Journal of Machine Learning Research 3 (2002) 441-461 Submitted 11/01; Published 11/02

Limitations of Learning Via Embeddings in Euclidean Half Spaces

Shai Ben-David SHAI@CS.TECHNION.AC.IL
Department of Computer Science, Technion
Haifa 32000, Israel

Nadav Eiron NADAV@US.IBM.COM
IBM Almaden Research Center, 650 Harry Road
San Jose, CA 95120, USA

Hans Ulrich Simon SIMON@LMI.RUHR-UNI-BOCHUM.DE
Fakultät für Mathematik, Ruhr-Universität Bochum
D-44780 Bochum, Germany

Editor: Philip M. Long

Abstract

The notion of embedding a class of dichotomies in a class of linear half spaces is central to the support vector machines paradigm. We examine the question of determining the minimal Euclidean dimension and the maximal margin that can be obtained when the embedded class has a finite VC dimension.

We show that an overwhelming majority of the family of finite concept classes of any constant VC dimension cannot be embedded in low-dimensional half spaces. (In fact, we show that the Euclidean dimension must be almost as high as the size of the instance space.) We strengthen this result even further by showing that an overwhelming majority of the family of finite concept classes of any constant VC dimension cannot be embedded in half spaces (of arbitrarily high Euclidean dimension) with a large margin. (In fact, the margin cannot be substantially larger than the margin achieved by the trivial embedding.) Furthermore, these bounds are robust in the sense that allowing each image half space to err on a small fraction of the instances does not imply a significant weakening of these dimension and margin bounds.

Our results indicate that any universal learning machine, which transforms data into the Euclidean space and then applies linear (or large margin) classification, cannot enjoy any meaningful generalization guarantees that are based on either VC dimension or margins considerations. This failure of generalization bounds applies even to classes for which "straight forward" empirical risk minimization does enjoy meaningful generalization guarantees.

Keywords: Concept Learning, Embeddings in Half Spaces, Large Margin Classification

1. Introduction

Half spaces, or hyper-planes, have been at the center of the computational learning theory research since the introduction of the Perceptron algorithm by Rosenblatt (1958, 1962) and Minsky and Papert (1988). This interest in half spaces has led to a multitude of results concerning the learnability of these classes. In an attempt to harness the results achieved for this concept class in more general cases, one may consider a (more or less) universal learning paradigm that works by embedding other concept classes in half spaces. E.g., Support Vector Machines (SVMs) are based on the idea of embedding complex concept classes in half spaces and then applying efficient half spaces learning algorithms.

However, there may be a cost to pay for learning via such embeddings. The best known sample-independent bounds on the generalization ability of a hypothesis generated by a learning algorithm depend on the VC dimension of the concept class from which hypotheses are drawn. For half spaces this equals the Euclidean dimension over which these half spaces are defined. The first question addressed by this research is:

> Given a concept class (that is, some domain set and a family of dichotomies over this set), what is the minimal dimension of half spaces into which it can be embedded?

SVM theory offers a partial remedy to this problem. The margins of a hypothesis half space w.r.t. a given training sample can be used to compute a bound on the generalization quality of the hypothesis. If classification occurs with large enough margins then good generalization can be guaranteed regardless of the Euclidean dimension of these half spaces (For example, see Vapnik 1998, Freund and Schapire 1999, Mason et al. 2000.) This leads us to the second question that we discuss:

> Given a concept class, can the domain points and the class of concepts be embedded in some class of half spaces (of arbitrarily high Euclidean dimension) in such a way that there will be some significant margin separating the images of the sample points from the half spaces that are the images of the concepts of the class?

In this work we obtain strong negative answers to both questions. We prove that for "most classes" of any fixed VC dimension no embedding can obtain either a dimension or margins that are significantly better than those obtained by the trivial embedding. For classes that exhibit this kind of behavior, the generalization that can be guaranteed by SVM's is too weak to be of any practical value. Such examples exist also for classes of small VC dimension, in which case learning by empirical risk minimization over the original class would yield much better generalization bounds.

Before we elaborate any further on our results, let us explain the basic framework that we work in. Consider an SVM specialist faced with some learning task. The first significant decision she makes is the choice of kernel (or embedding) to be used for mapping the original feature vectors of the training set into a Euclidean space (where later half space learning algorithms will be applied). Assuming no prior knowledge on the nature of the learning task, one can readily see that the best possible embeddings are those mapping each example into a separate Euclidean dimension. We call such an embedding a *trivial embedding.* It is easy to see that trivial embeddings cannot yield any useful generalization. The other extreme case is when the learner bases her choice of embedding on the full knowledge of the sample labels. In this case the redundant function that maps all positively labeled examples to one point and all the negatively labeled examples to another, achieves perfect loading of the data, alas, once again it yields no generalization. Practical reality is somewhere between these two extremes: the learner does assume some prior knowledge and uses it for the choice of embedding. It is not at all clear how to model this prior knowledge. In this work we consider the case where the prior knowledge available to the learner is encapsulated as a collection of possible dichotomies of the domain feature vectors. The learner assumes that the correct labeling of the examples is close to one of these dichotomies. This modeling is very common in COLT research, and the collection of dichotomies is known as the *concept class*. Given such a class, the learner wishes to find an embedding of the instance space into a Euclidean space, so that every dichotomy in the class will be realized by a half space over the images of the examples.

We assume that both the instance space and the concept class are finite, and denote their cardinalities by n and m respectively. For sake of simplicity, we mainly focus on the case $m = n$. The general case $m \geq n$ is briefly discussed in Section 6. Our main results are as follows:

In Section 3 we show that, as n and $m = n$ grow unboundedly, an overwhelming majority of the family of finite concept classes of any constant VC dimension d cannot be embedded in the class of r-dimensional half spaces, unless r (as a function in n) is asymptotically larger than $n^{1-1/d-1/2^d}$. Note that, for large values of d, this lower bound approaches the trivial upper bound n achieved by the trivial embedding.

In Section 4 we address the issue of the margins obtainable by embeddings. We show that, as n and $m = n$ grow unboundedly, an overwhelming majority of the family of finite concept classes of constant VC dimension d cannot be embedded in the class of half spaces (of arbitrarily high dimension) with margin γ, unless γ (as a function in n) is asymptotically smaller than $\sqrt{1/n^{1-1/d-1/2^d}}$. Note that, for large values of d, this upper bound on γ approaches the trivial lower bound $1/\sqrt{n}$ achieved by the trivial embedding.

Furthermore, we show that our impossibility results remain qualitatively the same if the notion of embedding is relaxed, so that for every concept in the original class there exists a half space that classifies *almost all* of the embedded points like the original concept (rather than demanding the existence of a half space that achieves perfect fit with the concept dichotomy).

For large values of d, the lower bounds proven in Section 5 are almost tight because (as mentioned above) they approach the trivial upper bound n. For small values of d, namely $d = 4$ or $d = 6$, we show (in a less trivial manner) in Section 5 that these lower bounds are tight up to a logarithmic factor.

Our results indicate 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. Although we address only two generalization bounds (namely, the VC dimension and margin bounds), we believe that the phenomena that we demonstrate applies to other generalization bounds as well. Our results may be interpreted as showing that for a typical (or “random” or “common”) concept class of small VC dimension, an embedding of the class into a class of linearly separable dichotomies, inevitably introduces a significant degree of over-fitting. While half spaces embeddings may be desirable from the computational point of view, since there are efficient algorithms (like SVM’s) for learning via such embeddings, there are cases in which they result in a loss of the sample complexity bounds that do exist for learning small VC dimension classes using empirical risk minimization.

To clarify the implications of our results, we would like to mention that these results do not, of course, render learning machines of this type (like SVMs) useless. In fact, if most of the *important* classes could be nicely embedded, who cares about the vast majority?[1] Instead, our results indicate that the design of a "universal" learning machine (based on embeddings in half spaces) is an overly ambitious goal if it is pursued without further restrictions.

