On the Limitations of Embedding Methods

Shahar Mendelson*

Abstract. We show that for any class of functions H which has a reasonable combinatorial dimension, the vast majority of small subsets of the combinatorial cube can not be represented as a Lipschitz image of a subset of H, unless the Lipschitz constant is very large. We apply this result to the case when H consists of linear functionals of norm at most one on a Hilbert space, and thus show that "most" classification problems can not be represented as a reasonable Lipschitz loss of a kernel class.

1 Introduction

The aim of this article is to investigate the limitations of embedding methods (or, as we prefer to call them here, representation methods), which are commonly used in Machine Learning. Our focus is not on the statistical side, but rather on the degree by which embedding methods can be used to approximate subsets of the combinatorial cube. To be more precise, consider a class of functions H, which we call the *base class*, defined on a metric space (Ω, d_Ω), and let ϕ be a Lipschitz function with a Lipschitz constant at most L. One can represent a subset $A \subset \{-1,1\}^n$ in H using ϕ if there are $t_1, ..., t_n \in \Omega$, such that for every $v \in A$ there is some $h_v \in H$ for which $\phi(h_v(t_j)) = v(j)$, where $v(j)$ is the j-th coordinate of v. Hence, if this is the case, we were able to represent A as a Lipschitz image (with constant at most L) of a subset of H.

In the context of Learning Theory, one should think of ϕ as a loss functional, and the question we wish to focus on is which classification problems (each problem corresponds to a subset of the combinatorial cube) can be represented in a useful manner. One could view the representation as a richness parameter of subsets of the cube. If a subset is a Lipschitz image of a subset of H (i.e. a loss class associated with a subset of H), it has to be simple.

Having this in mind, it seems likely that for a representation to be useful one needs two key ingredients. First of all, the class H has to be simple and canonical in some sense - otherwise, there is no point in using it. The second is that the Lipschitz constant of ϕ is not "too large"; if it is, the distortion caused by ϕ might make the image very rich, even if H is simple.

* Centre for Mathematics and its Applications, The Australian National University, Canberra, ACT 0200, Australia. email: shahar.mendelson@anu.edu.au

The natural example in this context is margin based classifiers. For every $\gamma > 0$, define the γ margin function (which has a Lipschitz constant of $1/\gamma$) as

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

The base class H consists of linear functionals of norm one on a Hilbert space and ϕ is generated by the desired margin. Our question in this restricted setup is as follows.

Question 1 *Set B_{ℓ_2} to be the unit ball in the Hilbert space ℓ_2. Let $A \subset \{-1,1\}^n$, $|A| = N$ (representing a classification problem), and let $\gamma > 0$. Can one find $x_1, ..., x_n \in B_{\ell_2}$ and $x_1^*, ..., x_N^* \in B_{\ell_2}$ such that for every i, j, $x_i^*(x_j) \geq \gamma$ if $a_i(j) = 1$ and $x_i^*(x_j) \leq -\gamma$ if $a_i(j) = -1$, where $a_i(j)$ is the j-th component of the i-th element of A?*

The original motivation for this study was to understand whether embedding (or kernel) techniques using the margin or a more general loss functional could serve as a generic method in Learning Theory.

Roughly speaking, in kernel methods one embeds Ω in the unit ball of a Hilbert space using the feature map, and considers the set of linear functionals in the dual unit ball as the class of functions H. It is therefore natural to ask (though, unfortunately, this question has never been studied extensively) which classification problems can be captured as a margin loss of a kernel. Of course, a small Lipschitz constant of ϕ is translated to a large margin.

