum) classes with VC dimension d. Take a family F of $\binom{N}{d}/(d+1)$ subsets of $[N]$ of size $(d+1)$ so that every subset of size d is contained in exactly one of them. Such families exist by a recent breakthrough result of Keevash [41], provided the trivial divisibility conditions hold (i.e. for all $0 \leq k < d$ that $\binom{N-k}{d-k}/(d+1-k)$ is an integer) and $N > N_0(d)$. His proof also gives that there are $N^{(1+o(1))\binom{N}{d}/(d+1)}$ such families.

Now, construct a class C by taking all subsets of cardinality at most $d-1$, and for each $(d+1)$-subset in the family F take it and all its subsets of cardinality d besides one. The VC dimension of C is indeed d. The number of possible Cs that can be constructed this way is at least the number of families F. Therefore, the number of classes of VC dimension d is at least the number of Fs:

$$N^{(1+o(1))\binom{N}{d}/(d+1)} = N^{\Omega(N^d/d^{d+1})}.$$

Note we can actually use in a similar manner, any family of $d+1$ subsets such that no d subset is contained in two of them. One can show using the Rödl-Nibble (see e.g. [1]) that there are many such families, and hence we do not really require neither the divisibility condition nor Keevash's result.

□

Proof of Theorem 12. For the upper bound we use the known fact that every maximum class is a connected subgraph of the boolean cube [36]. Thus, to upper bound the number of maximum classes of VC dimension d it is enough to upper bound the number of connected subgraphs of the N-dimensional cube of size f. It is known (see, e.g., Lemma 2.1 in [4]) that the number of connected subgraphs of size k in a graph with m vertices and maximum degree D is at most $m(eD)^k$. In our case, plugging $k = f$, $m = 2^N$, $D = N$ yields the desired bound $2^N(eN)^f = N^{(1+o(1))f}$.

For the lower bound, note that in the proof of Theorem 11 the constructed classes were of size f, and therefore maximum classes. Therefore, there are at least $N^{(1+o(1))\binom{N}{d}/(d+1)}$ maximum classes of VC dimension d. □

3.4.4 Counting graphs

Proof of Theorem 13. The key observation is that whenever we split the vertices of a $U(d+1)$-free graph into two disjoint sets of equal size, the bipartite graph between them defines a matrix of VC dimension at most d. Therefore, by Lemma 20, there is a reordering of the rows of the matrix so that the number of sign changes in every column is at most $O(N^{1-1/d})$. It follows that after such a reordering the number of possible columns is at most

$$\binom{N}{O(N^{1-1/d})} = 2^{O(\frac{1}{d}N^{1-1/d}\log N)}$$

Hence, the number of such bipartite graphs is at most

$$T(N,d) = 2^{O(\frac{1}{d}N^{2-1/d}\log N)}.$$

By a known lemma of Shearer [25], this implies that the total number of $U(d+1)$-free graphs on N vertices is less than $T(N,d)^2 = 2^{O(\frac{1}{d}N^{2-1/d}\log N)}$. For completeness, we include the simple details. The lemma we use is the following.

Lemma 30 ([25])**.** *Let $\mathcal{F}$ be a family of vectors in $S_1 \times S_2 \cdots \times S_n$. Let $\mathcal{G} = \{G_1, \ldots, G_m\}$ be a collection of subsets of $[n]$, and suppose that each element $i \in [n]$ belongs to at least k members of $\mathcal{G}$. For each $1 \le i \le m$, let $\mathcal{F}_i$ be the set of all projections of the members of $\mathcal{F}$ on the coordinates in G_i. Then*

$$|\mathcal{F}|^k \le \prod_{i=1}^{m} |\mathcal{F}_i|.$$

In our application, $n = \binom{N}{2}$ and $S_1 = \ldots = S_n = \{0,1\}$. The vectors represent graphs on N vertices, each vector being the characteristic vector of a graph on N labeled vertices. The set $[n]$ corresponds to the set of all $\binom{N}{2}$ potential edges. The family $\mathcal{F}$ represents all $U(d+1)$-free graphs. The collection $\mathcal{G}$ is the set of all complete bipartite graphs with $N/2$ vertices in each color class. Each edge $i \in [n]$ belongs to at least (in fact a bit more than) half of them, i.e., $k \ge m/2$. Hence,

$$|\mathcal{F}| \le \left(\prod_{i=1}^{m} |\mathcal{F}_i|\right)^{2/m} \le \left((T(N,d))^m\right)^{2/m},$$