Most of our results are based on counting arguments and therefore only show the *existence* of 'hard-to-embed' classes (and that, indeed, they are the common case). However, in Section 4.1 we discuss the (non-)embeddability of specific concept classes.

We believe that the design of analytic tools, that allow the study of embeddability of a given concept class, will deepen the understanding of the embeddability question further. (See Forster 2001, Forster et al. 2001 as first steps in this direction.)

2. Definitions

In this section, we formally define the central notions of this paper. We start with the general notion of a concept class and the general notion of an embedding of one concept class into another (Subsection 2.1). Then we pass to geometric notions like hyper-planes, half spaces, and margins (Subsection 2.2). Afterwards we are prepared to formulate the general problem of embedding an arbitrary finite concept class (represented by a Boolean matrix) of a fixed VC dimension in the class of half spaces (Subsection 2.3). At the end of this section, we present some notions related to a famous problem of Zarankiewicz (Subsection 2.4). These notions are needed for proof technical reasons.

2.1 Concept Classes and Embeddings

The set of all functions from $\mathcal{X}$ to $\{0,1\}$ is denoted by $2^{\mathcal{X}}$. A function from $\mathcal{X}$ to $\{0,1\}$ is also called a *concept* over *domain* $\mathcal{X}$; each $\mathcal{C} \subseteq 2^{\mathcal{X}}$ is called a *concept class* over domain $\mathcal{X}$. Whenever we find it convenient, we identify a concept $f : \mathcal{X} \to \{0,1\}$ with the set $\{x \in \mathcal{X} : f(x) = 1\}$ (and vice versa).

The central notion discussed in this paper is the notion of embedding of one concept class into another.

Definition 1 *A concept class $\mathcal{C} \subseteq 2^{\mathcal{X}}$ over a domain $\mathcal{X}$ is embeddable in another concept class $\mathcal{C}' \subseteq 2^{\mathcal{X}'}$ over a domain $\mathcal{X}'$ iff there exists a function $\psi : \mathcal{X} \mapsto \mathcal{X}'$ such that*

$$\forall f \in \mathcal{C}, \exists g \in \mathcal{C}', \forall x \in \mathcal{X} \; f(x) = g(\psi(x)) \ .$$

We also present some results on approximate embeddings. These are embeddings in which some of the points in every concept class may be mis-classified by the embedding. Formally:

Definition 2 *A concept class $\mathcal{C} \subseteq 2^{\mathcal{X}}$ over a domain $\mathcal{X}$ is η-approximately embeddable in another concept class $\mathcal{C}' \subseteq 2^{\mathcal{X}'}$ over a domain $\mathcal{X}'$ iff there exists a function $\psi : \mathcal{X} \mapsto \mathcal{X}'$*

[1] Indeed, classes often studied in Computational Learning Theory research, such as Monomials and Decision Lists, can be embedded in Euclidean space of dimension exponentially smaller than that required for the majority of classes.

such that

$$\forall f \in \mathcal{C}, \exists g \in \mathcal{C}' : |\{x \in \mathcal{X} : g(\psi(x)) \neq f(x)\}| \leq \eta|\mathcal{X}| \ .$$

2.2 Separating Hyper-planes, Half Spaces and Margins

Consider the r-dimensional Euclidean domain $\mathbb{R}^r$. The Euclidean norm is denoted by $\|\cdot\|$. $S^{r-1} := \{x \in \mathbb{R}^r : \|x\| = 1\}$ denotes the unit sphere in $\mathbb{R}^r$; $B^r := \{x \in \mathbb{R}^r : \|x\| \leq 1\}$ denotes the closed unit ball. The *Euclidean hyper-plane* induced by "weight vector" $w \in \mathbb{R}^r$ and "threshold" $t \in \mathbb{R}$ is the set

$$H(w,t) = \{x \in \mathbb{R}^r : w \cdot x = t\} \ .$$

$H(w,t)$ splits $\mathbb{R}^r$ into two half-spaces

$$H_+(w,t) = \{x \in \mathbb{R}^r : w \cdot x > t\} \text{ and } H_-(w,t) = \{x \in \mathbb{R}^r : w \cdot x < t\} \ .$$

A hyper-plane or half space with threshold $t = 0$ is called *homogeneous.* By definition, the concept class of *r-dimensional Euclidean half spaces* consists of all half spaces of the form $H_+(w,t)$ for some $w \in \mathbb{R}^r$ and some $t \in \mathbb{R}$. The subclass of *r-dimensional homogeneous Euclidean half spaces* consists of all homogeneous half spaces of the form $H_+(w,0)$ for some $w \in \mathbb{R}^r$. In the sequel, we assume without loss of generality that a weight vector w is normalized to be of Euclidean length 1.

Definition 3 *Consider two finite sets of points in the r dimensional unit ball B^r, say S_- and S_+. We say hyper-plane $H(w,t)$* separates S_- from S_+ with margin γ *if $S_+ \subseteq H_+(w,t)$, $S_- \subseteq H_-(w,t)$, and γ is the Euclidean distance between H and $S_+ \cup S_-$:*

$$\gamma = \min\{|w \cdot x - t| : x \in S_+ \cup S_-\} \ .$$

According to Definition 1, an embedding of a finite concept class $\mathcal{C} \subseteq 2^{\mathcal{X}}$ in r-dimensional Euclidean half spaces is obtained by mapping each $x \in X$ to a point $u_x \in \mathbb{R}^r$ and by finding for each concept $f \in \mathcal{C}$ a weight vector $w = w_f$ and a threshold $t = t_f$ such that $H(w,t)$ separates $\{u_x : x \in \mathcal{X} \text{ and } f(x) = 0\}$ from $\{u_x : x \in \mathcal{X} \text{ and } f(x) = 1\}$. We may assume without loss of generality that the points u_x belong to the unit ball B^r (contract $\mathbb{R}^r$ by a scaling factor if necessary).[2]

Definition 4 *We say that a finite concept class $\mathcal{C}$* can be embedded in r-dimensional (homogeneous) half spaces with margin γ *if there exists an embedding that maps each $x \in X$ to a point $u_x \in B^r$ such that, for each $f \in \mathcal{C}$, there exists a (homogeneous) hyper-plane that separates $\{u_x : x \in \mathcal{X} \text{ and } f(x) = 0\}$ from $\{u_x : x \in \mathcal{X} \text{ and } f(x) = 1\}$ with margin γ.*

The latter definition can be relaxed by allowing a smaller margin (including mis-classifications) for an η-fraction of all instances:

Definition 5 *We say that a finite concept class $\mathcal{C}$* can be η-approximately embedded in r-dimensional (homogeneous) half spaces with margin γ *if there exists an embedding that maps each $x \in X$ to a point $u_x \in B^r$ such that, for each $f \in \mathcal{C}$, there exists a subset $M \subseteq X$ of cardinality at most $\eta|\mathcal{X}|$ and a (homogeneous) hyper-plane that separates $\{u_x : x \in \mathcal{X} \setminus M \text{ and } f(x) = 0\}$ from $\{u_x : x \in \mathcal{X} \setminus M \text{ and } f(x) = 1\}$ with margin γ.*

[2] Without this assumption, the margins associated with the embedding can be made arbitrarily large by means of scaling.

2.3 The Embedding Problem for Matrices of Small VC Dimension

Recall that $\mathcal{S} \subseteq \mathcal{X}$ is called *shattered* by $\mathcal{C}$ if for each function $f \in 2^{\mathcal{S}}$ there exists a function $g \in \mathcal{C}$ such that $f(x) = g(x)$ for all $x \in \mathcal{S}$. The *VC dimension* of C is the size of the maximum subset of X that is shattered by $\mathcal{C}$ (or "infinity" if there exist arbitrarily large finite shattered sets).

Throughout the paper, we use Boolean matrices to represent finite concept classes. A class $\mathcal{C}$ is represented by a matrix F of size $m \times n$, where $|\mathcal{C}| = m$, $|\mathcal{X}| = n$, and $F_{i,j}$ is the value of the ith concept on the jth instance.