The first result (at least as far as this author is aware of) that showed the limitations of kernel methods in the context of Question 1 is due to Ben-David, Eiron and Simon [2]. They proved the remarkable fact that for every n, the vast majority of subsets of $\{-1,1\}^n$ with n elements and VC dimension at most d can not be represented for a nontrivial γ in ℓ_2. To be exact, the authors showed that for a fixed d, only a vanishing fraction (at most $\sim 2^{-cn}$) of such subsets can be represented in ℓ_2 with a margin better than $1/n^\alpha$, where $\alpha = 1/2 - 1/2d - 1/2^{d-1}$. It is easy to check that $\{-1,1\}^n$ itself is represented in ℓ_2 for $\gamma = 1/\sqrt{n}$; thus, most of the small subsets of $\{-1,1\}^n$ in the sense of VC theory are not an image of a kernel class with the margin loss - unless the margin is extremely small, i.e., close to the scale at which the entire cube is represented in ℓ_2.

The basis for this result and for others of the same flavour [4, 6, 9] has to do with incompatibility of structures. On one hand, one selects H to have a simple structure. On the other, there are various notions of simplicity for subsets of the combinatorial cube. Representation methods are an attempt of imposing one structure on the other. For example, the hope that kernel methods are universal in some sense, means that every reasonable classification problem (e.g. a subset of the combinatorial cube with a small VC dimension) can be represented as a reasonable Lipschitz image of a class of linear functionals. Unfortunately, it turns out that this is impossible unless the subset of the cube has a very special structure.

Not only is it impossible to find such a linear structure in most small subsets of the cube, we show here that the situation is equally bad even if one replaces the kernel class with any other simple class H. It turns out that unless H itself contains a large "cubic" structure, (and in which case, H is no longer simple), the vast majority of small subsets of the combinatorial cube are not a reasonable Lipschitz image of a subset of H. Our aim is to find the quantitative connection between the "richness" of the set H, the Lipschitz constant of the "loss" ϕ and the number of subsets of the cube that one can reconstruct using H and ϕ with such a Lipschitz constant. The richness parameter we use for H is a variant of the combinatorial dimension, and was introduced by Pajor in [11].

Definition 1. *We say that $\{t_1, ..., t_n\}$ is ε P-shattered by H if there are sets $V_+, V_- \subset \mathbb{R}$ satisfying $d(V_+, V_-) \geq \varepsilon$, such that for every $J \subset \{1, ..., n\}$ there is $h_J \in H$ for which $h_J(t_j) \in V_+$ if $j \in J$ and $h_J(t_j) \in V_-$ otherwise. We denote by $P - VC(H, \varepsilon)$ the largest cardinality of a set which is ε P-shattered by H.*

Note that this definition extends the notion of level shattering, in which $V_+ = [\alpha + \varepsilon, \infty)$ and $V_- = (-\infty, \alpha - \varepsilon]$ for some fixed α.

Here, we denote by $VC(H, \varepsilon)$ the combinatorial dimension (also known in Learning Theory literature as the *fat-shattering* dimension) of the class H at level ε.

To compare the notion of P-shattering with the standard combinatorial dimension, let us recall the definition of packing and covering numbers, which will be required throughout this article.

Definition 2. *If (Y, d) is a metric space and $K \subset Y$, then for every $\varepsilon > 0$, $N(\varepsilon, K, d)$ is the minimal number of open balls (with respect to the metric d) needed to cover K.*

A set is ε-separated with respect to a metric d if the distance between every two distinct points in the set is larger than ε. We denote the maximal cardinality of an ε-separated subset of Y by $D(\varepsilon, Y, d)$.

It is possible to show [7] that if H is a class of functions bounded by 1 then for any probability measure μ,

$$D(\varepsilon, H, L_2(\mu)) \leq \left(\frac{2}{\varepsilon}\right)^{K \cdot VC(H, c\varepsilon)}, \tag{1.1}$$

where K and c are absolute constants.

