

Embedding with a Lipschitz function

Shahar Mendelson*

Research School of Information Sciences and Engineering
The Australian National University
Canberra, ACT 0200, Australia
shahar.mendelson@anu.edu.au

Abstract

We investigate a new notion of embedding of subsets of $\{-1, 1\}^n$ in a given normed space, in a way which preserves the structure of the given set as a class of functions on $\{1, \dots, n\}$. This notion is an extension of the *margin* parameter often used in Nonparametric Statistics. Our main result is that even when considering “small” subsets of $\{-1, 1\}^n$, the vast majority of such sets do not embed in a better way than the entire cube in any normed space that satisfies a minor structural assumption.

Keywords: embedding with a Lipschitz function, margin, random sets, local theory of normed spaces.

1 Introduction

The question we investigate is how well a subset of the discrete cube $\{-1, 1\}^n$ embeds in a given normed space. Unlike the usual notions of embeddings, in which the metric structure of the given set is preserved, we are interested in embeddings which preserve the structure of the set when viewed as a class of functions on $\{1, \dots, n\}$. To be precise, let F be a class of N binary valued functions on $\{1, \dots, n\}$ (and thus may be considered as a subset of $\{-1, 1\}^n$ of cardinality N). For a normed space X , let B_X be its unit ball put B_{X^*} to be the unit ball of the dual space X^* .

*I would like to thank B. Klartag, A. Litvak, G. Luria, G. Schechtman, A. Shraibman and N. Tomczak-Jaegermann for many stimulating discussions. I would also like to thank M. Rudelson and R. Vershynin who allowed me to use a result from their yet unpublished manuscript. Finally, many thanks to the referees for their valuable comments. This research was supported in part by an Australian Research Council Discovery Grant.

Definition 1.1 A class $F = \{f_i\}_{i=1}^N \subset \{-1, 1\}^n$ embeds with a Lipschitz function with a constant L in a normed space X if there are sets $\{x_1, \dots, x_n\} \subset B_X$ and $\{x_1^*, \dots, x_N^*\} \subset B_{X^*}$ and a Lipschitz function $\phi : \mathbb{R} \rightarrow \mathbb{R}$ such that $\|\phi\|_{\text{lip}} < L$ and for every $1 \leq i \leq N$ and $1 \leq j \leq n$,

$$\phi(x_i^*(x_j)) = f_i(j).$$

Given a space X , a constant is called *trivial* if the class of all functions on $\{1, 2, \dots, n\}$ embeds in X with that constant. We shall refer to *the trivial constant* as the infimum of all trivial constants.

This definition generalizes a slightly different notion of embedding which originated in Statistical Learning Theory. There, one considers *embedding with margin*, in which a specific Lipschitz function is used.

Definition 1.2 A class $F = \{f_i\}_{i=1}^N \subset \{-1, 1\}^n$ embeds with margin γ in a normed space X if there are sets $\{x_1, \dots, x_n\} \subset B_X$ and $\{x_1^*, \dots, x_N^*\} \subset B_{X^*}$ such that

$$\begin{aligned} x_i^*(x_j) &> \gamma \quad \text{if } f_i(j) = 1, \\ x_i^*(x_j) &< -\gamma \quad \text{if } f_i(j) = -1. \end{aligned}$$

As above, a margin is *trivial* if the class of all functions on $\{1, 2, \dots, n\}$ embeds in X with that margin.

In other words, an embedding consists of mappings of $\{1, \dots, n\}$ into B_X and of F into B_{X^*} in a way which preserves the structure of F as a class of functions. For example, the closer the separation parameter γ (i.e. the margin) is to 1, the closer the values $x_i^*(x_j)$ are to the original values $f_i(j)$. The general case is a natural extension of the margin, though it has a less clear geometric interpretation. Indeed, if F embeds with margin γ in X then it embeds with a Lipschitz function with a constant $1/\gamma$, using the function

$$\phi_\gamma(t) = \begin{cases} 1 & \text{if } t \geq \gamma, \\ \frac{t}{\gamma} & \text{if } -\gamma < t < \gamma, \\ -1 & \text{if } t \leq -\gamma. \end{cases}$$

One of our aims in this article is to present a first step in a systematic study of this notion of embedding, and in particular, to explore the connections between the geometry of the class, the geometry of the space and the best constant with which a class embeds in the space.

The motivation for investigating this notion comes from Statistical Learning Theory, where margin based prediction bounds have been thoroughly studied (see, e.g. [6, 3, 8] and references therein). From the learning theoretical

perspective, the advantages in embedding a class with large margin are both algorithmic and theoretical. If a class embeds, say in a Hilbert space, it has a representation as a rather simple set, given by a collection of linear functionals - and this turns out to be very useful algorithmically [15, 4]. Moreover, the fact that an embedding is with large margin (that is, a margin which is significantly larger than the trivial one) gives important statistical information on the class. Indeed, the margin can be used to upper bound other, more direct complexity parameters of the class F , such as the expectation of the supremum of the Rademacher process indexed by F (see, e.g. [6] for more details), and those estimates are central in obtaining prediction bounds which are at the heart of Learning Theory.

It is possible to derive similar prediction bounds if F embeds with a small Lipschitz constant in any normed space which has a reasonable structure (for example, a space with a bounded type 2 constant). All in all, as far as Statistical Learning is concerned, there is no real theoretical reason to restrict the analysis of embedding based methods to margin in a Hilbert space.

Although embedding based methods have been extensively studied, what was left unresolved is the following basic question: which classes embed with a large margin (or more generally, embed with a Lipschitz function with a small constant) and in what spaces? Because of this stumbling block, applied statistical approaches such as *support vector machines* and *kernel methods* resorted to creating ad-hoc classes of thresholded linear functionals without being able to identify the subsets of $\{-1, 1\}^n$ those classes represent [18, 15, 4].

It was believed by parts of the learning community that embedding based methods were universal in the sense that any solvable prediction problem could be solved via an embedding; that there is a fixed space X such that any class of binary valued functions which is small enough to allow a solution of a prediction problem, embeds in X in a way which yields the appropriate prediction bound. To check the validity of such a belief, one has to compare the constant with which a class F embeds in X to a parameter that captures the difficulty of a prediction problem that uses F , namely, the *Vapnik-Chervonenkis* dimension.

Definition 1.3 For every $\sigma \subset \{1, \dots, n\}$, the coordinate projection of F onto σ is defined as $P_\sigma F = \{(f(i))_{i \in \sigma} \mid f \in F\}$. The *Vapnik-Chervonenkis (vc) dimension* of F is the largest cardinality of a set $\sigma \subset \{1, \dots, n\}$, such that $P_\sigma F = \{-1, 1\}^{|\sigma|}$.

This leads to the main question we address here.

Question 1 *Is the optimal embedding constant of F in X equivalent in some sense to the Vapnik-Chervonenkis dimension of F ? In other words, does a “small” vc dimension imply the existence of an embedding with a small constant in X ?*

Our main result is a negative answer to Question 1 for a large variety of spaces. We show that the ability to embed a class with a constant that is significantly better than the trivial one is a much stronger property than the class just having a “small” vc dimension.

Previously it was demonstrated in [1] that asymptotically, the vast majority of subsets of $\{-1, 1\}^n$ with n elements and of a fixed vc dimension do not embed in a Hilbert space with a margin significantly better than the trivial one. To be exact, the authors showed that for a fixed d , only a vanishing fraction (at most $\sim 2^{-cn}$) of the set of subsets of $\{-1, 1\}^n$ with n elements and vc dimension at most d embed in ℓ_2 with a margin better than $1/n^\alpha$, where $\alpha = 1/2 - 1/2d - 1/2^{d-1}$. A part of the proof shows that a random subset of $\{-1, 1\}^n$ with N elements does not embed in ℓ_2 with a margin better than $c\sqrt{(\log N)/n}$ for suitable absolute constant c , as long as $N/n^2 \rightarrow \infty$. A new approach, based on operator ideal theory, is used in [7] to prove that if $N \geq cn$, a random subset of $\{-1, 1\}^n$ with N elements only embeds in ℓ_2 with the trivial margin of c_1/\sqrt{n} .

The results presented in [1] were the first indication that the vc dimension and the margin are very different complexity measures, at least as far as embeddings into a Hilbert space are concerned. Unfortunately, the approach of [1] uses the particular structure of a Hilbert space and cannot be extended beyond this case, even for the margin.

Here, we show that a random subset of $\{-1, 1\}^n$ with $N \geq cn \log n$ elements does not embed with a Lipschitz function in a nontrivial manner in many spaces (including ℓ_p^n for $1 \leq p < \infty$ and spaces with a bounded type 2 constant). Since the vc dimension of such classes is at most $c \log n$, our result shows that the two notions of size - the embedding constant and the vc dimension are indeed different for those spaces.

The article is organized as follows. Section 2 is devoted to basic definitions and notation. In section 3 we present the geometric characterization of the optimal embedding constant in a general Banach space X . We show that it is determined by the ability to factor the identity operator between \mathbb{R}^n endowed with a norm which represents the class and ℓ_∞^n , through a dual space to an n -dimensional subspace of X .

In section 4 we study a weaker notion of embedding with a Lipschitz function, called *soft embedding*, in which for every $1 \leq i \leq N$ one has

control on $x_i^*(x_j)$ only on subset of $\{1, \dots, n\}$ of cardinality proportional to n (see Section 3 for the exact definition). We prove that a random class of cardinality $N \geq cn \log n$ only trivially embeds in the weaker sense of the soft embedding in a large family of Banach spaces. In fact, the proof we present holds for any space for which the Banach-Mazur distance between any proportional subspace and ℓ_1^{cn} is essentially the same as the distance of the space to ℓ_1^n . In particular, this property can be verified for ℓ_p^n and any space with a bounded type 2 constant.