Definition 6 *Let $\mathcal{D}(m, n, d)$ denote the family of Boolean matrices with m rows and n columns that have VC dimension smaller than d.*

Definition 7 *Let $\mathcal{E}(m, n, r)$ denote the family of Boolean matrices with m rows and n columns that can be embedded in the class of r-dimensional Euclidean half spaces.*

As mentioned in the introduction, we study embeddings of concept classes (represented by matrices) with low VC dimension into the class of Euclidean half spaces. Our basic approach will be to derive a lower bound on the Euclidean dimension r from the inequality $|\mathcal{D}(m, n, d)| \leq |\mathcal{E}(m, n, r)|$ (or slight variations of this inequality).

2.4 Some Technical Notions

We conclude this section with two technical notions that shall be needed in our proofs in Section 3.

Definition 8 *Let M be an $m \times n$ size matrix over $\{0, 1\}$. We say that M contains a* 1-monochromatic rectangle of size $s \times t$ *if there are sets of indices $A \subseteq \{1, \dots, m\}$ and $B \subseteq \{1, \dots, n\}$, where $|A| = s$ and $|B| = t$, such that for all $i \in A$ and all $j \in B$, the i, jth entry $m_{i,j}$ of M is* 1.

Definition 9 *Let $\mathcal{Z}(m, n, s, t)$ denote the family of Boolean matrices with m rows and n columns that do not contain a 1-monochromatic rectangle of size $s \times t$.*
Let $z(m, n, s, t)$ denote the maximum number of 1-entries in any matrix in $\mathcal{Z}(m, n, s, t)$.

Note that we may interpret $z(m, n, s, t)$ as the maximum number of edges in a bipartite graph G, whose vertex classes have size m and n, respectively, subject to the condition that G does not contain a complete bipartite $s \times t$ subgraph.

The following observation relates these combinatorial notions to classes of small VC dimension.

Lemma 10 $\mathcal{Z}(m, n, 2^d, d) \subseteq \mathcal{D}(m, n, 2d)$. *In other words, if a concept class has VC dimension $2d$ (and above) then its matrix representation contains a 1-monochromatic rectangle of size $2^d \times d$.*

Proof Consider a matrix F that shatters a set $Y = \{y_1 \dots y_{2d}\}$ of size $|Y| = 2d$. By definition, it means that F contains concepts with every possible assignment in those $2d$ places, including 2^d concepts that assign 1 to $y_1 \dots y_d$ and any other combination to the rest of the y's. These concepts define a 1-monochromatic rectangle in F with 2^d rows and d columns. Therefore, any matrix that does not contain such a rectangle has VC dimension smaller than $2d$. ■

3. An Asymptotic Lower Bound on the Euclidean Dimension Needed by Embeddings in Half Spaces

This section presents our results concerning the minimal dimension required to embed general matrices (representing concept classes) of fixed VC dimension in half spaces. Our proofs use known results for a combinatorial problem known as "the problem of Zarankiewicz".

In order to provide lower bounds on the dimensions of half spaces needed for embeddings, we shall show that there are many matrices of any fixed VC dimension. We shall compare these bounds with known upper bounds on the number of classes that can be embedded in half spaces in any fixed dimension Euclidean space.

The problem of determining $z(m,n,s,t)$ was first suggested (for specific values of s,t), by Zarankiewicz (1951), and later became known as "the problem of Zarankiewicz". Bollobás (1978) provides the following bounds, which are valid for all $2 \leq s \leq m$, $2 \leq t \leq n$:

$$\begin{aligned} z(m,n,s,t) &< (s-1)^{1/t}(n-t+1)m^{1-1/t} + (t-1)m & (1) \\ z(m,n,s,t) &\geq l(m,n,s,t) := \left\lfloor \left(1 - \frac{1}{s!t!}\right) m^{1-\alpha} n^{1-\beta} \right\rfloor & (2) \end{aligned}$$

where

$$\alpha := \alpha(s,t) := \frac{s-1}{st-1} \text{ and } \beta := \beta(s,t) := \frac{t-1}{st-1} \ .$$

Since the class $\mathcal{Z}(m,n,s,t)$ of matrices (viewed as bipartite graphs) is closed under edge deletion, the following inequality obviously holds:

$$|\mathcal{Z}(m,n,s,t)| \geq 2^{z(m,n,s,t)} \geq 2^{l(m,n,s,t)} = 2^{\left\lfloor (1-\frac{1}{s!t!}) m^{1-\alpha} n^{1-\beta} \right\rfloor}$$

Assume that at least a fraction $0 < \lambda \leq 1$ of the matrices in $\mathcal{Z}(m,n,s,t)$ is embeddable in the class of r-dimensional half spaces. It follows that

$$|\mathcal{E}(m,n,r)| \geq \lambda 2^{z(m,n,s,t)} \geq \lambda 2^{l(m,n,s,t)} \ . \qquad (3)$$

If inequality (3) is violated for every $r < r_0$, we may conclude that less than a fraction λ of the matrices in $\mathcal{Z}(m,n,s,t)$ can be embedded in the class of half spaces unless we embed into at least r_0 Euclidean dimensions. We will use this basic counting argument several times in what follows.

On the other hand, there are known bounds on the number of matrices of size $m \times n$ that may be embedded in half spaces.

Theorem 11 *(Alon et al., 1985) For every n, m, r:*

$$|\mathcal{E}(m,n,r)| \leq \min_{h \leq mn} \left(8 \left\lceil \frac{mn}{h} \right\rceil \right)^{(n+m)r+h+m} \qquad (4)$$

The remainder of this section is devoted to a first application of the basic counting argument. For sake of simplicity, we restrict ourselves to the case $m = n$ and postpone the general case $m \geq n$ to Section 6.

Theorem 12 *Let $s, t \geq 2$ be arbitrary but fixed constants. Then, for all sufficiently large n, the following holds. Only a vanishing fraction $2^{-l(n,n,s,t)/2}$ of the matrices from the family $\mathcal{Z}(n, n, s, t)$ is embeddable in the class of r-dimensional Euclidean half spaces unless*

$$r = \Omega\left(\frac{n^{1-\alpha(s,t)-\beta(s,t)}}{\log n}\right) = \omega\left(n^{1-1/s-1/t}\right).$$

Proof Let γ be an arbitrary but fixed constant such that $\alpha + \beta < \gamma \leq 1$ (like $\gamma = 1/s + 1/t$ for instance), let $m = n$, $h = n^{2-\gamma}$, and $\lambda = 2^{-l(n,n,s,t)/2}$. From (3), we get

$$|\mathcal{E}(n, n, r)| \geq \lambda 2^{l(n,n,s,t)} = 2^{l(n,n,s,t)/2} \quad . \tag{5}$$

From (4), we get

$$|\mathcal{E}(m, n, r)| \leq \left(8\left\lceil\frac{n^2}{h}\right\rceil\right)^{2nr+h+n} = (8\lceil n^{\gamma}\rceil)^{2nr+n^{2-\gamma}+n} \quad . \tag{6}$$

According to the definition of $l(m, n, s, t)$ in (2):

$$l(n, n, s, t) = \left\lfloor\left(1 - \frac{1}{s!t!}\right)n^{2-\alpha-\beta}\right\rfloor \tag{7}$$

Combining (5), (6),(7) and taking logarithms, we get

$$n(2r + n^{1-\gamma} + 1)\log(8\lceil n^{\gamma}\rceil) \geq \frac{1}{2}\left\lfloor\left(1 - \frac{1}{s!t!}\right)n^{1-\alpha-\beta}n\right\rfloor \geq \frac{n}{2}\left\lfloor\left(1 - \frac{1}{s!t!}\right)n^{1-\alpha-\beta}\right\rfloor \quad .$$

Cancelation by n finally yields

$$(2r + n^{1-\gamma} + 1)\log(8\lceil n^{\gamma}\rceil) \geq \frac{1}{2}\left\lfloor\left(1 - \frac{1}{s!t!}\right)n^{1-\alpha-\beta}\right\rfloor \quad . \tag{8}$$