Assume that H consists of functions bounded by 1. Then, one can verify (see the proof of Theorem 4) that if $\{t_1, ..., t_n\}$ is ε P-shattered, there is a set $H' \subset H$, $|H'| \geq 2^{cn}$ which is $\varepsilon/4$-separated in $L_2(\mu_n)$, where μ_n is the empirical measure supported on $\{t_1, ..., t_n\}$. Hence, by (1.1),

$$cn \leq \log D\left(\varepsilon/4, H, L_2\left(\mu_n\right)\right) \leq K \cdot VC(H, c'\varepsilon) \log\left(\frac{2}{\varepsilon}\right),$$

implying that for any $\varepsilon < 1$

$$P - VC(H, \varepsilon) \leq K \cdot VC(H, c\varepsilon) \log\left(\frac{2}{\varepsilon}\right),$$

for suitable absolute constants K and c.

In the reverse direction, if H is a convex and symmetric class of functions (that is, if the fact that $f \in H$ implies that $-f \in H$), then for any $\varepsilon > 0$, $VC(H, \varepsilon) \leq P - VC(H, \varepsilon)$. Indeed, in this case the combinatorial dimension can be attained by taking the fixed levels $\alpha_i = 0$ (see, e.g. [8]), and thus if a set is ε-shattered, it is also ε P-shattered.

The main result we present here connects the P-dimension of the base class H with the ability to represent "many" subsets of $\{-1,1\}^n$ as a Lipschitz image of that class, using a function with a small Lipschitz constant. The notion of representation we focus on here is rather weak, and originated from the *soft margin*.

Definition 3. *Let H be a class of functions on Ω, and set $1/2 < \delta \leq 1$. If A is a subset of $\{-1,1\}^n$, $|A| = N$, we say that A can be (L, δ) weakly represented in H if there are $x_1, ..., x_n \in \Omega$, $h_1, ...h_N \in H$ and $\phi : \mathbb{R} \to \mathbb{R}$ such that*

1. *$\|\phi\|_{lip} \leq L$, and*
2. *for every $1 \leq i \leq N$ there is a set $J_i \subset \{1, ..., n\}$ of cardinality $|J_i| \geq \delta n$, and for every i and $j \in J_i$, $\phi(h_i(x_j)) = a_i(j)$, where $a_i(j)$ is the j-th component of the i-th element in A.*

To formulate our main result, let (Ω^n, d_n) be the n product of Ω endowed with the metric

$$d_n(u, v) = \frac{1}{n} \sup_{h \in H} \sum_{i=1}^{n} |h(u_i) - h(v_i)|,$$

where $u = (u_1, ..., u_n)$, $v = (v_1, ..., v_n)$ and $u_i, v_i \in \Omega$. For every integer $N \leq 2^n$, the probability measure we use on the subsets of $\{-1,1\}^n$ of cardinality N is the counting probability measure.

Theorem 1. *There exist absolute constants k and k', and for every $1/2 < \delta \leq 1$ there are constants $c(\delta), c'(\delta), c''(\delta)$ and $n_0(\delta)$ depending only on δ for which the following holds. Let H be a class of functions on Ω which are bounded by 1. For every $L > 0$, if $n \geq n_0(\delta)$, $P - VC(H, k/L) \leq c(\delta)n$ and*

$$N \geq c(\delta) \max\left\{ \frac{k'L}{n}, \frac{\log N(c'(\delta)/L, \Omega^n, d_n)}{n} \right\},$$

then with probability at least $1 - \exp(-c''(\delta)Nn)$, a set $A \subset \{-1,1\}^n$ of cardinality N is not (L, δ) weakly represented in H.

The main novelty in Theorem 1, compared with results of a similar flavour (see, for example, [2, 4, 6, 9]), is in its nonlinear nature. Although its proof uses essentially the same ideas as in [9], what we do here goes beyond the situation where H is a class of linear functionals, which was studied in [9]. It also allows us to improve the best known estimates in what is arguably the most important case - when $H = B_{\ell_2}$.