Let us mention that the notion of embedding with a Lipschitz function is far from being fully understood. What seems to be an intriguing question which will not be addressed here is

Question 2 *Is there a useful geometric property of a class which governs the best constant with which it embeds in a given normed space?*

2 Preliminaries

In this section we present some notation and the preliminary background in Banach spaces theory we require. The interested reader could turn to [12, 16] as general references which cover the necessary background.

Throughout, all absolute constants are denoted by c or C . Their values may change from line to line or even within the same line. c_φ , C_φ and $C(\varphi)$ denote constants which depend only on the parameter φ , and $a \sim_\varphi b$ means that $c_\varphi b \leq a \leq C_\varphi b$. If the constants are absolute we use the notation $a \sim b$. For a set A , let $|A|$ be its cardinality and if $A, B \subset \mathbb{R}$, put $A+B = \{a+b | a \in A, b \in B\}$.

Given a real Banach space X , let B_X be its unit ball. The dual of X , denoted by X^* , consists of all the bounded linear functionals on X , endowed with the norm $\|x^*\| = \sup_{\|x\|=1} |x^*(x)|$. If X and Y are Banach spaces and $T : X \rightarrow Y$ is a linear operator, define the operator norm $\|T\| = \sup_{\|x\|_X=1} \|Tx\|_Y$.

A set K is called symmetric if the fact that $x \in K$ implies that $-x \in K$. The symmetric convex hull of K , denoted by $\text{absconv}(K)$, is the convex hull of $K \cup -K$. If $K \subset \mathbb{R}^n$ is bounded, convex and symmetric with a nonempty interior then K is a unit ball of the norm denoted by $\|\cdot\|_K$. From here on, we refer to such a set K as a *ball* and let $|K|$ be its (n -dimensional) volume (it will be clear from the context when this notation is used for the volume and when it is used to denote cardinality).

It is possible to show that the *polar* of K , defined by $K^\circ = \{x \in \mathbb{R}^n | \sup_{k \in K} \langle k, x \rangle \leq 1\}$, is the unit ball of the dual space of $(\mathbb{R}^n, \|\cdot\|_K)$.

For $1 \leq p < \infty$, let ℓ_p^n be \mathbb{R}^n endowed with the norm $\|\sum_{i=1}^n a_i e_i\|_p = (\sum_{i=1}^n |a_i|^p)^{1/p}$ and set ℓ_∞^n to be \mathbb{R}^n endowed with the norm $\|\sum_{i=1}^n a_i e_i\|_\infty = \sup_i |a_i|$; B_p^n is the unit ball of ℓ_p^n .

Recall that the *Banach-Mazur distance* between two isomorphic Banach spaces X and Y is defined as $d(X, Y) = \inf \|T\| \cdot \|T^{-1}\|$, where the infimum is taken with respect to all isomorphisms between X and Y . It is easy to see that if X, Y and Z are isomorphic, then $d(X, Z) \leq d(X, Y) \cdot d(Y, Z)$. A well known theorem due to F. John [12], states that for any n -dimensional normed space X , $d(X, \ell_2^n) \leq \sqrt{n}$.

We say that a Banach space X is *finitely represented* in a Banach space Y if for every $\varepsilon > 0$ and every finite dimensional subspace $X_0 \subset X$ there is a subspace $Y_0 \subset Y$ such that $d(X_0, Y_0) \leq 1 + \varepsilon$. In other words, one can find isomorphic copies of each finite dimensional subspace of X in Y , and the isomorphism is arbitrarily close to being an isometry.

An example, which is one of the consequences of Dvoretzky's Theorem is that ℓ_2 is finitely represented in any infinite dimensional Banach space [12].

A property of Banach spaces that will be used throughout this article is called *type*. Some basic facts concerning the concept of type may be found, for example, in [12].

Definition 2.1 *A Banach space X has type p if there is some constant C such that for every integer n and every $x_1, \dots, x_n \in X$,*

$$\mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\| \leq C \left(\sum_{i=1}^n \|x_i\|^p \right)^{1/p}, \quad (2.1)$$

where $(\varepsilon_i)_{i=1}^n$ are independent Rademacher random variables (that is, symmetric $\{-1, 1\}$ -valued). The smallest constant for which (2.1) holds is called the *type p constant* of X and is denoted by $T_p(X)$.

Clearly, (2.1) holds for any Banach space for $p = 1$, with $T_1(X) = 1$. Setting

$$p^* = \sup\{p : X \text{ has type } p\},$$

it is possible to show that $1 \leq p^* \leq 2$. If $p^* = 1$ then X is said to have only a trivial type. By considering $x_i = e_i$ – which are the standard unit vectors in \mathbb{R}^n , it is evident that $T_p(\ell_1^n) \geq n^{1-1/p}$, and this turns out to be the correct estimate of $T_p(\ell_1^n)$ [16].

An important fact known as the Maurey-Pisier Theorem [12], is that if X is infinite dimensional then ℓ_{p^*} is finitely represented in X . In particular, if X only has a trivial type, it contains “almost isometric” copies of ℓ_1^n for every n .

3 Embedding a class in a normed space

Given a class of functions F and a normed space X , let $L(F, X)$ be the infimum of the set of Lipschitz constants by which F embeds with a Lipschitz function in X . It is easy to see that for every space and any class with at least two elements, $L(F, X) \geq 1$.

Observe that F embeds with a constant L using the sets $\{x_1^*, \dots, x_N^*\}$ and $\{x_1, \dots, x_n\}$ if and only if there are sets $W_+, W_- \subset [-1, 1]$ such that $d(W_+, W_-) > 2/L$, and for every $1 \leq i \leq N$ and $1 \leq j \leq n$, $x_i^*(x_j) \in W_+$ if $f_i(j) = 1$ and $x_i^*(x_j) \in W_-$ if $f_i(j) = -1$. Hence, every Lipschitz function ϕ that can be used in an embedding is “coded” by such separated subsets of $[-1, 1]$.

A weaker notion of embedding is called *soft embedding*, in which one only needs to control the behavior of each x_i^* on a subset of $\{x_1, \dots, x_n\}$ of cardinality proportional to n .

Definition 3.1 *Given $L \geq 1$ and $1/2 < \delta \leq 1$, a class $F \subset \{-1, 1\}^n$ (L, δ)-softly embeds with a Lipschitz function in a space X if the following holds. For every $1 \leq i \leq N$ there is a set $A_i \subset \{1, \dots, n\}$ with $|A_i| \geq \delta n$, sets $\{x_1, \dots, x_n\} \subset B_X$ and $\{x_1^*, \dots, x_N^*\} \subset B_{X^*}$ and a function $\phi : \mathbb{R} \rightarrow \mathbb{R}$, such that $\|\phi\|_{\text{lip}} < L$ and*

$$\phi(x_i^*(x_j)) = f_i(j) \quad \text{if } j \in A_i.$$

For $1/2 < \delta \leq 1$, denote by $L(F, X, \delta)$ the infimum of the constants L with which F (L, δ)-softly embeds with a Lipschitz function in X .

A result we require is a representation theorem that identifies the constant with which F embeds in X .

Definition 3.2 *Let $F \subset \{-1, 1\}^n$ with $|F| = N$, and fix $L \geq 1$. A set $T = \{t_i : 1 \leq i \leq N\} \subset \mathbb{R}^n$ L -represents F , if there are sets $W_+, W_- \subset [-1, 1]$ with $d(W_+, W_-) > 2/L$ such that for every $1 \leq i \leq N$, $t_i = (t_{i,j})_{j=1}^n \in T$ satisfies that $t_{i,j} \in W_+$ if $f_i(j) = 1$ and $t_{i,j} \in W_-$ when $f_i(j) = -1$.*

For $1/2 < \delta \leq 1$, a set T (L, δ)-softly represents F , if for every $1 \leq i \leq N$ there is a set A_i as in Definition 3.1, and $t_{i,j}$ are as above for $j \in A_i$.

Theorem 3.3 *A class $F \subset \{-1, 1\}^n$ L -embeds with a Lipschitz function in X if and only if there is a set T which L -represents F and an n -dimensional subspace E of X , such that the operator $\text{Id} : (\mathbb{R}^n, \|\cdot\|_{\text{absconv}(T)}) \rightarrow \ell_\infty^n$ factors through E^* .*

In other words, there is a subspace $E \subset X$ of dimension n and operators $v : (\mathbb{R}^n, \|\cdot\|_{\text{absconv}(T)}) \rightarrow E^*$ and $u : E^* \rightarrow \ell_\infty^n$, such that $\|u\|, \|v\| \leq 1$ and $uv = Id$.

Proof. For the sake of simplicity, assume that $\dim(X) = n$; the general case follows an identical path to this one. Suppose that F embeds in X using $(x_j)_{j=1}^n \subset B_X$, $(x_i^*)_{i=1}^n \subset B_{X^*}$ and a Lipschitz function ϕ satisfying that $\|\phi\|_{\text{lip}} < L$. Applying a perturbation argument (in which ϕ might be slightly changed) it can be assumed that $\{x_1, \dots, x_n\}$ are linearly independent. Let $W_+ = [-1, 1] \cap \{\phi = 1\}$ and $W_- = [-1, 1] \cap \{\phi = -1\}$. For every $1 \leq i \leq n$ set $t_i = (x_i^*(x_j))_{j=1}^n$ and put $T = \{t_i : 1 \leq i \leq n\}$. Then $t_{i,j} \in W_+$ if $f_i(j) = 1$ and $t_{i,j} \in W_-$ if $f_i(j) = -1$.