Now Theorem 12 follows immediately (using $0 < \alpha + \beta < \gamma \leq 1$) because $n^{1-\alpha-\beta}$ grows asymptotically faster than $n^{1-\gamma}\log n$. ■

With some additional effort, Theorem 12 can be generalized to approximate embeddings:

Corollary 13 *Let $s, t \geq 2$ be arbitrary but fixed constants, and let γ be an arbitrary constant such that $\alpha(s,t) + \beta(s,t) < \gamma \leq 1$. ($\gamma = 1/s + 1/t$ would be a possible choice.) Then, for all sufficiently large n, the following holds. Only a vanishing fraction $2^{-l(n,n,s,t)/2}$ of the matrices from the family $\mathcal{Z}(n, n, s, t)$ is $n^{-\gamma}$-approximately embeddable in the class of r-dimensional Euclidean half spaces unless*

$$r = \Omega\left(\frac{n^{1-\alpha(s,t)-\beta(s,t)}}{\log n}\right) = \omega\left(n^{1-1/s-1/t}\right).$$

Proof Suppose that K is an arrangement of n half spaces and n vectors in $\mathbb{R}^r$ that represents a matrix F of $\mathcal{Z}(n,n,s,t)$. Then K represents a matrix F' $n^{-\gamma}$-approximately if, for all $i = 1, \ldots, n$, the Hamming distance between the ith row in F and the ith row in F' is at most $n^{-\gamma}n = n^{1-\gamma}$. The number of the matrices that are $n^{-\gamma}$-approximately represented by K is therefore upper-bounded by

$$\left(n^{n^{1-\gamma}}\right)^n = n^{n^{2-\gamma}}.$$

In order to complete the proof, we may therefore perform similar calculations as before, except that we have to expand the upper bound in Theorem 11 by the additional factor $n^{n^{2-\gamma}}$. These calculations end up at inequality

$$(2r + n^{1-\gamma} + 1)\log\left(8\left\lceil n^{\gamma}\right\rceil\right) + n^{1-\gamma}\log(n) \geq \frac{1}{2}\left\lfloor\left(1 - \frac{1}{s!t!}\right)n^{1-\alpha-\beta}\right\rfloor$$

instead of (8). Thus Euclidean dimension r exhibits the same asymptotic growth as before. ■

Combined with Lemma 10, this implies:

Corollary 14 *Let $d \geq 2$ be arbitrary but fixed. Let γ be an arbitrary constant such that $\alpha(2^d, d) + \beta(2^d, d) < \gamma \leq 1$. ($\gamma = 1/d + 1/2^d$ would be a possible choice.) Then, for all sufficiently large n, the following holds. Only a vanishing fraction $2^{-l(n,n,2^d,d)/2}$ of the matrices from the family $\mathcal{D}(n,n,2d)$ is $n^{-\gamma}$-approximately embeddable in the class of r-dimensional Euclidean half spaces unless*

$$r = \Omega\left(\frac{n^{1-\alpha(2^d,d)-\beta(2^d,d)}}{\log n}\right) = \omega\left(n^{1-1/2^d-1/d}\right).$$

4. Upper Bounds on the Margin Attainable by Embeddings in Half Spaces

In this section we prove some upper bounds on the margin that an embedding of an arbitrary class in half spaces may yield. We are going to employ two different techniques: a bound based on a concrete parameter of the class, and a combinatorial counting argument over the family of classes.

4.1 A Concrete Bound as a Function of Online Mistake Bounds

We present a rather simple technique that yields non-trivial upper bounds on the margins that can be obtained for certain specific classes. The idea is to use the online learning complexity of the class.

Recall that the online (or Mistake Bound) learning task for a class C of functions from some domain X to $\{0,1\}$ is defined by a game between a ‘teacher’ and a ‘student’. The teacher picks some function $c \in C$. Now the game runs in steps. At each step i the teacher picks some $x_i \in X$ and presents it to the student. The student returns a label $l_i \in \{0,1\}$

and passes it to the teacher, who then tells the student the value $c(x_i)$ and picks x_{i+1}. The cost of such a run of the game is $|\{i : l_i \neq c(x_i)\}|$. The Mistake Bound complexity of a class C is the minimum over all students strategies of the maximum over all teacher strategies of the cost of the run that is produced by these playing strategies. We denote it by $\mathrm{MB}(C)$.

Now, how does it relate to embeddings and margins? Let us recall the following well known result:

Theorem 15 *(Novikoff, 1962) Let $S = ((x_1, b_1), \ldots, (x_s, b_s))$ be a sequence of $\{0, 1\}$ labeled points in the unit ball in $\mathbb{R}^n$. If there exists a hyper-plane that separates $\{x_i : b_i = 0\}$ from $\{x_i : b_i = 1\}$ with margin $\geq \gamma$, then the online Perceptron algorithm makes at most $4/\gamma^2$ many mistakes on S. If, in addition, the separating hyper-plane is homogeneous, then the upper bound improves to $1/\gamma^2$.*

Tying these notions together we readily get:

Theorem 16 *Pick any $\gamma > 0$. If a class C can be embedded in the class of (not necessarily homogeneous) half spaces (in any dimension) with margin γ, then $MB(C) \leq 4/\gamma^2$. If, in addition, the embedding uses only homogeneous half spaces, then $MB(C) \leq 1/\gamma^2$.*

Proof The trick is to apply Novikoff's theorem about the Perceptron algorithm. Let $\psi : X \mapsto \mathbb{R}^n$ be an embedding that achieves margins above γ for the class C. Now let the learner use the following strategy: upon receiving a point x_{i+1} run the perceptron algorithm on $(\psi(x_1), c(x_1)), \ldots (\psi(x_i), c(x_i))$ and let l_{i+1} be the label given to $\psi(x_{i+1})$ by the half space that the perceptron algorithm produces. By Novikoff's theorem, if there exists a half space that separates the images of the points that c labels 1 from the images of the points that c labels 0 with margin $\geq \gamma$, then the perceptron algorithm (and therefore, our student) will make at most $4/\gamma^2$ mistakes (and at most $1/\gamma^2$ mistakes if the half space happens to be homogeneous). ■

The above simple result can be readily applied to demonstrate that some of the simplest classes cannot be embedded in half spaces with good margins. For example, let I^n be the class of all initial segments of $(1, \ldots, n)$. Note that the VC dimension of I^n is 1 regardless of the value of n. Just the same, it is not hard to see (and proven by Maass and Turán 1992) that $\mathrm{MB}(I^n) = \lfloor \log(n) \rfloor$.

Corollary 17 *I^n cannot be embedded in the class of half spaces (in any dimension) with margin above $2/\sqrt{\log(n)}$. Furthermore, I^n cannot be embedded in the class of homogeneous half spaces (in any dimension) with margin above $1/\sqrt{\log(n)}$*

In spite of the simplicity of the above result, it has quite striking consequences for learning methods. Namely, for some of the most simple concept classes, while Empirical Risk Minimization suffices for learning them efficiently (due to their constant VC dimension and simple structure), once they are embedded in half spaces, the generalization bound that relies on margins will grow to infinity with the size of the instance space!

Note however that, as the mistake bound of a class C is always bounded from above by $\log(|C|)$ (due to the Halving algorithm), the above idea cannot be used for obtaining smaller upper bounds on the values of obtainable margins.[3]

In the following subsection, we turn to a counting technique, that yields stronger bounds, but shows only existence of classes (rather then providing bounds for concrete classes).