as desired. □

4 Concluding remarks and open problems

We have given explicit examples of $N \times N$ sign matrices with small VC dimension and large sign rank. However, we have not been able to prove that any of them has sign rank exceeding $N^{1/2}$. Indeed this seems to be the limit of Forster's approach, even if we do not bound the VC dimension. Forster's theorem shows that the sign rank of any $N \times N$ Hadamard matrix is at least $N^{1/2}$. It is easy to see that there are Hadamard matrices of sign rank significantly smaller than linear in N. Indeed, the sign rank of the 4×4 signed identity matrix is 3, and hence the sign rank of its k'th tensor power, which is an $N \times N$ Hadamard matrix with $N = 4^k$, is at most $3^k = N^{\log 3/\log 4}$ (a similar argument was given by [34] for the Sylvester-Hadamard matrix). It may well be, however, that some Hadamard matrices have sign rank linear in N, as do random sign matrices, and it will be very interesting to show that this is the case for some such matrices. It will also be interesting to decide what is the correct behavior of the sign rank of the incidence graph of the points and lines of a projective plane with N points. We have seen that it is at least $\Omega(N^{1/4})$ and at most $O(N^{1/2})$.

Using our spectral technique we can give many additional explicit examples of matrices with high sign rank, including ones for which the matrices not only have VC dimension 2, but are more restricted than that (for example, no 3 columns have more than 6 distinct projections).

We have shown that the maximum sign rank $f(N, d)$ of an $N \times N$ matrix with VC dimension $d > 1$ is at most $O(N^{1-1/d})$, and that this is tight up to a logarithmic factor for $d = 2$, and close to being tight for large d. It seems plausible to conjecture that $f(N, d) = \tilde{\Theta}(N^{1-1/d})$ for all $d > 1$.

We have also showed how to use this upper bound to get a nontrivial approximation algorithm for the sign rank. It will be interesting to fully understand the computational complexity of computing the sign rank.

Finally we note that most of the analysis in this paper can be extended to deal with $M \times N$ matrices, where M and N are not necessarily equal, and we restricted the attention here for square matrices mainly in order to simplify the presentation.

Acknowledgements

We wish to thank Rom Pinchasi, Amir Shpilka, and Avi Wigderson for helpful discussions and comments.

References

[1] N. Alon and J.H. Spencer. *The Probabilistic Method.* Wiley Series in Discrete Mathematics and Optimization. Wiley, 2015.

[2] Noga Alon. Eigenvalues and expanders. *Combinatorica*, 6(2):83–96, 1986.

[3] Noga Alon. Eigenvalues, geometric expanders, sorting in rounds, and Ramsey theory. *Combinatorica*, 6(3):207–219, 1986.

[4] Noga Alon. A parallel algorithmic version of the local lemma. *Random Struct. Algorithms*, 2(4):367–378, 1991.

[5] Noga Alon, József Balogh, Béla Bollobás, and Robert Morris. The structure of almost all graphs in a hereditary property. *J. Comb. Theory, Ser. B*, 101(2):85–110, 2011.

[6] Noga Alon, Peter Frankl, and Vojtech Rödl. Geometrical realization of set systems and probabilistic communication complexity. In *26th Annual Symposium on Foundations of Computer Science, Portland, Oregon, USA, 21-23 October 1985*, pages 277–280, 1985.

[7] Noga Alon, David Haussler, and Emo Welzl. Partitioning and geometric embedding of range spaces of finite Vapnik-Chervonenkis dimension. In *Proceedings of the Third Annual Symposium on Computational Geometry, Waterloo, Ontario, Canada, June 8-10, 1987*, pages 331–340, 1987.

[8] Noga Alon and V. D. Milman. Eigenvalues, expanders and superconcentrators (extended abstract). In *25th Annual Symposium on Foundations of Computer Science, West Palm Beach, Florida, USA, 24-26 October 1984*, pages 320–322, 1984.

[9] Noga Alon and V. D. Milman. λ_1, isoperimetric inequalities for graphs, and superconcentrators. *J. Comb. Theory, Ser. B*, 38(1):73–88, 1985.

[10] Noga Alon, Lajos Rónyai, and Tibor Szabó. Norm-graphs: Variations and applications. *J. Comb. Theory, Ser. B*, 76(2):280–290, 1999.

[11] Richard P. Anstee, Lajos Rónyai, and Attila Sali. Shattering news. *Graphs and Combinatorics*, 18(1):59–73, 2002.

[12] Rosa I. Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 63(2):161–182, 2006.

[13] Hans-Jürgen Bandelt, Victor Chepoi, Andreas W. M. Dress, and Jack H. Koolen. Combinatorics of lopsided sets. *Eur. J. Comb.*, 27(5):669–689, 2006.