Consider the operator $U : \mathbb{R}^n \rightarrow \mathbb{R}^n$ given by $Ue_i = x_i$. Since $(x_j)_{j=1}^n$ are linearly independent then U is invertible and each $x \in \mathbb{R}^n$ can be written as $x = \sum_{j=1}^n \alpha_j x_j$. Hence, for every $1 \leq i \leq n$,

$$x_i^*(x) = x_i^* \left(\sum_{j=1}^n \alpha_j x_j \right) = \sum_{j=1}^n \alpha_j t_{i,j} = \left\langle \sum_{j=1}^n \alpha_j e_j, \sum_{j=1}^n t_{i,j} e_j \right\rangle. \quad (3.1)$$

Since $\|x\| \leq 1$ if and only if $\sum_{j=1}^n \alpha_j Ue_j \in B_X$, or equivalently, if $\sum_{j=1}^n \alpha_j e_j \in U^{-1}B_X$, then by (3.1), the definition of the dual norm and the fact that $(U^{-1}B_X)^\circ = U^*B_{X^*}$,

$$\begin{aligned} \|x_i^*\|_{X^*} &= \sup_{\|x\|_X \leq 1} x_i^*(x) = \sup_{y \in U^{-1}B_X} \left\langle \sum_{j=1}^n t_{i,j} e_j, y \right\rangle \\ &= \left\| \sum_{j=1}^n t_{i,j} e_j \right\|_{(U^{-1}B_X)^\circ} = \left\| \sum_{j=1}^n t_{i,j} e_j \right\|_{U^*B_{X^*}}. \end{aligned}$$

Therefore, $\|x_i^*\|_{X^*} \leq 1$ if and only if $(U^*)^{-1} \left(\sum_{j=1}^n t_{i,j} e_j \right) \in B_{X^*}$. Setting $v = (U^*)^{-1}$ when considered as $v : (\mathbb{R}^n, \|\cdot\|_{\text{absconv}(T)}) \rightarrow X^*$, then $\|v\| \leq 1$. To complete the proof, one needs to show that $\|U^*\|_{X^* \rightarrow \ell_\infty^n} \leq 1$ (and to set $u = U^*$). By duality, it suffices to show that $\|U\|_{\ell_1^n \rightarrow X} \leq 1$, i.e., that $Ue_i \in B_X$ for every i , which is the case by the definition of U .

The reverse direction follows a similar argument and its proof is omitted. ■

In a similar manner, one can formulate and prove a “soft” version of Theorem 3.3.

Theorem 3.4 A class $F \subset \{-1, 1\}^n$ (L, δ) -softly embeds with a Lipschitz function in X if and only if there is an n -dimensional subspace $E \subset X$ and some set T which (L, δ) -softly represents F , such that $Id : (\mathbb{R}^n, \|\cdot\|_{\text{absconv}(T)}) \rightarrow \ell_\infty^n$ factors through E^* .

Remark 3.5 Observe that the proofs of Theorem 3.3 and of Theorem 3.4 (which is not presented here) show that if x_1, \dots, x_n can be used in the embedding, the identity operator can be factored using $u = U^*$ and $v = (U^*)^{-1}$, where $Ue_i = x_i$.

3.1 Embedding with margin

Probably the most interesting example in the context of Statistical Learning of a Lipschitz function that can be used to embed a class is the margin.

Definition 3.6 Given $0 < \gamma, \delta \leq 1$, the class $F \subset \{-1, 1\}^n$ (γ, δ) -softly embeds with margin in X if the following holds. For every $1 \leq i \leq N$ there is a set $A_i \subset \{1, \dots, n\}$ of cardinality $|A_i| \geq \delta n$, and there are sets $\{x_1, \dots, x_n\} \subset B_X$ and $\{x_1^*, \dots, x_N^*\} \subset B_{X^*}$ such that

$$\begin{aligned} x_i^*(x_j) &> \gamma \text{ if } f_i(j) = 1 \text{ and } j \in A_i, \\ x_i^*(x_j) &< -\gamma \text{ if } f_i(j) = -1 \text{ and } j \in A_i. \end{aligned}$$

Selecting $\delta = 1$ in this definition, one controls all the values $x_i^*(x_j)$ and the usual notion of embedding with margin is recovered.

Given a class F and a normed space X , denote by $m(F, X)$ the maximal margin with which F embeds in X . If $X = \ell_2$, the maximal margin is denoted by $m(F)$.

It is clear that if a class embeds with margin γ then it Lipschitz embeds with a constant $L \leq 1/\gamma$. Hence, for every normed space X and every class F , $L(F, X) \leq 1/m(F, X)$.

Another observation is that the margin has the following geometric interpretation. If F embeds with margin γ in X and if we consider the subsets $B_i \subset \{1, \dots, n\}$ defined by $B_i = \{j : f_i(j) = 1\}$, then for every $1 \leq i \leq N$,

$$d_{\|\cdot\|}(\text{conv}(\{x_j : j \in B_i\}), \text{conv}(\{x_j : j \notin B_i\})) \geq 2\gamma. \quad (3.2)$$

Indeed, if $y \in \text{conv}(\{x_j : j \in B_i\})$ and $z \in \text{conv}(\{x_j : j \notin B_i\})$ then $\|y - z\| \geq x_i^*(y - z) \geq 2\gamma$.

Up to a constant factor 2, the reverse direction also holds true in ℓ_2^n ; that is, if $\{x_1, \dots, x_n\} \subset B_2^n$ and

$$d_{\|\cdot\|}(\text{conv}(\{x_j : j \in B_i\}), \text{conv}(\{x_j : j \notin B_i\})) \geq 2\gamma$$

then F embeds with margin $\gamma/2$ in ℓ_2^n . Indeed, by the Hahn-Banach Theorem there are functionals $(x_i^*) \subset B_{X^*}$ that separate the sets $K_i^+ = \text{conv}(\{x_j : j \in B_i\})$ and $K_i^- = \text{conv}(\{x_j : j \notin B_i\})$ in the sense that for every $1 \leq i \leq N$ there is some s_i such that $x_i^* \geq s_i + \gamma$ on K_i^+ and $x_i^* \leq s_i - \gamma$ on K_i^- . Since s_i need not be 0, one has to increase the dimension by 1, by mapping each x_j to $y_j = (x_j \oplus 1) \in \ell_2^{n+1}$ and x_i^* to a functional y_i^* on ℓ_2^{n+1} defined as $y_i^*(x \oplus a) = x_i^*(x) - a s_i$. Note that necessarily $|s_i| \leq 1 - \gamma$ and thus $\|y_j\| \leq \sqrt{2}$ and $\|y_i^*\| \leq (1 + (1 - \gamma)^2)^{1/2} \leq \sqrt{2}$. Therefore, the set $(y_i^*/\sqrt{2})_{i=1}^N, (y_j/\sqrt{2})_{j=1}^n$ define an embedding of F with margin $\gamma/2$ in ℓ_2^{n+1} . Since the span of $(y_j)_{j=1}^n$ is isomorphic to a subspace of ℓ_2^n (which could be ℓ_2^n itself), this is actually an embedding of F with margin $\gamma/2$ in ℓ_2^n .

A similar proof shows that for any space X , if $\{x_1, \dots, x_n\} \subset B_X$ and

$$d_{\parallel}(\text{conv}(\{x_j : j \in B_i\}), \text{conv}(\{x_j : j \notin B_i\})) \geq 2\gamma,$$

then for any space $Y \neq X$ which contain X as a subspace, F embeds in Y with margin $\gamma/4$.

Again, one can define the notion of a representation of F in terms of the margin one is interested in.

Definition 3.7 Let $F \subset \{-1, 1\}^n$ with $|F| = N$. A set $T = \{t_i : 1 \leq i \leq N\} \subset \mathbb{R}^n$ γ -represents F if for every $1 \leq i \leq N$, $t_i = (\varepsilon_{i,j} b_{i,j})_{j=1}^n \in T$, where $\varepsilon_{i,j} = f_i(j)$ and $b_{i,j} \geq \gamma$.

For $0 < \gamma, \delta \leq 1$ the set T (γ, δ) -softly represents F if for every $1 \leq i \leq N$ there is a set A_i as in Definition 3.6, and $t_i = (\varepsilon_{i,j} b_{i,j})_{j=1}^n$ satisfies that $\varepsilon_{i,j}$ are as above and $b_{i,j} \geq \gamma$ for $j \in A_i$.

Definition 3.7 (for $\delta = 1$) could be understood in the following way: the elements of F define the set of “significant quadrants” \mathcal{Q}_F in the unit cube B_∞^n . Each significant quadrant $Q \in \mathcal{Q}_F$ is represented by an element x in that quadrant, where all the absolute values the coordinates of x are larger than γ .

The following characterizes the ability to (γ, δ) -softly embed with margin a given class in a Banach space. Its proof is analogous to the proof of Theorem 3.3 and is omitted.

Theorem 3.8 For $0 < \gamma, \delta \leq 1$, a class $F \subset \{-1, 1\}^n$ (γ, δ) -softly embeds with margin in X if and only if there is an n -dimensional subspace $E \subset X$ and a set T which (γ, δ) -softly represents F , such that the operator $\text{Id} : (\mathbb{R}^n, \|\cdot\|_{\text{absconv}(T)}) \rightarrow \ell_\infty^n$ factors through E^* .

The first corollary which follows from Theorem 3.8 (and could be shown directly with a different proof), is that every class F has a “natural” Banach space which is completely compatible with the geometry of F , and in which F embeds with margin 1.

Corollary 3.9 *For any class $F \subset \{-1, 1\}^n$ there is a Banach space X of dimension n such that $m(F, X) = 1$. In particular, if $|F| \geq 2$, $L(F, X) = 1$.*

Proof. For $1 \leq i \leq n$ define $t_i = (\varepsilon_{i,j})_{j=1}^n$ by $\varepsilon_{i,j} = f_i(j)$. Thus, T 1-represents F . Let $X = (\mathbb{R}^n, \|\cdot\|)$ where $\|x\|_X = \sup_{t \in T} |\langle t, x \rangle|$. Therefore, the dual unit ball of X is $B_{X^*} = \text{absconv}(T)$ and the identity operator factors through X^* by taking $u = v = Id$. ■

Next, one should consider the margin with which any class embeds in X . We will be interested in comparing $m(F, X)$ with this *trivial margin*, which is attained for $F = \{-1, 1\}^n$. To identify the trivial margin, set

$$\rho_X(n) = \inf_{E \subset X, \dim(E)=n} d(E, \ell_1^n), \quad (3.3)$$

where $d(X, Y)$ is the Banach-Mazur distance between X and Y . Hence, $\rho_X(n)$ measures the minimal distance between an n -dimensional subspace of X and ℓ_1^n . If X is n -dimensional then $\rho_X(n) = d(X, \ell_1^n)$.