4.2 Strong Margin Bounds for the Majority of Classes

In this section we are going to translate the lower bounds of Section 3 on the dimension of embeddings, into bounds on obtainable margins. The translation is done via the random projections technique. We use the following result:

Lemma 18 *(Arriaga and Vempala, 1999) Let $u \in \mathbb{R}^r$ be arbitrary but fixed. Let $R = (R_{i,j})$ be a random $(k \times r)$-matrix such that the entries $R_{i,j}$ are i.i.d. according to the normal distribution $N(0,1)$. Consider the* random projection $u_R := \frac{1}{\sqrt{k}}(Ru) \in \mathbb{R}^k$*. Then the following holds for every constant $\gamma > 0$:*

$$\Pr_R\left[\left|\|u_R\|^2 - \|u\|^2\right| \geq \gamma\|u\|^2\right] \leq 2e^{-\gamma^2 k/8}.$$

Corollary 19 *Let $w, x \in \mathbb{R}^r$ be arbitrary but fixed. Let $R = (R_{i,j})$ be a random $(k \times r)$-matrix such that the entries $R_{i,j}$ are i.i.d. according to the normal distribution $N(0,1)$. Then the following holds for every constant $\gamma > 0$:*

$$\Pr_R\left[|w_R \cdot x_R - w \cdot x| \geq \frac{\gamma}{2}\left(\|w\|^2 + \|x\|^2\right)\right] \leq 4e^{-\gamma^2 k/8}.$$

Proof Consider the events

$$\begin{aligned} \left|\|w_R + x_R\|^2 - \|w + x\|^2\right| &< \gamma\|w + x\|^2 && (9)\\ \left|\|w_R - x_R\|^2 - \|w - x\|^2\right| &< \gamma\|w - x\|^2. && (10)\end{aligned}$$

According to Lemma 18 (applied to $u = w + x$ and $u = w - x$, respectively), the probability of a violation of (9) or (10) is upper-bounded by $4e^{-\gamma^2 k/8}$. It suffices therefore to derive $|w_R \cdot x_R - w \cdot x| < \frac{\gamma}{2}(\|w\|^2 + \|x\|^2)$ from (9) and (10). From

$$\|w + x\|^2 = \|w\|^2 + 2w \cdot x + \|x\|^2 \text{ and } \|w - x\|^2 = \|w\|^2 - 2w \cdot x + \|x\|^2, \qquad (11)$$

we conclude that

$$\|w + x\|^2 - \|w - x\|^2 = 4w \cdot x. \qquad (12)$$

Clearly, the analogous relation holds for the random projections:

$$\|w_R + x_R\|^2 - \|w_R - x_R\|^2 = 4w_R \cdot x_R. \qquad (13)$$

[3] Using a much more involved method, it has been shown recently by Forster et al. (2001) that the maximum possible margin for an embedding of I^n in the class of homogeneous half spaces is exactly $n\left(\sum_{l=1}^{n}\left(\sin\frac{\pi(2l-1)}{2n}\right)^{-1}\right)^{-1} = \frac{\pi}{2\ln n} + \theta\left(\frac{1}{(\ln n)^2}\right)$.

Applying (12), (13), the triangle inequality, (9), (10), and (11) (in this order), we accomplish the proof as follows:

$$\begin{aligned} |w_R \cdot x_R - w \cdot x| &= \frac{1}{4} \left| \|w_R + x_R\|^2 - \|w + x\|^2 + \|w - x\|^2 - \|w_R - x_R\|^2 \right| \\ &\leq \frac{1}{4} \left(\left| \|w_R + x_R\|^2 - \|w + x\|^2 \right| + \left| \|w_R - x_R\|^2 - \|w - x\|^2 \right| \right) \\ &< \frac{\gamma}{4} \left(\|w + x\|^2 + \|w - x\|^2 \right) \\ &= \frac{\gamma}{2} \left(\|w\|^2 + \|x\|^2 \right) . \end{aligned}$$

■

From Corollary 19, the following result is easily obtained:

Corollary 20 *Let $\mathcal{C}$ be a set of $m = |\mathcal{C}|$ homogeneous half spaces of dimension r, and let $\mathcal{X}$ be a set of $n = |\mathcal{X}|$ points in the unit ball B^r. Assume that the smallest distance between a point from $\mathcal{X}$ and the homogeneous hyper-plane H associated with a half space H_+ from $\mathcal{C}$ is at least γ. Let R be a random $(k \times r)$-matrix such that the entries $R_{i,j}$ are i.i.d. according to the normal distribution $N(0,1)$. Then the following holds:*

1. $\Pr_R \left[\exists\, w \in \mathcal{C}, \exists\, x \in \mathcal{X} \ :\ sgn(w \cdot x) \neq sgn(R^T w \cdot R^T x) \right] \leq 4mne^{-\gamma^2 k/8}$.
2. *If $\gamma > \sqrt{8 \ln(4mn)/k}$, then $\mathcal{C}$ can be embedded in the class of k-dimensional half spaces.*

Proof Note that $\gamma > \sqrt{8 \ln(4mn)/k}$ is equivalent to $4mne^{-\gamma^2 k/8} < 1$. The second statement is therefore an immediate consequence of the first statement. The first statement can be shown as follows. Let $H_+(w,0)$ such that $\|w\| = 1$ be a fixed homogeneous half space from $\mathcal{C}$ and x be a fixed point from $\mathcal{X}$. By assumption, the distance between x and $H(w,0)$ is at least γ: $|w \cdot x| \geq \gamma$. If a random projection changes the sign of $w \cdot x$, it must change the value of $w \cdot x$ by at least γ. According to Corollary 19, the probability for this to happen is bounded by $4e^{-\gamma^2 k/8}$. Since there are mn choices for $H(w,t)$ and x, the total probability for a change of at least one of the signs is bounded by $4mne^{-\gamma^2 k/8}$. ■

Note that this result is independent of the original dimension r, and depends only on the margin γ and the dimension into which we embed, k. From Corollaries 20 and 14, we immediately obtain the main result of this section:

Theorem 21 *Let $d \geq 2$ be arbitrary but fixed. Then, for all sufficiently large n, the following holds. Only a vanishing fraction $2^{-l(n,n,2^d,d)/2}$ of the matrices from the family $\mathcal{D}(n,n,2d)$ is $n^{-\gamma}$-approximately embeddable in the class of half spaces (of arbitrarily large dimension) with a margin γ unless*

$$\gamma = O\left(\sqrt{\frac{\ln(n) \log(n)}{n^{1-\alpha(2^d,d)-\beta(2^d,d)}}} \right) = o\left(\sqrt{\frac{1}{n^{1-1/2^d-1/d}}} \right) .$$

We briefly note (without proof) that one can derive the following result from Corollary 20 and the counting arguments given in the paper of Alon et al. (1985).

Theorem 22 *For all sufficiently large n, the following holds. Only a vanishing fraction of the Boolean matrices of size $n \times n$ is embeddable in the class of half spaces (of arbitrarily large dimension) with a margin γ unless*

$$\gamma = O\left(\sqrt{\frac{\ln(n)}{n}}\right).$$

5. Tight Bounds for Classes of Low VC Dimension

We turn now to the question of what positive results can be achieved to complement our negative results on the dimension required for embedding. Low dimension embeddings for specific "interesting" classes can usually constructed by ad-hoc techniques. For instance, the class of monomials over n boolean variables can be embedded in half spaces of dimension n by associating a Euclidean dimension with each of the boolean variables, and treating the "true" value as 1 and the "false" value as -1. The half space associated with a monomial will simply have a 1 corresponding to a positive literal appearing in the monomial, a -1 corresponding to a negative literal appearing in the monomial, and a 0 for a variable that does not appear in the monomial. In fact, this embedding is not only of low dimension (of the order of $\log(|\mathcal{X}|)$), but also achieves margin $\Theta(1/n)$. Similarly, boolean decision lists on n boolean variables can be embedded in n Euclidean dimensions. The sample point transformation is the same as described above for monomials. The concept transformation is similar, but uses, for each variable, a value that has an exponential dependency on the location of the variable in the decision list, instead of the constants 1 and -1.

Contrary to these ad-hoc embeddings for specific classes, the type of results we seek is for a more general family of classes:

> For some fixed d, all matrices (or classes) of size $n \times n$ and VC dimension $2d$ may be embedded in half spaces of dimension $r(d,n)$, for some function $r(d,n) = O(n^{1-1/2^d-1/d})$.