[14] Ronen Basri, Pedro F. Felzenszwalb, Ross B. Girshick, David W. Jacobs, and Caroline J. Klivans. Visibility constraints on features of 3d objects. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 1231–1238, 2009.

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

[16] Shai Ben-David and Michael Lindenbaum. Localization vs. identification of semi-algebraic sets. *Machine Learning*, 32(3):207–224, 1998.

[17] Albrecht Beutelspacher and Ute Rosenbaum. *Projective geometry - from foundations to applications*. Cambridge University Press, 1998.

[18] Amey Bhangale and Swastik Kopparty. The complexity of computing the minimum rank of a sign pattern matrix. *CoRR*, abs/1503.04486, 2015.

[19] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Classifying learnable geometric concepts with the Vapnik-Chervonenkis dimension (extended abstract). In Juris Hartmanis, editor, *Proceedings of the 18th Annual ACM Symposium on Theory of Computing, May 28-30, 1986, Berkeley, California, USA*, pages 273–282. ACM, 1986.

[20] Béla Bollobás and A. J. Radcliffe. Defect Sauer results. *J. Comb. Theory, Ser. A*, 72(2):189–208, 1995.

[21] Bernhard E. Boser, Isabelle Guyon, and Vladimir Vapnik. A training algorithm for optimal margin classifiers. In David Haussler, editor, *Proceedings of the Fifth Annual ACM Conference on Computational Learning Theory, COLT 1992, Pittsburgh, PA, USA, July 27-29, 1992.*, pages 144–152. ACM, 1992.

[22] William G. Brown. On graphs that do not contain a Thomsen graph. *Canadian Mathematical Bulletin*, 9:281–85, 1966.

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

[24] Bernard Chazelle and Emo Welzl. Quasi-optimal range searching in space of finite VC-dimension. *Discrete & Computational Geometry*, 4:467–489, 1989.

[25] Fan R. K. Chung, Ronald L. Graham, Peter Frankl, and James B. Shearer. Some intersection theorems for ordered sets and graphs. *J. Comb. Theory, Ser. A*, 43(1):23–37, 1986.

[26] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[27] J. Dodziuk. Difference equations, isoperimetric inequality and transience of certain random walks. *Trans. Am. Math. Soc.*, 284:787–794, 1984.

[28] Thorsten Doliwa, Gaojian Fan, Hans Ulrich Simon, and Sandra Zilles. Recursive teaching dimension, VC-dimension and sample compression. *Journal of Machine Learning Research*, 15(1):3107–3131, 2014.

[29] Thorsten Doliwa, Hans-Ulrich Simon, and Sandra Zilles. Recursive teaching dimension, learning complexity, and maximum classes. In Marcus Hutter, Frank Stephan, Vladimir Vovk, and Thomas Zeugmann, editors, *Algorithmic Learning Theory, 21st International Conference, ALT 2010, Canberra, Australia, October 6-8, 2010. Proceedings*, volume 6331 of *Lecture Notes in Computer Science*, pages 209–223. Springer, 2010.

[30] Sally Floyd and Manfred K. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

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

[32] Jürgen Forster, Matthias Krause, Satyanarayana V. Lokam, Rustam Mubarakzjanov, Niels Schmitt, and Hans-Ulrich Simon. Relations between communication complexity, linear arrangements, and computational complexity. In Ramesh Hariharan, Madhavan Mukund, and V. Vinay, editors, *FST TCS 2001: Foundations of Software Technology and Theoretical Computer Science, 21st Conference, Bangalore, India, December 13-15, 2001, Proceedings*, volume 2245 of *Lecture Notes in Computer Science*, pages 171–182. Springer, 2001.

[33] Jürgen Forster, Niels Schmitt, Hans-Ulrich Simon, and Thorsten Suttorp. Estimating the optimal margins of embeddings in Euclidean half spaces. *Machine Learning*, 51(3):263–281, 2003.

[34] Jürgen Forster and Hans-Ulrich Simon. On the smallest possible dimension and the largest possible margin of linear arrangements representing given concept classes. *Theor. Comput. Sci.*, 350(1):40–48, 2006.

[35] Peter Frankl. Traces of antichains. *Graphs and Combinatorics*, 5(1):295–299, 1989.

[36] Bernd Gärtner and Emo Welzl. Vapnik-Chervonenkis dimension and (pseudo-)hyperplane arrangements. *Discrete & Computational Geometry*, 12:399–432, 1994.

[37] David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded Vapnik-Chervonenkis dimension. *J. Comb. Theory, Ser. A*, 69(2):217–232, 1995.

[38] David Haussler and Emo Welzl. ϵ-nets and simplex range queries. *Discrete & Computational Geometry*, 2:127–151, 1987.