In Section 3 we will present a detailed survey of the known estimates when $H = B_{\ell_2}$, but for now, let us formulate

Corollary 1. *Let $H = B_{\ell_2}$, considered as a set of linear functionals on $\Omega = B_{\ell_2}$. For any $1/2 < \delta \leq 1$, if $n \geq n_0(\delta)$ and $N \geq c(\delta)n$, then with probability at least $1 - \exp(-c''(\delta)Nn)$, $A \subset \{-1,1\}^n$ with $|A| = N$ is not $(c'(\delta)\sqrt{n}, \delta)$ weakly represented in H.*

To put Corollary 1 in the right perspective, $\{-1,1\}^n$ itself is represented in B_{ℓ_2} with a constant $\sqrt{n}$. And, in fact, one can use the margin function $\phi_{1/\sqrt{n}}$ for the representation. However, by Corollary 1, for any $1/2 < \delta \leq 1$ and a slightly smaller constant (which depends on δ), the vast majority of even the very small subsets of $\{-1,1\}^n$ are not weakly represented in B_{ℓ_2}.

The rest of this article is devoted to the proofs of Theorem 1 and Corollary 1. Although it is possible to find truly nonlinear applications, (for example, when H is the set of Lipschitz functions with constant 1 on the unit ball in $\mathbb{R}^d$), we decided not to present them here, as they involve routine estimates.

We end the introduction with a notational convention. Throughout, all absolute constants are denoted by c or k. Their values may change from line to line or even within the same line. $C(\varphi)$ denotes constants which depend only on the parameter φ. For a set A, let $|A|$ be its cardinality and if A, B are subsets of a vector space, put $A + B = \{a + b | a \in A,\ b \in B\}$.

2 Proof of Theorem 1

The first step in the proof of Theorem 1 is a covering argument. Here, one shows that is suffices to control a fine enough net in (Ω^n, d_n) and a finite set of Lipschitz functions.

2.1 Covering

We shall construct a finite approximating set to the set of all "meaningful" Lipschitz functions $\phi : \mathbb{R} \to \mathbb{R}$ and all possible elements $x = (x_1, ..., x_n) \in \Omega^n$ that can be used in an (L, δ) weak representation. Since H consists of functions which are bounded by 1, it is enough to consider Lipschitz functions ϕ that map $[-1,1]$ to $[-1,1]$. Indeed, if the range of ϕ exceeds $[-1,1]$, it is possible to compose it with the retraction onto $[-1,1]$ without increasing the Lipschitz constant. For every fixed L one can identify each "legal" ϕ with the pair of nonempty subsets of $[-1,1]$, $W_+ = \{t \ | \phi(t) = 1\}$ and $W_- = \{t \ | \phi(t) = -1\}$, such that $d(W_+, W_-) > 2/L \equiv \gamma$. Divide the interval $[-1,1]$ to intervals with disjoint interiors $Y_i = [a_i, b_i]$, where $b_i = a_{i+1}$, and each Y_i has length at most $\gamma/10$. One can find such decomposition with at most cL intervals Y_i for some absolute constant c.

Recall that $A + B = \{a + b : a \in A,\ b \in B\}$, and for every ϕ, define ϕ' as follows. If Y_i intersects $W_+ + (-\gamma/100, \gamma/100)$ then $\phi' = 1$ on that set, and if it intersects $W_- + (-\gamma/100, \gamma/100)$ it is -1 on that set. On the complement, which is a finite union of intervals, define ϕ' as the linear interpolation of the boundary values at each interval. Clearly, $W_+ \subset \{\phi' = 1\}$, $W_- \subset \{\phi' = -1\}$, $\|\phi'\|_{lip} < cL$,

and there are at most 3^{cL} different functions ϕ'. Denote this set of functions by Φ' and let $D_n(\varepsilon)$ be an ε cover of (Ω^n, d_n).