Obviously, such a result would be interesting primarily for low values of d, where the difference between $r(d,n)$ and n (the dimension required by the trivial embedding) is significant. While we cannot present a general positive result, we do show that, for specific values of s,t, there exist sub-families of $\mathcal{Z}(n,n,s,t)$ that can be embedded in half spaces of a dimension matching the corresponding lower bound. Although this result is weaker than can ideally be hoped for, it shows that there are non-trivial cases, where the smallest Euclidean dimension needed to embed a family of matrices can be determined quite accurately.

The main results in this section are as follows:

Theorem 23 *For all n, the class of matrices $\mathcal{Z}(n,n,2,2)$ contains a sub-family $\mathcal{F}_{2\times 2}(n)$ that can be embedded in half spaces of dimension $O(n^{1/2})$, but cannot be embedded in half spaces of dimension $o(n^{1/2}/\log(n))$.*

Theorem 24 *For all n, the family of matrices $\mathcal{Z}(n,n,3,3)$ contains a sub-family $\mathcal{F}_{3\times 3}(n)$ that can be embedded in half spaces of dimension $O(n^{2/3})$, but cannot be embedded in half spaces of dimension $o(n^{2/3}/\log(n))$.*

The proofs of these theorems are given in Section 5.1 and 5.2. As in Section 3, the lower bounds are obtained by the basic counting argument. The upper bounds are obtained by exploiting the relationship between communication complexity and embeddings from the paper of Paturi and Simon 1986). Section 5.1 presents the sub-families $\mathcal{F}_{2\times2}(n)$ and $\mathcal{F}_{3\times3}(n)$ and applies the basic counting argument to them. Section 5.2 presents the corresponding embeddings.

5.1 Lower Bounds for Classes of Low VC Dimension

We would like to demonstrate that the bound we achieve for the Zarankiewicz matrices can be matched by an actual embedding for matrices of this type. The reason such results can only be expected for specific (small) values of s and t is that, as commented in the book of Bollobás (1978), the general lower bound for the Zarankiewicz problem is far from being tight. We therefore consider specific values of s, t for which better lower bounds are known. Furthermore, for these cases, constructions of the graphs that demonstrate the lower bound on $z(m, n, s, t)$ are also known (unlike the general lower bound, whose proof is not constructive). We consider two such specific cases, and show that for these cases we can construct an embedding in dimension close to our lower bound (using the improved results for the Zarankiewicz problem available for these cases).

The first case we tackle concerns the class of graphs $\mathcal{Z}(n, n, 2, 2)$, namely, bipartite graphs with two vertex sets of equal cardinality that do not contain a quadrilateral. For this specific case, Bollobás (1978) shows the following construction:

Let q be a prime power, and let $\mathrm{PG}(2, q)$ be the projective plane over a field of order q. Let V_1 be the set of points in $\mathrm{PG}(2, q)$ and V_2 be the set of lines in $\mathrm{PG}(2, q)$. An edge (v_1, v_2) is included in the graph iff the point v_1 is incident to the line v_2. It is immediate to verify that this graph indeed does not contain quadrilaterals (as any two points can only be incident to a single line).

The number of points, as well as the number of lines, in the projective plane, assuming we take $q = p$, a prime, is:

$$n = \frac{p^3 - 1}{p - 1} = p^2 + p + 1.$$

It is well-known that each point is incident to exactly $p - 1$ lines. We conclude that, for each prime p and $n = p^2 + p + 1$, there exists a Boolean $(n \times n)$-matrix F_n with $(p - 1)n$ 1-entries that does not contain a 1-monochromatic rectangle of size 2×2. Note that the latter property is preserved by flipping 1-entries to 0. Denote by $\mathcal{F}_{2\times2}(n)$ the family of all matrices of size $n \times n$, where $n = p^2 + p + 1$, that are constructed from F_n by flipping some of the 1-entries into zeros. We conclude that $\mathcal{F}_{2\times2}(n) \subseteq \mathcal{Z}(n, n, 2, 2)$ and

$$z(n, n, 2, 2) \geq n(p - 1) \geq n^{3/2}(1 - o(1))$$

A straightforward application of our basic counting argument shows that $r = \Omega(n^{1/2}/\log n)$ Euclidean dimensions are needed to embed each matrix from $\mathcal{F}_{2\times2}(n)$ in the class of r-dimensional half spaces. In the next subsection, we show that $O(n^{1/2})$ Euclidean dimensions are enough.

Another specific construction that appears in the book of Bollobás (1978) is tailored to the case $s = t = 3$. Again the construction is via geometry spaces over finite fields. This time the affine geometry space $\mathrm{AG}(3,p)$ of dimension 3 over the field $\mathrm{GF}(p)$ is used. For the construction we choose an element q in $\mathrm{GF}(p)$ which is a quadratic residue if and only if -1 is not a quadratic residue. We then define $S(x)$, for a point $x \in \mathrm{AG}(3,p)$, to be the sphere consisting of points y that satisfy:

$$\sum_{i=1}^{3}(x_i - y_i)^2 = q. \tag{14}$$

We can now construct a bipartite graph, with $n = p^3$ vertices in each of the vertex sets V_1 and V_2. We connect the edge between vertices $x \in V_1$ and $y \in V_2$ iff $x \in S(y)$ (or, equivalently, $y \in S(x)$). The resulting matrix, say F'_n, contains no 1-monochromatic rectangle of size 3×3. The number of 1-entries in F'_n is $p^5 - p^4$. Let us denote the family of matrices of size $n \times n$ obtained from F'_n by flipping some of 1-entries into zeros by $\mathcal{F}_{3\times 3}(n)$. Again, we have, $\mathcal{F}_{3\times 3}(n) \subseteq \mathcal{Z}(n,n,3,3)$, and a straightforward application of our basic counting argument shows that $r = \Omega(n^{2/3}/\log n)$ Euclidean dimensions are needed to embed each matrix from $\mathcal{F}_{3\times 3}(n)$ in the class of r-dimensional half spaces. In the next subsection, we show that $O(n^{2/3})$ Euclidean dimensions are enough.

5.2 Constructing Embeddings Through Communication Protocols

To construct embeddings we use a well-known connection between probabilistic communication complexity and embedding in half spaces. We use the model of unbounded error, two sided, communication complexity (see the paper of Paturi and Simon 1986). In this model, two players P_0 and P_1 are trying to compute a Boolean function $f(x,y)$, where P_0 is given $x \in \{0,1\}^n$ as input and P_1 is given $y \in \{0,1\}^n$ as input. Each player has unlimited computational power, and may realize any distribution on the messages it transmits to the other player. A protocol is said to calculate a function $f(x,y)$, if for any possible input pair (x,y), with probability exceeding $1/2$ (over the randomness of the players), the protocol will output the value of $f(x,y)$. For a communication protocol $\mathcal{A}$, we denote $C(\mathcal{A})$ its communication complexity, defined as the maximum over all possible inputs of the number of bits exchanged between P_0 and P_1 during the run of the protocol. For a function f, we define its unbounded error communication complexity to be:

$$C_f := \min_{\mathcal{A}_f} C(\mathcal{A}_f)$$

where the minimum is taken over all protocols $\mathcal{A}_f$ that correctly compute f.

The function f to be computed in such a communication protocol may also be represented by a square Boolean matrix F, where the entry $F(x,y)$ contains the value of $f(x,y)$. Paturi and Simon prove the following result (which is cited here with terminology adapted to this paper):

Theorem 25 *(Paturi and Simon, 1986) Let F be the matrix of a Boolean function f. Let r be the smallest dimension in which there is an embedding of the class represented by F into hyper-planes. Then, the unbounded error probabilistic communication complexity C_f for the function f satisfies:*

$$\lceil \log(r) \rceil \le C_f \le \lceil \log(r) \rceil + 1$$

Therefore, each communication protocol in this model for a function f with matrix F implicitly represents an embedding of F in the class of half spaces of dimension exponential in the communication complexity of the protocol (and vice versa). Let us now present communication protocols for the functions whose matrices were introduced in the previous subsection.