Lemma 3.10 *If $F = \{-1, 1\}^n$ then $m(F, X) = 1/\rho_X(n)$. In particular, $L(F, X) \leq \rho_X(n)$.*

Proof. Let $T \subset B_\infty^n$ be a set which γ -represents F . By Theorem 3.8, F embeds with margin γ in X if and only if there is a subspace $E \subset X$ with $\dim(E) = n$ and an operator $U : \mathbb{R}^n \rightarrow \mathbb{R}^n$, such that $\text{absconv}(T) \subset U^*B_{E^*} \subset B_\infty^n$. Since T γ -represents $\{-1, 1\}^n$ then $\gamma B_\infty^n \subset \text{absconv}(T)$, implying that

$$\rho_X(n) \leq d(E, \ell_1^n) = d(E^*, \ell_\infty^n) \leq 1/\gamma.$$

The reverse inequality follows a similar path, by choosing subspaces for which the infimum in the definition of $\rho_X(n)$ is almost attained. ■

Observe that the trivial margin for $X = \ell_2$ is $1/\sqrt{n}$. Also, if ℓ_1 is finitely represented in X , then X contains almost isometric copies of ℓ_1^n for an arbitrarily large n . Hence, for any $\gamma < 1$, any class embeds with margin γ in X .

A somewhat less trivial application of Theorem 3.8 is that classes which are “low dimensional” embed with a large margin in ℓ_2 . For every $d > 0$, let

$N(F, \sqrt{d}B_2^n)$ be the minimal number of translates of the ball $\sqrt{d}B_2^n$ centered at elements of F needed to cover F .

Recall that F is called the d -cross if there is some $x \in \{-1, 1\}^n$ such that

$$F = \{y \in \{-1, 1\}^n : d_H(x, y) \leq d\},$$

where d_H is the Hamming metric. A (k, d) -cross is the union of k such d -crosses.

Theorem 3.11 *If $\rho_d = \max\{d, N(F, \sqrt{d}B_2^n)\}$ then F embeds with margin $\gamma \geq \sup_d \frac{1}{8\sqrt{\rho_d}}$ in ℓ_2 . In particular, a (k, d) -cross embeds with margin $\gamma \geq c/\sqrt{\max\{d, k\}}$, implying that it L -embeds with a Lipschitz function in ℓ_2 for $L \leq C\sqrt{\max\{d, k\}}$.*

Proof. Fix d , denote by $v_1, \dots, v_N \in \{-1, 1\}^n$ the elements of F , let $k = N(F, \sqrt{d}B_2^n)$ and assume that v_1, \dots, v_k are the elements of the cover. One can clearly assume that $k \leq n$, otherwise the assertion of the theorem is immediate.

Consider the operator $R : \ell_2^k \rightarrow \ell_\infty^n$ given by $Re_i = v_i/8\sqrt{\rho_d}$ and note that if $\|a\|_2 \leq 1$ then

$$\left\| \sum_{i=1}^k a_i Re_i \right\|_\infty = \frac{1}{8\sqrt{\rho_d}} \sup_l \left\langle \sum_{i=1}^k a_i v_i, e_l \right\rangle \leq \frac{1}{8} \sqrt{\frac{k}{\rho_d}} \leq \frac{1}{8},$$

implying that $RB_2^k \subseteq \frac{1}{8}B_\infty^n$.

Let \mathcal{E} be any ellipsoid with principal axes u_1, \dots, u_k . For every $r > 0$, let \mathcal{E}' be an n -dimensional ellipsoid with k of its principal directions being the same as those of \mathcal{E} and its axes lengths are $\max\{\|u_i\|_2, r\}$ for $i = 1, \dots, k$, and r for the other $n - k$ axes. Then, $\mathcal{E} + rB_2^n \subset 2\mathcal{E}'$ and $\mathcal{E}' \subset 2(\mathcal{E} + rB_2^n)$. Hence, for every $1 \leq i \leq N$ and selecting $\mathcal{E} = RB_2^k$ and $r = 1/8$,

$$\frac{v_i}{8\sqrt{\rho_d}} \in \mathcal{E} + \frac{1}{8}B_2^n \subset 2\mathcal{E}' \subset 4 \left(\mathcal{E} + \frac{1}{8}B_2^n \right) \subset \frac{1}{2} (B_\infty^n + B_2^n) \subset B_\infty^n,$$

showing that the set $T = \{v_i/8\sqrt{\rho_d}\}$ which $1/8\sqrt{\rho_d}$ -represents F satisfies that

$$\text{absconv}(T) \subset 2\mathcal{E}' \subset B_\infty^n.$$

The assertion now follows from Theorem 3.8 for $\delta = 1$. ■

Theorem 3.8 has a geometric interpretation, which is a way of expressing the fact that the identity operator factors through E^* .

Corollary 3.12 *Let $F \subset \{-1, 1\}^n$ and put \mathcal{Q}_F to be the set of the “significant quadrants” defined by F . For every $0 < \gamma \leq 1$, F embeds with margin γ in X if and only if the following holds. There is a projection P of rank n such that $P_E B_{X^*} \subset B_\infty^n$, and for every $Q \in \mathcal{Q}_F$ there is some $x \in P_E B_{X^*} \cap Q$, which satisfies that $|x_j| \geq \gamma$ for every $1 \leq j \leq n$.*

If X is a Hilbert space then all of its unit ball’s projections are ellipsoids. Therefore, Corollary 3.12 means that embedding F with margin γ in ℓ_2 is equivalent to finding an ellipsoid $\mathcal{E} \subset B_\infty^n$ whose intersection with each $Q \in \mathcal{Q}_F$ contains a point with “large” coordinates.

Another corollary of Theorem 3.8 is that the best margin can be bounded using the vc dimension of the class. Therefore, the ability to embed a class with a large margin is formally stronger than the class having a small vc dimension.

Lemma 3.13 *Let $F \subset \{-1, 1\}^n$ with $\text{vc}(F) = d$. Then, for every space X , $m(F, X) \leq 1/\rho_X(d)$.*

Proof. Assume that T γ -represents F and satisfies that $\text{absconv}(T) \subset P_E B_{X^*} \subset B_\infty^n$. Then, there is a coordinate projection P_σ such that

$$P_\sigma F = \{(f(i))_{i \in \sigma} : f \in F\} = \{-1, 1\}^d,$$

and in particular, $\gamma B_\infty^d \subset P_\sigma(\text{absconv}(T))$. Since $P_\sigma B_\infty^n = B_\infty^d$, it follows that there is a projection P of rank d such that $\gamma B_\infty^d \subset P B_{X^*} \subset B_\infty^d$, and passing to the dual spaces, $\rho_X(d) \leq 1/\gamma$. ■

Remark 3.14 Since $\rho_{\ell_2}(d) = 1/d^{1/2}$, a (k, d) -cross satisfies that $m(F) \sim 1/d^{1/2}$ for $k \leq d$

4 Soft embeddings of random classes

In this section we present our main result, that for a large variety of Banach spaces X , a random class with few elements only trivially embeds with a Lipschitz function in X . In fact, we show that this is the case even for soft embeddings. The model we use for a random class is simply a random subset of $\{-1, 1\}^n$.

Definition 4.1 Let η be a $\{-1, 1\}$ -valued symmetric random variable. For every $1 \leq i \leq N$, let $f_i = \sum_{j=1}^n \eta_{i,j} e_j$, where $(\eta_{i,j})$ are independent $\{-1, 1\}$ -valued random variables distributed as η . If we set $F = \{f_1, \dots, f_N\}$ then F is a random subset of $\{-1, 1\}^n$, which we call a random class of functions.

Here, we extend the following theorem from [7] on the inability to embed with large margin a random class in a Hilbert space.

Theorem 4.2 There are constants C and c such that if $N \geq cn$ and if $|F| = N$, then with probability larger than $1 - e^{-N/8}$,

$$m(F) \leq C/\sqrt{n} = Cm(\{-1, 1\}^n).$$

Theorem 4.2 is extended in three directions. First, we show that for a random class, not only is the optimal margin close to the trivial one, but that a random class does not embed with a Lipschitz function with a nontrivial constant (the trivial being $\rho_X(n)$). Second, our result holds even for soft embeddings. Finally, it holds for a large variety of Banach spaces and not just for a Hilbert space.

For the proof we require a slightly larger random class than in Theorem 4.2, with $N \geq Cn \log n$ elements.

Theorem 4.3 For any $1/2 < \delta \leq 1$ there is an integer n_δ and positive constants $C(\delta)$, $c'(\delta)$ and c , such that for every $n \geq n_\delta$ and any n -dimensional Banach space X the following holds. With probability at least $1 - e^{-cN}$ a random class F with $|F| = N \geq C(\delta)n \log n$ does not (L, δ) -softly embed with a Lipschitz function in X with a constant smaller than $c'(\delta)\rho_X$ ($c'(\delta)n/\log n$). Moreover, if X has type p then F does not (L, δ) -softly embed with a Lipschitz function in X for $L \leq c'(\delta)n^{1-1/p}/T_p(X)$.

As we have shown, every $F \subset \{-1, 1\}^n$ embeds in any Banach space X with margin larger than $1/\rho_X(n)$, and thus, a random class satisfies that

$$c'(\delta) \cdot \rho_X \left(c'(\delta) \frac{n}{\log n} \right) \leq L(F, X, \delta) \leq \rho_X(n). \quad (4.1)$$

For most “reasonable” spaces the gap between $\rho_X(n)$ and $\rho_X(cn/\log n)$ is not large. We will be particularly interested in spaces for which the distances between sections of X of dimension cn and ℓ_1^{cn} change smoothly in c , since for those spaces the gap between the upper and lower bounds in (4.1) is small. The smoothness assumption we require is that there is a positive function θ_X , such that for every $0 < t < 1$,

$$\rho_X(tn) \geq \theta_X(t)\rho_X(n). \quad (4.2)$$

It turns out that many Banach spaces satisfy (4.2), including all the ℓ_p^n spaces and spaces with a bounded type 2 constant. In fact, for spaces with a bounded type 2 constant the superfluous logarithmic factor in (4.1) can be completely removed.