Lemma 1. *There exists an absolute constant k and for every $1/2 < \delta \leq 1$ there is a constant $k'(\delta)$ for which the following holds. Let $A \subset \{-1,1\}^n$ and assume that $x = (x_1, ..., x_n) \in \Omega^n$ and ϕ can be used in an (L, δ) representation of A. If δ' satisfies $\delta' - 1/2 = (\delta - 1/2)/2$, then there are $y = (y_1, ..., y_n) \in D_n(k'(\delta)/L)$ and $\phi' \in \Phi'$ which can be used to (kL, δ') weakly represent A.*

Proof. Let $\rho > 0$ be a constant which will be specified later, set ϕ' to be as above, and put y such that $d_n(x, y) \leq \rho\gamma$ for $\gamma = 2/L$. By the definition of d_n, $\sup_{h \in H} \sum_{i=1}^{n} |h(x_i) - h(y_i)| < \rho\gamma n$. Thus, for any ρ and every $h \in H$ there is a set $J_h \subset \{1, ..., n\}$, $|J_h| \geq (1 - 1000\rho)n$, such that on J_h, $|h(x_i) - h(y_i)| < \gamma/1000$. Note that by the definition of ϕ', if $h(x_i) \in W_+$ (resp, $h(x_i) \in W_-$), then $\phi'(h(y_i)) = 1$ (resp. $\phi'(h(y_i)) = -1$).

Let $h_1, ..., h_N$ be functions that can be used to represent A. Then, since the functions can be used in a (L, δ)-weak representation, it is evident that for every i, there is a set J_i, $|J_i| \geq \delta n$ for $\delta > 1/2$ such that $\phi(h_i(x_j)) = a_i(j)$. Setting δ' by $\delta' - 1/2 = (\delta - 1/2)/2$, then for ρ sufficiently small, $|J_i \cap J_{h_i}| \geq \delta' n$, and on that intersection, $\phi'(h_i(y_j)) = a_i(j)$, as claimed. ■

From Lemma 1 it is clear that it suffices to show that A is not represented using any $(\phi', y) \in \Phi' \times D_n(k'(\delta)/L)$, and there are at most $3^{kL} \cdot N(k'(\delta), \Omega^n, d_n)$ such pairs.

2.2 Controlling a "Rectangle"

A source of difficulty in the analysis of this problem stems in the "weakness" of the representation, namely, that one does not control every pair $h_i(x_j)$, but only a proportional set of indices for every $1 \leq i \leq N$. Next, we will show how to bypass this obstacle. We will show that there is a relatively small set $B \subset \{-1,1\}^n$, such that for any ϕ which has a Lipschitz constant at most L and $x \in \Omega^n$, if $A \subset \{-1,1\}^n$ is represented using (ϕ, x) then $A \subset B$. Note that the Lipschitz condition on ϕ is equivalent to having two sets, W_+ and W_- which are $2/L$ apart; on the first $\phi = 1$ and on the second $\phi = -1$.

The philosophy of the proof is as follows. Suppose that there is a "large set" $B \subset \{-1,1\}^n$, such that for every $v \in B$ there is some $h_v \in H$ for which $\phi(h_v(x_j)) = v(j)$ on δn coordinates. If this is the case, it is possible to find a large subset of B and a large subset of $\{1, ..., n\}$ on which *for every* i, j, $\phi(h_i(x_j)) = v_i(j)$. We will show that this implies that H has a large P-shattering dimension at a scale proportional to L, in contradiction to our assumption.

The combinatorial part of the proof could be described in the following way. If one views the vectors $(h_v(x_j))_{j=1}^n$ as rows in a matrix, and if in each row one can control δn of the entries for $\delta > 1/2$, then if there are enough rows in the matrix, one can find a large "rectangle", or sub-matrix on which one has complete control. The exact formulation of this claim is:

Lemma 2. *For every $1/2 < \delta \leq 1$ there exist constants α, β and n_0, all depending only on δ, for which the following holds. Assume that $n \geq n_0$, that T is an $m \times n$, $\{0,1\}$-valued matrix and that each row in T has at least δn entries that are 1. If we set $\Delta = \frac{1}{2}(\delta - \frac{1}{2})(1 - \log_2(3-2\delta)) > 0$, and if $m \geq 2^{n(1-\Delta)}$, then T contains an (s,t) sub-matrix of 1s, for $s \geq 2^{\beta n}$, $t \geq \alpha n$, and $\alpha + \beta \geq 1 + \Delta/2$.*

The proof is based on the following estimate on the so-called "problem of Zarankiewicz".