Recall that $F_n \in \mathcal{F}_{2\times 2}(n)$ denotes the matrix from the family $\mathcal{F}_{2\times 2}(n)$, $n = p^2 + p + 1$, that indicates the incidences between points and lines in $\mathrm{PG}(2,p)$. Assume processor P_0 has as input a (binary encoding of the) point x in $\mathrm{PG}(2,p)$, while P_1 has as input a (binary encoding of the) line y in $\mathrm{PG}(2,p)$. Our protocol for the matrix F_n is based on the following observation:

F_n has a 1 in position (x,y) if and only if the point x is incident to the line y. If we represent a point (a 1-dimensional vector subspace of $(\mathrm{GF}(p))^3$) by a vector in that subspace, and a line (a 2-dimensional vector subspace of $(\mathrm{GF}(p))^3$) by a vector that is orthogonal to the subspace, we have that x is incident to y if and only if $x \cdot y = 0$ (where "$\cdot$" denotes the inner product of the vector space $(\mathrm{GF}(p))^3$).

We can therefore use the following probabilistic communication protocol for the matrix F_n:

Protocol 1

1. Processor P_0 normalizes its input: if $x_1 \neq 0$, let $\hat{x} = x/x_1$. Otherwise, let $\hat{x} = x$.
2. Processor P_0 sends the value of $\hat{x}_1$ (one bit).
3. Processor P_0 sends the value of $\hat{x}_2$.
4. Processor P_1 solves the equation $\sum_{i=1}^{3} \hat{x}_i y_i = 0$ for $\hat{x}_3$. Denote this solution by z.
5. The processors run the protocol, EQ, for testing the equality of z and $\hat{x}_3$ (see the paper of Paturi and Simon 1986) and output the same bit as the EQ-protocol.

Theorem 26 *Protocol 1 is a probabilistic communication protocol for the matrix F_n that uses $3 + \lceil \log(p) \rceil = \frac{1}{2}\log(n) + O(1)$ bits of communication.*

Proof The correctness of the protocol is immediate: in step 4, processor P_1 has the values for y_1, y_2, y_3, $\hat{x}_1$, and $\hat{x}_2$ and can therefore solve the linear equation. From the observation above, a 1 in the matrix F_n corresponds to a solution of this equation. The EQ-protocol is then used to check whether $\hat{x}$ indeed solves this equation.

As for communication complexity, communicating the value of $\hat{x}_1$ takes just 1 bit (since its value is either 0 or 1). Communicating the value of $\hat{x}_2$ takes $\lceil \log(p) \rceil$ bits, and the EQ-protocol of Paturi and Simon requires two additional bits. ■

Note that a slight modification of this protocol can be used in the case that some of the 1-entries in the matrix were changed to zeros:
In step 4 above, a check should be made to see if the entry represented by the solution to this equation is 0. If this is the case, we can immediately output zero, even without running the EQ-protocol.

It follows that each matrix of the family $\mathcal{F}_{2,2}(n)$ can be computed by a protocol that exchanges $\frac{1}{2}\log n + O(1)$ bits. According to Theorem 25, each matrix from the family $\mathcal{F}_{2,2}(n)$ can be embedded in half spaces of dimension $O(n^{1/2})$. Theorem 23 immediately follows.

Let us now move to matrices from the class $\mathcal{F}_{3\times 3}(n)$, $n = p^3$, described in Subsection 5.1. Recall that F'_n is the matrix from $\mathcal{F}_{3\times 3}(n)$ that has a 1-entry in position (x, y) iff x and y satisfy relation (14). Assume, once more, that processor P_0 has as input a point $x \in \mathrm{AG}(3, p)$ while processor P_1 has as input a point $y \in \mathrm{AG}(3, p)$.

Before we describe a protocol for this matrix, let us mention a protocol for a problem we call EQ2 (for Equality-2). In this problem, processor P_0 is given an l-bit number x and processor P_1 is given two different l-bit numbers[4] (z, z'). The function EQ2 is given by

$$\mathrm{EQ2}(x, z, z') = 1 \iff (x = z \vee x = z').$$

Note that we assumed $z \neq z'$.

Lemma 27 *There exists a probabilistic communication protocol for EQ2 that uses* 5 *bits of communication (regardless of* l*).*

Proof Paturi and Simon provided a two-dimensional half space embedding for the matrix induced by the equality function, EQ, that checks whether two given l-bit numbers x and z are equal. Clearly, this embedding can be converted into a three-dimensional homogeneous half space embedding of EQ or, alternatively, ¬EQ. In other words, we may represent x as $(\xi_1, \xi_2, \xi_3) \in \mathbb{R}^3$ and z as $(\zeta_1, \zeta_2, \zeta_3) \in \mathbb{R}^3$ such that

$$x = z \iff \sum_{i=1}^{3} \xi_i \zeta_i < 0 \text{ and } x \neq z \iff \sum_{i=1}^{3} \xi_i \zeta_i > 0.$$

Making use of $z \neq z'$, it follows that

$$(x = z \vee x = z') \iff \sum_{i=1}^{3} \sum_{j=1}^{3} \xi_i \xi_j \zeta_i \zeta'_j = \left(\sum_{i=1}^{3} \xi_i \zeta_i\right) \cdot \left(\sum_{j=1}^{3} \xi_j \zeta'_j\right) < 0. \tag{15}$$

Equation (15) shows that the matrix induced by the function EQ2 can be embedded in the class of 9-dimensional half spaces. According to Theorem 25, there must be a probabilistic communication protocol that uses at most 5 bits of communication. ∎

We refer to the protocol for function EQ2 as the EQ2-protocol in what follows. Now that we have the EQ2-protocol (exchanging at most 5 bits), we are ready to introduce our protocol for the matrix F'_n:

Protocol 2

1. Processor P_0 sends the values of x_1 and x_2 to processor P_1.

[4] While the model of Paturi and Simon requires inputs to both processors to be of the same length, we may assume that P_0 is given two l-bit numbers, and ignores one of them in its computation.

2. Given x_1, x_2, y_1, y_2, y_3, Processor P_1 solves equation (14) for x_3 and finds (at most) two solutions. If no solutions exist, output 0. Otherwise, denote the solutions by z and z'. Processor P_1 informs P_0 whether $z = z'$ or $z \neq z'$ (one bit).

3. If $z = z'$, the processors run the EQ-protocol such as to check whether $x_3 = z$. If $z \neq z'$, the processors run the EQ2-protocol such as to check whether $x_3 = z$ or $x_3 = z'$. They output the same bit as the EQ-protocol or the EQ2-protocol, respectively, does.

Theorem 28 *Protocol 2 is a probabilistic communication protocol for the matrix F'_n, and uses $6 + 2\lceil \log(p) \rceil = (2/3) \log n + O(1)$ bits of communication.*

Proof The communication complexity of Protocol 2 is immediate: Processor P_0 sends x_1 and x_2, which are both elements of $\mathrm{GF}(p)$ and therefore require $\lceil \log(p) \rceil$ bits each. Processor P_1 sends one bit in order to inform P_0 of whether $z = z'$ or not. Afterwards, the processors either run the EQ-protocol (at the expense of 2 bits) or the EQ2-protocol (at the expense of 5 bits). This sums up to at most $6 + 2\lceil \log(p) \rceil$ bits of communication.

The correctness of the protocol is also immediate. Equation (14), solved by P_1 in step 2 of the protocol, coincides with the equation that was used to define the 1-entries of the matrix F'_n. ■

Again, a slight modification of the protocol may be used for matrices in $\mathcal{F}_{3\times 3}(n)$ that had some of their 1-entries flipped to 0:
Either P_1 knows, after receiving x_1 and x_2, that the result is 0 (if all solutions to the equation of step 2 correspond to entries that have been flipped to 0), or it knows that one of the two possible solutions to this equation (say, w.l.o.g., z'), corresponds to an entry that was flipped to 0. In the latter case, the EQ-protocol can be used to check whether $x_3 = z$. It follows that each matrix of the family $\mathcal{F}_{3,3}(n)$ can be computed by a protocol that exchanges $(2/3) \log n + O(1)$ bits. According to Theorem 25, each matrix from the family $\mathcal{F}_{3,3}(n)$ can be embedded in half spaces of dimension $O(n^{2/3})$. Theorem 23 immediately follows.