To make the presentation smoother, we first formulate and prove a slightly simpler result than Theorem 4.3, dealing with embedding with margin. In particular, we show that under assumption (4.2), it is impossible to embed a random class with a margin better than $c/\rho_X(n)$. We then indicate the modifications required for the proof of Theorem 4.3.

Theorem 4.4 *For any $1/2 < \delta \leq 1$ there is an integer n_δ and positive constants $C(\delta)$, $c'(\delta)$ and c such that for every $n \geq n_\delta$ and any n -dimensional Banach space X the following holds. With probability at least $1 - e^{-cN}$ a random class F with $|F| = N \geq C(\delta)n \log n$ does not (γ, δ) -softly embed with margin in X for*

$$\gamma \geq \frac{1}{\rho_X(c'(\delta)n)}.$$

In particular, if X satisfies (4.2), the assertion holds for

$$\gamma \geq \frac{1}{\theta_X(c'(\delta))\rho_X(n)}.$$

Before presenting the proof, we require the following preliminary result, on the so-called “problem of Zarankiewicz”.

Lemma 4.5 [2] *Let G be a bipartite graph with (m, n) vertices and denote by $Z(m, n, s, t)$ the maximal number of edges in G such that G does not contain an (s, t) -complete bipartite subgraph. Then,*

$$Z(m, n, s, t) \leq (s-1)^{1/t}(n-t+1)m^{1-1/t} + (t-1)m.$$

Another fact we need is the well known Sauer-Shelah Lemma (see, for example, [17]).

Lemma 4.6 *Let $F \subset \{-1, 1\}^n$. If $d = \text{vc}(F)$ then for every $\sigma \subset \{1, \dots, n\}$, $|P_\sigma F| \leq \sum_{i=0}^d \binom{|\sigma|}{i} \leq (e|\sigma|/d)^d$, where the last inequality holds if $|\sigma| \geq d$.*

The key observation in the proof of Theorem 4.4 is the following:

Lemma 4.7 *For every $1/2 < \delta \leq 1$ there exist a constant $c(\delta)$ and an integer n_δ for which the following holds. Let $n \geq n_\delta$, $\gamma \geq 1/\rho_X(c(\delta)n)$, $x_1, \dots, x_n \in B_X$ and $\Delta = \frac{1}{2}(\delta - \frac{1}{2})(1 - \log_2(3 - 2\delta)) > 0$. There is a set $Q \subset \{-1, 1\}^n$ of cardinality at most $2^{n(1-\Delta)}$, such that if F (γ, δ) -softly embeds with margin in X using x_1, \dots, x_n , then $F \subset Q$.*

Proof. Let $U : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be given by $Ue_i = x_i$ and set $K = U^*B_{X^*} \subset B_\infty^n$. Define a set $\mathcal{Q}_K \subset \{-1, 1\}^n$ as follows: $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n) \in \mathcal{Q}_K$ if there are $A_\varepsilon \subset \{1, \dots, n\}$ and $y_\varepsilon \in K$ such that $|A_\varepsilon| \geq \delta n$, and if $y_\varepsilon(j)$ is the j -th coordinate of y_ε then

$$\text{sign}(y_\varepsilon(j)) = \varepsilon_j \quad \text{and} \quad |y_\varepsilon(j)| \geq \gamma > \frac{1}{\rho_X(c(\delta)n)} \quad \text{if } j \in A_\varepsilon. \quad (4.3)$$

From the proof of Theorem 3.8 (which is based on the same idea as the proof of Theorem 3.3) it is evident that the set \mathcal{Q}_K contains the points in $\{-1, 1\}^n$ that can possibly be (γ, δ) -softly embedded with margin using x_1, \dots, x_n . Thus, if F embeds in X using x_1, \dots, x_n then $F \subset \mathcal{Q}_K$. What is left to show is that $|\mathcal{Q}_K| \leq 2^{n(1-\Delta)}$.

To that end, assume that $|\mathcal{Q}_K| > 2^{n(1-\Delta)}$ and define a bipartite graph in the following manner. One side consists of the elements of \mathcal{Q}_K and the other side is the elements of $\{1, \dots, n\}$. There is an edge between ε and j if and only if $\text{sign}(y_\varepsilon(j)) = \varepsilon_j$ and $|y_\varepsilon(j)| \geq \gamma$. Using the notation of Lemma 4.5, $m > 2^{n(1-\Delta)}$ and the graph contains at least δmn edges. Hence, by Lemma 4.5, if s and t satisfy

$$\delta mn > (s-1)^{1/t}(n-t+1)m^{1-1/t} + (t-1)m \quad (4.4)$$

then G contains a complete (s, t) bipartite subgraph. Setting $\alpha = \delta - 1/2$, $t-1 = \alpha n$ and $(s-1) = 2^{\beta n}$, an easy computation shows that

$$\beta < 1 + \alpha \log_2 \left(\frac{\delta - \alpha}{1 - \alpha} \right) + \frac{1}{n} \left(\log_2 \left(\frac{\delta - \alpha}{1 - \alpha} \right) - \Delta n \right) \quad (4.5)$$

is enough to ensure (4.4). Note that one can choose $\beta > 0$ satisfying (4.5) such that, for $n \geq n_\delta$, $\alpha + \beta \geq 1 + \Delta/2$.

In particular, for every $\delta > 1/2$ there is some n_δ and constants α and β , all of which depend only on δ , for which the following holds: if $n \geq n_\delta$, then $\alpha + \beta - 1 \geq \Delta/2$, there is a set \mathcal{Q}_1 of cardinality at least $2^{\beta n}$ such that $\mathcal{Q}_1 \subset \mathcal{Q}_K$ and a set J of αn coordinates in $\{1, \dots, n\}$ which satisfy that for every $j \in J$ and $\varepsilon \in \mathcal{Q}_1$, $\text{sign}(y_\varepsilon(j)) = \varepsilon_j$ and $|y_\varepsilon(j)| \geq \gamma$.

Consider the coordinate projection $P_J \mathcal{Q}_1$. Since $|\mathcal{Q}_1| \geq 2^{\beta n}$ and $|J| = \alpha n$, then $|P_J \mathcal{Q}_1| \geq 2^{\beta n} / 2^{n-\alpha n}$. Indeed, any point in $P_J \mathcal{Q}_1$ is the image of at most $2^{n-\alpha n}$ elements in $\{-1, 1\}^n$. As $\alpha + \beta - 1 \geq \Delta/2$, it is evident that $|P_J \mathcal{Q}_1| \geq 2^{n\Delta/2}$. Applying the Sauer-Shelah Lemma, there is a subset $J_1 \subset J$, such that $|J_1| \geq c(\delta)n$ and $P_{J_1} \mathcal{Q}_1 = P_{J_1} P_J \mathcal{Q}_K = \{-1, 1\}^{|J_1|}$. From this, it immediately follows that

$$\gamma B_\infty^{|J_1|} \subset P_{J_1} (\text{conv}(\{y_\varepsilon : \varepsilon \in \mathcal{Q}_K\})) \subset P_{J_1} K,$$

and thus $\gamma B_\infty^{|J_1|} \subset P_{J_1} K \subset B_\infty^{|J_1|}$. By duality, there is a subspace $E \subset X$, such that $\dim(E) \geq c(\delta)n$ and $d(E, \ell_1^{\dim(E)}) \leq 1/\gamma$, implying that $\rho_X(c(\delta)n) \leq 1/\gamma$. This contradicts the choice of γ , proving that $|\mathcal{Q}_K| \leq 2^{n(1-\Delta)}$. ■

Proof (of Theorem 4.4). To prove that it is only possible to embed a class with a trivial margin, it suffices to consider a fine enough net in B_X . Indeed, observe that if $\dim(X) = n$ then by John's Theorem, $d(X, \ell_1^n) \leq n$, and thus the trivial margin is at worst $1/n$ (in fact, it was proved in [5] that if $\dim(X) = n$ then $d(X, \ell_1^n) \leq cn^{5/6}$ showing that the trivial margin is at least $c/n^{5/6}$). Let $D \subset B_X$ be a $1/n$ net with respect to the norm $\|\cdot\|_X$. For any $r > 1$, if there is an embedding of F in X with (soft) margin $\gamma \geq r/n$, then replacing x_1, \dots, x_n with appropriate points in the net, F embeds with margin $(r-1)/\gamma$ using elements of D . Therefore, it suffices to show that no such embedding exists using $\{x_1, \dots, x_n\} \in D^n$. By a simple volumetric estimate, $|D^n| \leq e^{C'n^2 \log n}$ for a suitable absolute constant C' .

Moreover, one needs only to consider the subset of D^n which yield a margin larger than $1/\rho_X(c(\delta)n)$ for a suitable constant $c(\delta)$, which is chosen to be the constant from Lemma 4.7.

Fix such a set $\{x_1, \dots, x_n\} \in D^n$. By Lemma 4.7, there exists a set \mathcal{Q} of cardinality at most $2^{n(1-\Delta)}$ which satisfies that if F (γ, δ)-softly embeds with margin in X using $\{x_1, \dots, x_n\}$, then $F \subset \mathcal{Q}$. Hence, if F is a random class with N elements, the probability that $\{x_1, \dots, x_n\}$ can be used to embed F with a soft margin larger than $1/\rho_X(c(\delta)n)$ is upper bounded by $e^{-c(\delta)Nn}$. Therefore, the probability that F embeds using some "legal" $\{x_1, \dots, x_n\} \in D^n$ is at most $e^{-c(\delta)Nn} |D^n| \leq e^{-c(\delta)Nn} e^{C'n^2 \log n} \leq e^{-cN}$ for our choice of $N \geq C(\delta)n \log n$.