Lemma 3. *[3] 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,*

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

Proof of Lemma 2. Assume that $m > 2^{n(1-\Delta)}$ and define a bipartite graph in the following manner. One side consists of the rows of T and the other side is the elements of $\{1,...,n\}$. There is an edge between a the i-th row and $\{j\}$ if and only if $T_{i,j} = 1$. Using the notation of Lemma 3, the graph contains at least δmn edges. Hence, by Lemma 3, if s and t satisfy

$$\delta mn > (s-1)^{1/t}(n-t+1)m^{1-1/t} + (t-1)m \tag{2.1}$$

then G contains a complete (s,t) bipartite subgraph, which corresponds to T having an $s \times t$ sub-matrix of 1s. Setting $\alpha = \delta - 1/2$, $t-1 = \alpha n$ and $(s-1) = 2^{\beta n}$, an easy computation shows that

$$\beta < 1 + \alpha \log_2\left(\frac{\delta - \alpha}{1-\alpha}\right) + \frac{1}{n}\left(\log_2\left(\frac{\delta-\alpha}{1-\alpha}\right) - \Delta n\right) \tag{2.2}$$

is enough to ensure (2.1). Note that one can choose $\beta > 0$ satisfying (2.2) such that, if $n \geq n_0$, $\alpha + \beta \geq 1 + \Delta/2$, as claimed. ∎

Theorem 2. *For every $1/2 < \delta \leq 1$ there are constants $c(\delta)$ and n_0 depending only on δ, for which the following holds. Fix $n \geq n_0$ and $L > 0$, assume that $P - VC(H, 2/L) \leq c(\delta)n$ and set $\Delta = \frac{1}{2}(\delta - \frac{1}{2})(1 - \log_2(3-2\delta))$. If $x = (x_1,...,x_n) \in \Omega^n$ and ϕ is a Lipschitz function with constant at most L, there is a set $B \subset \{-1,1\}^n$, $|B| \leq 2^{n(1-\Delta)}$, such that if x and ϕ can be used to (L,δ) weakly represent A, then $A \subset B$.*

Proof. Let $c(\delta)$ be a constant which will be specified later, set n_0 to be as in Lemma 2 and assume that $P - VC(H, 2/L) \leq c(\delta)n$. Note that $v \in \{-1,1\}^n$ can be (L,δ)-weakly represented using x and ϕ if and only if there is $h_v \in H$ and $J_v \subset \{1,...,n\}$ such that $|J_v| \geq \delta n$ and for every $j \in J_v$, $\phi(h_v(x_j)) = v(j)$. Define B as the set of all such elements v, and thus, if A can be (L,δ) weakly represented using (ϕ, x) then $A \subset B$. Assume that $|B| > 2^{(1-\Delta)n}$, and define the $|B| \times n$ $\{0,1\}$-valued matrix T by $T_{i,j} = 1$, if $j \in J_{v_i}$. Applying Lemma 2 (and using its notation), T contains an (s,t) sub-matrix of 1s, where $s \geq 2^{\beta n}$,

$t \geq \alpha n$ and $\alpha + \beta \geq 1 + \Delta/2$. In other words, since $n \geq n_0$, there is a set $B' \subset B$, $|B'| \geq 2^{\beta n}$ and a set $J \subset \{1, ..., n\}$, $|J| \geq \alpha n$ such that for every $v \in B'$ there is $h_v \in H$ and for every $j \in J$, $\phi(h_v(x_j)) = v_j$.