6. Conclusions and Open Problems

This work addresses the issue of what success guarantees can be proved for SVM like learning, from the assumption that the classification of examples is close to a dichotomy in some concept class of small VC dimension. Roughly speaking, we showed that neither the VC dimension nor margins can provide a guarantee for the generalization ability of the learning paradigm that embeds the feature space into Euclidean spaces and uses half-space learning algorithms on the embedded images. In particular, our results apply to SVM as well as other kernel based learning algorithms.

The most natural question that this research raises is what is the reason that SVMs do work so well in practice. We see two directions in which the search for an answer should be pursued: The first is to to look for other parameters that may guarantee generalization of learning paradigms, and may not be subject to pessimistic results as displayed here. The most natural candidate for this is the notion of sparsity (Ben-David, 2001). However there may be some other useful parameters to be discovered. The other potential research direction, and in a sense broader question, is the issue of modeling the learner's prior knowledge. In this paper we have been following the common COLT setup in which this knowledge is formalized as a concept class. While this formulation is convenient for mathematical analysis, it is not at all clear that it reflects many natural learning scenarios. There is definitely a place for other formal notions aiming to formalize a learner's prior knowledge, or bias, that may reflect some practically common aspects that are not modeled well enough by the notion of a concept class.

There is also a technical issue that, in spite of being a very natural extension of the questions that this paper answers, is not completely covered by our results:

We proved that only a vanishing fraction of the Boolean $(n \times n)$-matrices (representing concept classes) of constant VC dimension can be embedded in half spaces with an Euclidean dimension or a margin that is substantially better than the dimension or the margin of the trivial embedding. A natural question to ask is to what extent do these results carry over to non-quadratic matrices.

Since the class of $(m \times n)$-matrices such that $m \geq n$ contains the class of $(n \times n)$-matrices, the *full* family $\mathcal{D}(m, n, 2d)$ can clearly not be embedded in a lower-dimensional space (or with a larger margin) than the family $\mathcal{D}(n, n, 2d)$. Thus, the bounds $r = \Omega\left(\frac{n^{1-\alpha(2^d,d)-\beta(2^d,d)}}{\log n}\right)$ and $\gamma = O\left(\sqrt{\frac{\ln(n)\log(n)}{n^{1-\alpha(2^d,d)-\beta(2^d,d)}}}\right)$ from Corollary 14 and Theorem 21, respectively, apply for any family of the form $\mathcal{D}(l, k, d)$ as long as $n \leq \min\{k, l\}$. The point that remains unanswered by this simple consideration is the relative fraction of the matrices in such a class that cannot be embedded in half spaces of smaller dimension. We conjecture that only a vanishing fraction of the matrices from $\mathcal{D}(m, n, 2d)$ can be embedded in a substantially lower-dimensional space (or with a substantially larger margin). We must admit, however, that our current proof technique (the detour on the problem of Zarankiewicz) only leads to sort of "weak" generalizations. For instance, if $m = n^k$ for some constant k, the lower bound on the Euclidean dimension r from Corollary 13, namely

$$r = \Omega\left(\frac{n^{1-\alpha(s,t)-\beta(s,t)}}{\log n}\right) \quad ,$$

becomes

$$r = \Omega\left(\frac{n^{1-k\cdot\alpha(s,t)-\beta(s,t)}}{\log n}\right) \quad .$$

Likewise, the bound

$$r = \Omega\left(\frac{n^{1-\alpha(2^d,d)-\beta(2^d,d)}}{\log n}\right)$$

from Corollary 14 becomes

$$r = \Omega\left(\frac{n^{1-k\cdot\alpha(2^d,d)-\beta(2^d,d)}}{\log n}\right) \quad ,$$

a bound being completely useless for $k \geq d$. This raises the question whether less trivial embeddings in half spaces become possible when the size of the concept class comes close to the maximal possible size (which, by Sauer's Lemma, is $\sum_{i=0}^{d} \binom{n}{i} = \theta(n^d)$). We conjecture however that our proof technique, though quite successful when applied to quadratic matrices, does not show the right picture for matrices of arbitrary shape. We leave the problem of inventing more powerful techniques for the general case as an object of future research.

Another technical issue that seems to not be completely solved is tightening the lower bound on the problem of Zarankiewicz. As evident from (1) and (2), the upper and lower bounds on $z(m, n, s, t)$ are not tight. However, for our purposes, the gap between them is not very significant. In fact, using (1) instead of (2) in the proofs of Section 3 causes the minimal dimension required for embedding, as specified in Corollary 10, to be changed from $\omega\left(n^{1-1/2^d-1/d}\right)$ to $\omega\left(n^{1-1/d}\right)$. Therefore, any advances in proving better bounds on the problem of Zarankiewicz can only affect very minor changes in the results presented in Section 3.

Acknowledgments

This work has been supported in part by the ESPRIT Working Group in Neural and Computational Learning II, NeuroCOLT2, No. 27150. The authors gratefully acknowledge the support of the German-Israeli Foundation for Scientific Research and Development Grant I-403-001.06/95. Some of the research described in this work was done while the second author was a graduate student at the Department of Computer Science, Technion, Haifa, Israel. The third author was supported by the Deutsche Forschungsgemeinschaft Grant SI 498/4-1.

References

N. Alon, P. Frankl, and V. Rödl. Geometrical realization of set systems and probabilistic communication complexity. In *Proceedings of the 26th Symposium on Foundations of Computer Science*, pages 277–280, 1985.

Rosa I. Arriaga and Santosh Vempala. An algorithmic theory of learning: Robust concepts and random projection. In *Proceedings of the 40'th Annual Symposium on the Foundations of Computer Science*, pages 616–623, 1999.

Shai Ben-David. A priori generalization bounds for kernel based learning: Can sparsity save the day? Technical Report, Cornell University, 2001.

Béla Bollobás. *Extremal Graph Theory.* Academic Press, 1978.

Jürgen Forster. A linear bound on the unbounded error probabilistic communication complexity. In *Proceedings of the 16th Annual Conference on Computational Complexity*, pages 100–106, 2001.

Jürgen Forster, Niels Schmitt, and Hans Ulrich Simon. Estimating the optimal margins of embeddings in euclidean half spaces. In *Proceedings of the 14th Annual Workshop on Computational Learning Theory*, pages 402–415. Springer Verlag, 2001.

Yoav Freund and Robert Schapire. Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296, 1999.

Wolfgang Maass and György Turán. Lower bound methods and separation results for on-line learning models. *Machine Learning*, 9:107–145, 1992.

Llew Mason, Peter L. Bartlett, and Jonathan Baxter. Improved generalization through explicit optimization of margins. *Machine Learning*, 38(3):243–255, 2000.

Marvin L. Minsky and Seymour A. Papert. *Perceptrons.* The MIT Press, Cambrigde MA, third edition, 1988.

A. B. J. Novikoff. On convergence proofs for perceptrons. In *Proceedings of the Symposium of Mathematical Theory of Automata*, pages 615–622, 1962.

Ramamohan Paturi and Janos Simon. Probabilistic communication complexity. *Journal of Computer and System Sciences*, 33(1):106–123, 1986.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psych. Rev.*, 65:386–407, 1958.

F. Rosenblatt. *Principles and Neurodynamics: Perceptrons and the Theory of Brain Mechanisms.* Spartan Books, Washington, D.C., 1962.

Vladimir Vapnik. *Statistical Learning Theory.* Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, 1998.

K. Zarankiewicz. Problem P 101. *Colloq. Math.*, 2:301, 1951.