The second part of the claim is evident, since (4.2) implies that

$$\rho_X(c(\delta)n) \geq \theta_X(c(\delta)) \rho_X(n),$$

and thus

$$\gamma \leq \frac{1}{\theta_X(c(\delta)) \rho_X(n)}.$$

■

Finally, we turn to the proof of Theorem 4.3 which requires more preparation.

Some reflection on the previous proof reveals that the clear geometric interpretation of the margin yields the existence of a high dimensional coordinate projection of K that contains a cube. Indeed, it follows immediately

from the fact that $P_{J_1} Q_K = \{-1, 1\}^{|J_1|}$ and that on these coordinates one controls the signs and the absolute values of enough vectors y_ε . In the general case, an additional argument is needed to construct a “large cube”. The construction is based on a real-valued analog of the vc dimension, called the *combinatorial dimension*.

Definition 4.8 *Let G be a set of functions on Ω . For every $\varepsilon > 0$, a set $\sigma = \{x_1, \dots, x_m\} \subset \Omega$ is said to be ε -shattered by G , if there is some function $s : \sigma \rightarrow \mathbb{R}$ with the following property. For every $I \subset \{1, \dots, m\}$ there is some $g_I \in G$ for which $g_I(x_i) \geq s(x_i) + \varepsilon$ if $i \in I$ and $g_I(x_i) \leq s(x_i) - \varepsilon$ if $i \notin I$. The combinatorial dimension of G is the function*

$$\text{vc}(\varepsilon, G, \Omega) = \sup \{|\sigma| \mid \sigma \subset \Omega, \sigma \text{ is } \varepsilon\text{-shattered by } G\}.$$

If the underlying space is clear, the combinatorial dimension is denoted by $\text{vc}(\varepsilon, G)$.

In our case, Ω is the fixed coordinate system $\{e_1, \dots, e_n\}$ and each vector $v \in \mathbb{R}^n$ is identified with a function on Ω by $g_v(i) = \langle v, e_i \rangle$.

It is easy to show (see, e.g. [10]) that if G is a convex and symmetric set of functions, and if $\{x_1, \dots, x_m\}$ is ε -shattered by G , one can take $s(x_i) = 0$ for every i . In particular, if $K \subset \mathbb{R}^n$ is a convex, symmetric set and if $\sigma \subset \{e_1, \dots, e_n\}$ is ε -shattered by K (when considered as a set of functions on the coordinates), then $\varepsilon B_\infty^{|\sigma|} \subset P_\sigma K$.

Recall that if (X, d) is a metric space, a set $A \subset X$ is ε -separated if for every $a_1, a_2 \in A$, $d(a_1, a_2) \geq \varepsilon$. If μ is a probability measure, denote by $D(\varepsilon, G, L_2(\mu))$ the largest cardinality of an ε -separated subset of G with respect to the $L_2(\mu)$ metric.

Theorem 4.9 [9] *There are absolute constants C and c such that for any set G which consists of functions bounded by 1, every $0 < \varepsilon < 1$ and every probability measure μ ,*

$$D(\varepsilon, G, L_2(\mu)) \leq \left(\frac{2}{\varepsilon}\right)^{C \cdot \text{vc}(c\varepsilon, G)}.$$

Let μ be the uniform probability measure on $\{e_1, \dots, e_n\}$ and set L_2^n to be the corresponding L_2 space. Theorem 4.9 implies that if $K \subset B_\infty^n$ contains a large ε -separated set in L_2^n then there is a high dimensional coordinate projection $P_\sigma K$ which contains $c\varepsilon B_\infty^{|\sigma|}$.

Although Theorem 4.9 gives the best possible estimate in the general case, under some additional assumptions it is possible to remove the logarithm in the above estimate. This fact is a very recent result due to M. Rudelson and R. Vershynin.

Theorem 4.10 [14] *There is an absolute constant C for which the following holds. Let F be a set of functions and let $t > 0$. Assume that there is a decreasing function $v(t)$ and some $a > 2$ such that for every $s \geq t$, $\text{vc}(s, F) \leq v(s)$ and $v(as) \leq \frac{1}{2}v(s)$. Then, for any probability measure μ ,*

$$\log D(t, F, L_2(\mu)) \leq Cv(t) \log a.$$

Let $K = UB_{X^*} \subset B_\infty^n$ be convex and symmetric, and consider it to be a set of functions on the given coordinate structure $\{e_1, \dots, e_n\}$. Observe that if σ is t -shattered by K then $d(P_\sigma K, \ell_\infty^{|\sigma|}) \leq 1/t$, and thus $\rho_X(|\sigma|) \leq 1/t$. On the other hand, it is easy to verify that if X has type p then for every m , $\rho_X(m) \geq m^{1-1/p}/T_p(X)$. Thus,

$$\text{vc}(t, K, \{e_1, \dots, e_n\}) \leq \left(\frac{T_p(X)}{t} \right)^{\frac{p}{p-1}} \equiv v(t).$$

Since the assumptions of Theorem 4.10 hold for $a = 3$ and every t , it follows that for the uniform measure on $\{e_1, \dots, e_n\}$,

$$\log D(t, F, L_2(\mu)) \leq C \left(\frac{T_p(X)}{t} \right)^{\frac{p}{p-1}}. \quad (4.6)$$

Proof (of Theorem 4.3). The first step in the proof is to consider a finite approximating set to the set of all “meaningful” Lipschitz functions $\phi : \mathbb{R} \rightarrow \mathbb{R}$ and possible sets $\{x_1, \dots, x_n\} \subset B_X$ that can be used in the embedding. Clearly, it is enough to consider functions that map $[-1, 1]$ to $[-1, 1]$ (since $x_i^*(x_j) \in [-1, 1]$, and if the range of ϕ exceeds $[-1, 1]$ it is possible to compose ϕ with the retraction onto $[-1, 1]$). Also, one can assume that ϕ has a Lipschitz constant bounded by n , since $L(F, X) \leq L(\{-1, 1\}^n, X) \leq \rho_X(n) \leq n$. Thus, one can identify each “legal” ϕ with the pair of nonempty sets $W_+ = \{t \mid \phi(t) = 1\}$ and $W_- = \{t \mid \phi(t) = -1\}$, such that for $n \geq n_\delta$,

$$d(W_+, W_-) > \frac{1}{L} \geq \frac{C(\delta)}{\rho_X(c(\delta)n/\log(n))} \geq \frac{1}{n},$$

where $C(\delta)$, $c(\delta)$ and n_δ are positive constants depending only on δ which will be specified later.

For $-n^2 \leq i \leq n^2 - 1$, let $Y_i = [i/n^2, (i+1)/n^2]$ and set D to be a $1/4n$ net in B_X . Without loss of generality, one can assume that $(x_j)_{j=1}^n \in D^n$ and that $\{\phi = 1\}, \{\phi = -1\}$ are of the form $\bigcup_{i \in I} Y_i$. Indeed, if $\{z_1, \dots, z_n\} \in B_X$, $\{x_1^*, \dots, x_n^*\} \in B_{X^*}$ and ϕ is a Lipschitz function as above which can be used to (L, δ) -embed F , first replace $\{z_1, \dots, z_n\}$ by an appropriate approximation $(x_1, \dots, x_n) \in D^n$. Clearly, if $x_i^*(z_j) \in W_+$ then $x_i^*(x_j) \in W_+ + (-1/4n, 1/4n)$. Let I_+ be the set of integers $-n^2 \leq i \leq n^2 - 1$ such that $Y_i \cap (W_+ + (-1/4n, 1/4n))$ is nonempty and set V_+ to be the closure of $\bigcup_{i \in I_+} Y_i$. Similarly, define V_- , and let $\tilde{\phi} : [-1, 1] \rightarrow [-1, 1]$ be the piecewise linear function which is 1 on V_+ and -1 on V_- . It is easy to verify that $d(V_+, V_-) > 1/2L$ and that $\|\tilde{\phi}\|_{\text{lip}} < L$. Also, if $x_i^*(x_j) \in W_+$ (resp. W_-) then $x_i^*(x_j) \in V_+$ (resp. V_-), proving the assertion. Note that there are at most $e^{cn^2 \log n}$ triplets $((x_1, \dots, x_n), V_+, V_-)$ and fix a pair of sets V_+, V_- which are at least $\frac{C(\delta)}{\rho_X(c(\delta)n/\log(n))}$ apart and $(x_1, \dots, x_n) \in D^n$. In a similar manner to the proof of Theorem 4.4, put $K = U^*B_{X^*}$ where $Ue_i = x_i$. Applying Theorem 3.4, the proof will be concluded by showing that for every fixed triplet K, V_+, V_- , the probability that a random class embeds using the ‘‘position’’ K (in the sense of the factorization) and a Lipschitz function which is 1 on V_+ and -1 on V_- is at most $e^{-c(\delta)Nn}$.

To that end, define the set $\mathcal{Q} = \mathcal{Q}_{K, V_+, V_-} \subset \{-1, 1\}^n$ as follows: for every $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n) \in \{-1, 1\}^n$, $\varepsilon \in \mathcal{Q}$ if there is a set $A_\varepsilon \subset \{1, \dots, n\}$ of cardinality $|A_\varepsilon| \geq \delta n$ and some $y_\varepsilon \in K$ such that if $j \in A_\varepsilon$ then $y_\varepsilon(j) \in V_+$ if $\varepsilon_j = 1$ and $y_\varepsilon(j) \in V_-$ if $\varepsilon_j = -1$. If $\varepsilon \in \mathcal{Q}$, denote by y_ε any point in K with the above property. Using the same argument as in the previous proof, it remains is to show that $|\mathcal{Q}| \leq 2^{n(1-\Delta)}$, where Δ was defined in Lemma 4.7.