[39] Shlomo Hoory, Nathan Linial, and Avi Wigderson. Expander graphs and their applications. *Bull. Amer. Math. Soc.*, 43(04):439–562, August 2006.

[40] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mapping into Hilbert space. In *Conf. in modern analysis and probability*, volume 26 of *Contemporary Mathematics*, pages 189–206. American Mathematical Society, 1984.

[41] P. Keevash. The existence of designs. *CoRR*, abs/1401.366, 2014.

[42] János Komlós, János Pach, and Gerhard J. Woeginger. Almost tight bounds for epsilon-nets. *Discrete & Computational Geometry*, 7:163–173, 1992.

[43] Ilan Kremer, Noam Nisan, and Dana Ron. On randomized one-round communication complexity. *Computational Complexity*, 8(1):21–49, 1999.

[44] Eyal Kushilevitz and Noam Nisan. *Communication complexity*. Cambridge University Press, 1997.

[45] Dima Kuzmin and Manfred K. Warmuth. Unlabeled compression schemes for maximum classes. *Journal of Machine Learning Research*, 8:2047–2081, 2007.

[46] Troy Lee and Adi Shraibman. An approximation algorithm for approximation rank. In *Proceedings of the 24th Annual IEEE Conference on Computational Complexity, CCC 2009, Paris, France, 15-18 July 2009*, pages 351–357. IEEE Computer Society, 2009.

[47] Nathan Linial and Adi Shraibman. Learning complexity vs communication complexity. *Combinatorics, Probability & Computing*, 18(1-2):227–245, 2009.

[48] Satyanarayana V. Lokam. Complexity lower bounds using linear algebra. *Foundations and Trends in Theoretical Computer Science*, 4(1-2):1–155, 2009.

[49] Jirí Matousek. Intersection graphs of segments and $\exists\mathbb{R}$. *CoRR*, abs/1406.2636, 2014.

[50] Jirí Matousek, Emo Welzl, and Lorenz Wernisch. Discrepancy and approximations for bounded VC-dimension. *Combinatorica*, 13(4):455–466, 1993.

[51] N. E. Mnev. The universality theorems on the classification problem of configuration varieties and convex polytopes varieties. *Topology and Geometry*, 1346:527–544, 1989.

[52] Shay Moran. Shattering-extremal systems. *CoRR*, abs/1211.2980, 2012.

[53] Shay Moran and Manfred K. Warmuth. Labeled compression schemes for extremal classes. *CoRR*, abs/1506.00165, 2015.

[54] A. Nilli. On the second eigenvalue of a graph. *Discrete Mathematics*, 91(2):207 – 210, 1991.

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

[56] Alexander A. Razborov and Alexander A. Sherstov. The sign-rank of AC^0. *SIAM J. Comput.*, 39(5):1833–1855, 2010.

[57] J. Richter-Gebert. Mnëv's universality theorem revisited. *Sémin. Lothar. Comb. (electronic)*, B34h, 1995.

[58] Frank Rosenblatt. The perceptron–a perceiving and recognizing automaton. Technical Report 85-460-1, Cornell Aeronautical Laboratory, 1957.

[59] Benjamin I. P. Rubinstein and J. Hyam Rubinstein. A geometric approach to sample compression. *Journal of Machine Learning Research*, 13:1221–1261, 2012.

[60] J. Hyam Rubinstein, Benjamin I. P. Rubinstein, and Peter L. Bartlett. Bounding embeddings of VC classes into maximum classes. *CoRR*, abs/1401.7388, 2014.

[61] Norbert Sauer. On the density of families of sets. *J. Comb. Theory, Ser. A*, 13(1):145–147, 1972.

[62] Bernhard Schölkopf, Alexander J. Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998.

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

[64] Alexander A. Sherstov. Communication complexity under product and nonproduct distributions. *Computational Complexity*, 19(1):135–150, 2010.

[65] Peter W. Shor. Stretchabilltv of pseudolines is NP-hard. In Peter Gritzmann and Bernd Sturmfels, editors, *Applied Geometry And Discrete Mathematics, Proceedings of a DIMACS Workshop, Providence, Rhode Island, USA, September 18, 1990*, volume 4 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 531–554. DIMACS/AMS, 1990.

[66] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[67] Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998.

[68] H. E. Warren. Lower bounds for approximation by nonlinear manifolds. *Trans. Amer. Math. Soc.*, 133:167–178, 1968.

[69] Emo Welzl. Partition trees for triangle counting and other range searching problems. In Herbert Edelsbrunner, editor, *Proceedings of the Fourth Annual Symposium on Computational Geometry, Urbana-Champaign, IL, USA, June 6-8, 1988*, pages 23–33. ACM, 1988.