Consider the coordinate P_J projection (restriction) of B' onto J. Since $|B'| \geq 2^{\beta n}$ and $|J| \geq \alpha n$, then $|P_J B'| \geq 2^{\beta n}/2^{n-\alpha n}$. Indeed, any point in $P_J B'$ 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 $|P_J B'| \geq 2^{n\Delta/2}$. Applying the Sauer-Shelah Lemma, there is a subset $J_1 \subset J$, such that $|J_1| \geq c(\delta)n$ and $P_{J_1} B' = P_{J_1} P_J B = \{-1,1\}^{|J_1|}$.

Hence, for every $a \in \{-1,1\}^{|J_1|}$ there is some $h_a \in H$ such that $\phi(h_a(x_j)) = a(j)$ for every $j \in J_1$. Because $d(\{\phi = 1\}, \{\phi = -1\}) \geq 2/L$, it follows that $P - VC(H, 2/L) \geq |J_1| = c(\delta)n$, which contradicts our assumption. ■

Proof of Theorem 1. By Lemma 1 (and using its notation), it suffices to show that "most" subsets of the cube are not (kL, δ') weakly represented using any element from $\Phi' \times D_n(k'(\delta)/L, \Omega^n, d_n)$. The cardinality of this product set is at most $3^{cL}|D_n(k'(\delta)/L)|$ for an absolute constant c. Now, fix such a pair (ϕ, x). By the assumption of the Theorem, $P - VC(H, 2/(kL)) \leq c(\delta)n$, where $c(\delta)$ is selected as in Theorem 2, and set $\Delta' = \frac{1}{2}(\delta' - \frac{1}{2})(1 - \log_2(3 - 2\delta'))$. If $n \geq n_0(\delta')$, then by Theorem 2 applied to (kL, δ'), there is a set $B \subset \{-1,1\}^n$ of cardinality $|B| \leq 2^{n(1-\Delta')}$, such that if x and ϕ can be used to (kL, δ') weakly represent A, then $A \subset B$.

Clearly, the probability that a random point $v \in \{-1,1\}^n$ belongs to B is at most $|B|/2^n = 2^{-n\Delta'}$, and thus, if $|A| = N$, the probability that $A \subset B$ is at most $2^{-nN\Delta'}$. Therefore, if $3^{cL}|D_n(k'(\delta)/L)| \leq 2^{-nN\Delta'/2}$, it follows that with probability at least $1 - \exp(c'(\delta)Nn)$, A is not (L, δ) weakly represented in H. ■

3 Application: $H = B_{\ell_2}$

A natural base class which one should consider, and which was studied in [9], is $H = B_{X^*}$ - the dual unit ball of some n-dimensional Banach space X, acting as functionals on $\Omega = B_X$. Although we do not wish to focus on this general case, as it requires some knowledge in convex geometry, let us point out several relatively easy observations. Since H consists of Lipschitz functions of norm 1 then

$$d_n(u, v) \leq \max_{1 \leq i \leq N} \sup_{h \in H} |h(u_i) - h(v_i)| \leq \max_{1 \leq i \leq N} d_\Omega(u_i, v_i).$$

Therefore,

$$N(\varepsilon, \Omega^n, d_n) \leq (N(\varepsilon, \Omega, d_\Omega))^n, \tag{3.1}$$

and for $H = B_{X^*}$, $d_\Omega(u, v) = \|u - v\|_X$. Moreover, if X is an n dimensional Banach space then by a standard volumetric estimate (see, e.g. [12]), $N(\varepsilon, \Omega, d_\Omega) \leq (3/\varepsilon)^n$, implying that