Assume that $|\mathcal{Q}| > 2^{n(1-\Delta)}$. Applying Lemma 4.5 and the Sauer-Shelah Lemma just as in the proof of Lemma 4.7, there are subsets $\mathcal{Q}_1 \subset \mathcal{Q}$ and $J_1 \subset \{1, \dots, n\}$ such that $|J_1| \geq c(\delta)n$, $P_{J_1}\mathcal{Q}_1 = \{-1, 1\}^{|J_1|}$, and for every $\varepsilon \in \mathcal{Q}_1$ and every $j \in J_1$, $y_\varepsilon(j) \in V_+$ if $\varepsilon_j = 1$ and $y_\varepsilon(j) \in V_-$ if $\varepsilon_j = -1$.

It is well known that there is a set $\mathcal{Q}_2 \subset \{-1, 1\}^{|J_1|} \subset P_{J_1}\mathcal{Q}$ which is $|J_1|/4$ separated in the Hamming metric and is of cardinality $2^{c|J_1|}$ for a suitable absolute constant c . In particular, any two elements of \mathcal{Q} which are projected onto distinct elements in \mathcal{Q}_2 differ on at least $c(\delta)n$ coordinates.

Since $d(V_+, V_-) > 1/2L$, if ε and ε' differ on $j \in J_1$ then $|y_\varepsilon(j) - y_{\varepsilon'}(j)| > 1/2L$. Hence, for every $\varepsilon, \varepsilon' \in \mathcal{Q}$ such that $P_{J_1}\varepsilon \neq P_{J_1}\varepsilon'$ and $P_{J_1}\varepsilon, P_{J_1}\varepsilon' \in \mathcal{Q}_2$,

$$\|y_\varepsilon - y_{\varepsilon'}\|_{L_2^2}^2 \geq \frac{1}{n} \sum_{j \in J_1} |y_\varepsilon(j) - y_{\varepsilon'}(j)|^2 \geq \frac{|J_1|}{4nL^2} = \frac{c(\delta)}{L^2}.$$

Therefore, the set $\{y_\varepsilon | \varepsilon \in Q\} \subset K$ contains a subset of cardinality $2^{c'(\delta)n}$ which is $c(\delta)/L$ -separated in L_2^n , and thus

$$\begin{aligned} c(\delta)n &\leq \log D(c'(\delta)/L, K, L_2^n) \leq C(\delta) \cdot \text{vc}(c''(\delta)/L, K) \log(2L) \\ &\leq C(\delta) \text{vc}(c''(\delta)/L, K) \log n. \end{aligned}$$

As K is convex and symmetric, there is a set of coordinates $\sigma \subset \{1, \dots, n\}$ of cardinality $|\sigma| \geq c(\delta)n/\log n$ for which $\frac{c'(\delta)}{L} B_\infty^{|\sigma|} \subset P_\sigma K \subset B_\infty^{|\sigma|}$. Passing to the dual spaces it is evident that

$$\rho_X \left(c(\delta) \frac{n}{\log n} \right) \leq C(\delta)L,$$

which contradicts the choice of L .

For the last part of the theorem, if X has type p then by (4.6), for any $t > 0$, $\log D(t, F, L_2^n) \leq C \left(\frac{T_p(X)}{t} \right)^{\frac{p}{p-1}}$. On the other hand, if $|Q| > 2^{n(1-\Delta)}$ then K contains a subset of cardinality $2^{c'(\delta)n}$ which is $c(\delta)/L$ -separated in L_2^n , implying that $L \geq c(\delta)n^{1-1/p}/T_p(X)$. ■

The proofs of Lemma 4.7 and Theorem 4.3 yield a little more than what is claimed. In fact, it is evident from (4.5) that it is not necessary to use the full strength of the soft embedding assumption, in which for every $1 \leq i \leq N$ one controls the values of $x_i^*(x_j)$ on at least δn coordinates. All that is actually required is that out of the Nn pairs (i, j) , one controls $x_i^*(x_j)$ on at least δNn pairs. Thus, in the spirit of [7], one can consider a slightly weaker notion of soft embedding, in which $\phi(x_i^*(x_j)) = f_i(j)$ for at least δNn pairs, where $1/2 < \delta \leq 1$. Such a definition diverts from our goal of preserving the structure of the function class F , since the behavior of each individual $f \in F$ no longer matters. Instead, with this new definition, the embedding constant should be viewed as a complexity parameter of the $\{-1, 1\}$ -valued matrix $(f_i(j))$.

Note that this viewpoint explains the assumption that $1/2 \leq \delta$, which cannot be relaxed. Indeed, if A and B are $\{-1, 1\}$ -valued, $N \times n$ matrices then $\min\{d_H(A, B), d_H(A, -B)\} \leq 1/2$, where d_H is the natural Hamming distance on $N \times n$ matrices. Since embedding the matrix B (i.e., embedding the function class defined by the rows of B) is equivalent to embedding $-B$, one can take any matrix A , and by changing half of its entries arrive to a matrix that well embeds in X . Hence, it is impossible to obtain any kind of a hardness result in the spirit of Theorem 4.3 if one is allowed to change half the entries of $(f_i(j))$.

4.1 Stability of $\rho_X(tn)$

Next, we show that assumption (4.2) holds for a large variety of spaces. To that end, let us consider several properties of Banach spaces which are preserved if the spaces are “close”. The first such property is type.

It is easy to verify that if $T_p(X)$ is small for some $p > 1$ then X cannot be too close to ℓ_1^n . In fact, as we mentioned before, for any $1 < p \leq 2$, $d(X, \ell_1^n) \geq \frac{n^{1-1/p}}{T_p(X)}$ [16].

Another notion that can be used to bound the distance to ℓ_1^n is the *external volume ratio* of a ball K , defined as

$$\text{evr}(K) = \inf \left(\frac{|TB_2^n|}{|K|} \right)^{\frac{1}{n}},$$

where the infimum is with respect to all $T \in GL_n$ such that $K \subset TB_2^n$. One can show [13] that for every ball K there is a unique ellipsoid of minimal volume containing it. Hence, the external volume ratio is determined by the ratio of the volume of the ellipsoid of minimal volume containing K and the volume of K . It is well known [13] that $\max\{1, n^{\frac{1}{2}-\frac{1}{p}}\}B_2^n$ is the ellipsoid of minimal volume containing B_p^n , and by a simple volumetric estimate,

$$\text{evr}(B_p^n) \sim \begin{cases} n^{\frac{1}{p}-\frac{1}{2}} & 1 \leq p \leq 2, \\ 1 & 2 < p \leq \infty. \end{cases}$$

It is standard to verify the following lower estimate on the distance between two n -dimensional spaces using the external volume ratio.

Lemma 4.11 *Let X and Y be finite dimensional, isomorphic spaces and assume that $\text{evr}(B_X) \geq \text{evr}(B_Y)$. Then,*

$$d(X, Y) \geq \frac{\text{evr}(B_X)}{\text{evr}(B_Y)}.$$

In particular, $d(Y, \ell_1^n) \geq n^{1/2}/\text{evr}(B_Y)$.

Let us start by estimating $\rho_{\ell_p^n}(tn)$ for $1 \leq p \leq 2$. In that range, ℓ_p^n has type p with a constant $T_p(\ell_p^n) \leq C$ for a suitable absolute constant. Hence, for every integer m , $\rho_{\ell_p^n}(m) \geq m^{1-1/p}/T_p(\ell_p^n) \geq cm^{1-1/p}$. On the other hand, $\rho_{\ell_p^n}(n) = n^{1-1/p}$. Therefore,

$$\rho_{\ell_p^n}(tn) \geq ct^{1-1/p}n^{1-1/p} = \theta_{\ell_p^n}(t)\rho_{\ell_p^n}(n).$$

The case $2 < p \leq \infty$ is more delicate. We shall present two proofs to the fact that in this range, ℓ_p^n satisfies (4.2). The first one is very general and holds for all spaces with a bounded external volume ratio. The second one uses specific properties of ℓ_p^n and yields sharper bounds on the function θ , though the bounds are probably not optimal.

The following lemma is known to experts and presented for the sake of completeness.

Lemma 4.12 *There exists an absolute constant C for which the following holds. Let X be an n -dimensional space and let $E \subset \mathbb{R}^n$ be an m -dimensional subspace. Then,*

$$\text{evr}(B_X \cap E) \leq (C \cdot \text{evr}(B_X))^{\frac{n}{m}}.$$

Proof. Without loss of generality, assume that B_2^n is the ellipsoid of minimal volume containing B_X . Clearly, $B_X \cap E \subset B_2^n \cap E = B_2^m$ and $P_{E^\perp} B_X \subset P_{E^\perp} B_2^n = B_2^{n-m}$. By the Rogers-Shephard inequality ([13], Lemma 8.8), for every body $K \subset \mathbb{R}^n$ and a subspace E of dimension m ,

$$|K| \leq |K \cap E| \cdot |P_{E^\perp} K| \leq \binom{n}{m} |K|.$$

In particular, setting $K = B_X$,

$$|B_X| \leq |B_X \cap E| \cdot |P_{E^\perp} B_X| \leq |B_X \cap E| \cdot |B_2^{n-m}|.$$

Therefore,

$$\begin{aligned} \text{evr}(B_X \cap E) &\leq \left(\frac{|B_2^m|}{|B_X \cap E|} \right)^{\frac{1}{m}} \leq (\text{evr}(B_X))^{\frac{n}{m}} \left(\frac{|B_2^{n-m}| \cdot |B_2^m|}{|B_2^n|} \right)^{\frac{1}{m}} \\ &\leq (C \cdot \text{evr}(B_X))^{\frac{n}{m}}, \end{aligned}$$

where the last inequality holds because $|B_2^n|^{1/n} \sim n^{-1/2}$. ■

Corollary 4.13 *There exists an absolute constant C for which the following holds. Let X be an n -dimensional space such that $\text{evr}(B_X) = \alpha$ and $d(X, \ell_1^n) \leq n^{1/2}$. Then, for every $0 < t < 1$*

$$\rho_X(tn) \geq t^{\frac{1}{2}} (C\alpha)^{-\frac{1}{t}} \rho_X(n)$$

Proof. Let $E \subset \mathbb{R}^n$ of dimension tn . From Lemma 4.12, $\text{evr}(B_X \cap E) \leq (C\alpha)^{1/t}$, and thus

$$d(X \cap E, \ell_1^{tn}) \geq \frac{\text{evr}(B_1^{tn})}{\text{evr}(B_X \cap E)} \geq t^{\frac{1}{2}} (C\alpha)^{-\frac{1}{t}} n^{\frac{1}{2}} \geq t^{\frac{1}{2}} (C\alpha)^{-\frac{1}{t}} \rho_X(n). ■$$

Corollary 4.13 suffices to show that assumption (4.2) holds for ℓ_p^n in the range $2 < p \leq \infty$. Indeed, for such values of p , $\text{evr}(B_p^n) \leq C$ for some absolute constant C and $d(\ell_p^n, \ell_1^n) = n^{1/2}$. However, one can obtain a better bound on the external volume ratio of sections of ℓ_p^n and thus on the function $\theta_{\ell_p^n}(t)$.

Lemma 4.14 *There exists an absolute constant C such that for any $2 < p \leq \infty$, every integers $m < n$ and every subspace $E \subset \mathbb{R}^n$ of dimension m ,*

$$\text{evr}(B_p^n \cap E) \leq C \left(\frac{n}{m} \right)^{\frac{1}{2} - \frac{1}{p}}.$$

Proof. Recall that for $2 \leq p \leq \infty$, $n^{\frac{1}{2} - \frac{1}{p}} B_2^n$ is the ellipsoid of minimal volume containing B_p^n . If $E \subset \mathbb{R}^n$ of dimension m then by the Meyer-Pajor Theorem [11], $|B_p^m| \leq |B_p^n \cap E|$. Therefore,

$$\begin{aligned} \text{evr}(B_p^n \cap E) &\leq \left(\frac{|n^{\frac{1}{2} - \frac{1}{p}} B_2^m|}{|B_p^n \cap E|} \right)^{\frac{1}{m}} \leq \left(\frac{n}{m} \right)^{\frac{1}{2} - \frac{1}{p}} \text{evr}(B_p^m) \\ &\leq C \left(\frac{n}{m} \right)^{\frac{1}{2} - \frac{1}{p}}. \end{aligned}$$

■

In an identical way to the proof of Corollary 4.13 one can prove the following

Corollary 4.15 *There exists an absolute constant c such that for every integer n and any $0 < t < 1$,*

$$\rho_{\ell_p^n}(tn) \geq ct^{1 - \frac{1}{p}} \rho_{\ell_p^n}(n).$$

4.2 Spaces with type

It is evident from the second part of Theorem 4.3 that if X has a nontrivial type, it is possible to lower bound the constant with which a random class embeds in X . The case we focus on here is when $\rho_X(n)$ is of the same order of magnitude as the lower bound on L . First, consider the case $X = \ell_p^n$. Since $d(\ell_p^n, \ell_1^n) = n^{1-1/p}$ if $1 \leq p \leq 2$, and for $2 \leq p \leq \infty$, $d(\ell_p^n, \ell_1^n) = n^{1/2}$, then applying the bounds on the appropriate type constant for ℓ_p^n one obtains the following corollary (which, for the sake of simplicity, is presented for $\delta = 1$).

Corollary 4.16 *There are absolute constants c, c_1, c_2, C and an integer n_0 for which the following holds. If $n \geq n_0$ and $F \subset \{-1, 1\}^n$ is a random class with $N \geq cn \log n$ elements, then with probability at least $1 - e^{-c_1 N}$,*

1. *If $1 \leq p \leq 2$ then*

$$m(\{-1, 1\}^n, \ell_p^n) = n^{\frac{1}{p}-1} \leq m(F, \ell_p^n) \leq Cn^{\frac{1}{p}-1}$$

and

$$c_2 n^{1-\frac{1}{p}} \leq L(F, \ell_p^n) \leq n^{1-\frac{1}{p}}.$$

2. *If $2 \leq p \leq \infty$ then*

$$m(\{-1, 1\}^n, \ell_p^n) = n^{-\frac{1}{2}} \leq m(F, \ell_p^n) \leq Cn^{-\frac{1}{2}}$$

and

$$\frac{cn^{\frac{1}{2}}}{T_2(\ell_p^n)} \leq L(F, \ell_p^n) \leq n^{\frac{1}{2}}.$$

Another generic class of spaces for which the distance to ℓ_1^n can be bounded are spaces with a small type 2 constant. This is evident from the next lemma ([16], Theorem 42.3).

Lemma 4.17 *There is an absolute constant c and functions $\phi_n : [1, \infty) \rightarrow \mathbb{R}$ such that for every Banach space X ,*

$$d(X, \ell_1^n) \leq \phi_n(T_2(X)) n^{1/2}, \quad (4.7)$$

where $\phi_n(t) \leq c't$ if $t \leq \left(\frac{\log n}{c \log \log n}\right)^{1/2}$ and $\phi_n(t) \leq c't \log \log n$ for every t .

Corollary 4.18 *There exist absolute constants c, c_1 and c_2 , and for every $1/2 < \delta \leq 1$ there are constants $c_1(\delta)$ and $c_2(\delta)$ such that the following holds.*

Let X be an n -dimensional Banach space for which $T_2(X) \leq \left(\frac{\log n}{c \log \log n}\right)^{1/2}$. If F is a random class with $N \geq c_1(\delta)n \log n$ elements then with probability at least $1 - e^{-c_1 N}$,

$$c_2(\delta) \frac{n^{1/2}}{T_2(X)} \leq L(F, X, \delta) \leq c_2 T_2(X) n^{1/2}.$$

Moreover, by Lemma 4.17 it is clear that spaces with a small type 2 constant satisfy (4.2).

Corollary 4.19 *There are absolute constants c and c' for which the following holds. For every Banach space X of dimension n and every $0 < t < 1$,*

$$\rho_X(tn) \geq \frac{c't^{1/2}}{T_2^2(X)(\log \log n)^{1/2}} \rho_X(n).$$

5 Concluding Remarks

The statements and proofs presented in this article have a mild logarithmic looseness which is probably an artifact of the method of proof. The restriction that $|F| \geq cn \log n$ arises from the counting argument which is based on ε -nets and it is likely that this approach could be improved.

The second parasitic logarithmic term is due to the way a “large cube” is constructed in a coordinate projection of a position of B_{X^*} . It seems that in order to establish sharper bounds in the general case, one must identify specific geometric properties of subsets of $\{-1, 1\}^n$ that make them hard to embed. Our proof is based on bounding the cardinality of the subset on $\{-1, 1\}^n$ that can be embedded using a specific choice of $\{x_1, \dots, x_n\} \subset B_X$.

We conjecture that the following is true.

Conjecture 5.1 *There is an absolute constant c and for every $1/2 < \delta \leq 1$ there are constants $c_1(\delta)$ and $c_2(\delta)$ such that for any Banach space X and $N \geq c_1(\delta)n$ the following holds. With probability at least $1 - e^{-cN}$, a random class F with N elements satisfies that*

$$c_2(\delta)\rho_X(n) \leq L(F, X, \delta) \leq \rho_X(n).$$

Another natural question is whether embeddings using an arbitrary Lipschitz function truly yield a better result than the margin. It seems possible that the best strategy to embed a class could be always to use the margin function ϕ_γ . Unfortunately, we were not able to make much progress in that direction.

References

- [1] S. Ben-David, N. Eiron, H.U. Simon, Limitations of learning via embeddings in Euclidean half spaces, *Journal of Machine Learning Research* 3, 441-461, 2002.
- [2] B. Bollobás, *Extremal graph theory*, Academic Press, 1978.
- [3] Y. Freund, R. Schapire, Large margin classification using the perceptron algorithm, *Machine Learning* 37 (3) 277-296, 1999.
- [4] R. Herbrich, *Learning kernel classifiers*, MIT Press, 2002.
- [5] A.A. Giannopoulos, A note on the Banach-Mazur distance to the cube, in *Geometric aspects of Functional Analysis* (J. Lindenstrauss, V.D. Milman Eds), *Operator Theory: Advances and Applications*, vol 77, 67-73, 1995.
- [6] V. Koltchinskii, D. Panchenko, Empirical margin distributions and bounding the generalization error of combined classifiers, *Annals of Statistics*, 30
- [7] N. Linial, S. Mendelson, G. Schechtman, A. Shraibman, Complexity measures of sign matrices, preprint.
- [8] L. Mason, P.L. Bartlett, J. Baxter, Improved generalization bounds through explicit optimization of margins, *Machine Learning* 38 (3) 243-255, 2000.
- [9] S. Mendelson, R. Vershynin, Entropy and the combinatorial dimension, *Inventiones Mathematicae*, 152(1), 37-55, 2003.
- [10] S. Mendelson, G. Schechtman, The shattering dimension of sets of linear functionals, *Annals of Probability*, July 2004.
- [11] M. Meyer, A. Pajor, Sections of the unit ball of ℓ_p^n , *Journal of Functional Analysis* 80 (1), 109–123, 1988.
- [12] V.D. Milman, G. Schechtman, *Asymptotic theory of finite dimensional normed spaces*, *Lecture Notes in Mathematics* 1200, Springer, 1986.
- [13] G. Pisier, *The volume of convex bodies and Banach space geometry*, Cambridge University Press, 1989.
- [14] M. Rudelson, R. Vershynin, Combinatorics of random processes and sections of convex bodies, preprint.

- [15] B. Schölkopf, A.J. Smola, *Learning with kernels*, MIT Press, 2002.
- [16] N. Tomczak-Jaegermann, *Banach–Mazur distance and finite-dimensional operator Ideals*, Pitman monographs and surveys in pure and applied Mathematics 38, 1989.
- [17] A. W. Van der Vaart, J. A. Wellner, *Weak convergence and Empirical Processes*, Springer-Verlag, 1996.
- [18] V. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.