$$\frac{\log N(\varepsilon, \Omega^n, d_n)}{n} \leq cn \log(c'/\varepsilon).$$

As mentioned in the introduction, for every class of functions bounded by 1,

$$P - VC(H, \varepsilon) \leq K \cdot VC(H, c\varepsilon) \log(2/\varepsilon).$$

Hence, as long as $VC(H, 2/(kL)) \log(2L) \leq c(\delta)n$, the assumption of Theorem 1 holds. It turns out that up to a $\log(n)$ factor, the "critical level" L for which this assumption is still valid is determined by a notion of distance between Banach spaces. The critical L is proportional to the so-called *Banach-Mazur distance* between X and ℓ_1^n (we refer the reader to [9] for more details). On the other hand, if L is the distance between X and ℓ_1^n, then the entire cube $\{-1, 1\}^n$ is $(L, 1)$ represented in $H = B_{X^*}$. Thus, the situation one often encounters when H is the unit ball of the dual to an n dimensional Banach space is a surprising dichotomy. For $L = d(X, \ell_1^n)$, the entire cube, and thus all its subsets can be represented in H. For a slightly smaller constant, $c(\delta)L$, the vast majority of subsets of cardinality $N \approx c'(\delta)n \log n$ are not even $(c(\delta)L, \delta)$ weakly represented in H.

The case of $H = B_{\ell_2}$ has been studied, in one form on another, by several authors. A careful examination of the proof in [2] shows that only a vanishing fraction of the subsets of $\{-1, 1\}^n$ with N elements is represented in ℓ_2 with a margin better than $c\sqrt{(\log N)/n}$ for suitable absolute constant c, as long as $N/n^2 \to \infty$. This implies that, at least when ϕ is taken from the margin function family, and as long as $L \leq c\sqrt{n/\log N}$ and $N \geq cn^2$, most of the subsets of $\{-1, 1\}^n$ are not $(L, 1)$ weakly represented in B_{ℓ_2}.

A different approach, based on operator ideal theory, was used in [6] to prove that if $N \geq cn$, then with probability at least $1 - \exp(cN)$, a subset of $\{-1, 1\}^n$ with N elements is only represented in ℓ_2 with the trivial margin of $c_1/\sqrt{n}$; in other words, it improves [2] in the way N depends on n and because the restriction on L is the optimal one - $L \leq c\sqrt{n}$. However, it too only applies when the Lipschitz function is taken from the margin family.

These two results are limited since they are completely Hilbertian in nature. They do not extend to the case $H = B_{X^*}$ for a non-Hilbert space X, let alone to when H is not a class of linear functionals.

In [9], the method of proof (which is essentially the same as the proof of Theorem 1) enables one to deal with weak representation by an arbitrary Lipschitz function, and to treat $H = B_{X^*}$ for a general n-dimensional Banach space. For $H = B_{\ell_2}$ it was shown that if $N \geq c(\delta)n \log n$ then with probability at least $1 - \exp(c'(\delta)Nn)$ a subset of $\{-1, 1\}^n$ of cardinality N is not (L, δ)-weakly represented in B_{ℓ_2} if $L \leq c''(\delta)\sqrt{n}$. The price paid for the extension to an arbitrary Lipschitz function was that N was no longer linear in n. The main result of this section is to remove this parasitic logarithmic factor.

Although the analysis we present for B_{ℓ_2} goes beyond the Hilbertian case, it still only works under additional structural assumptions on the space X. And, though under such assumptions it is possible to remove the parasitic logarithmic factor, doing the same in the general case seems (at least to this author) a worthwhile challenge.

Theorem 3. *For every $1/2 < \delta \leq 1$, there exist constants $c(\delta)$, $c'(\delta)$, $c''(\delta)$ and $n_0(\delta)$, depending only on δ, for which the following holds. For every integer $n \geq n_0$, if $L \leq c(\delta)\sqrt{n}$ and $N \geq c'(\delta)n$, then with probability $1 - \exp(c''(\delta)nN)$, a set with N elements is not (L, δ) weakly represented in B_{ℓ_2}.*

Clearly, because of the structure of ℓ_2, it suffices to consider the n-dimensional Euclidean space ℓ_2^n, rather than the infinite dimensional one. Thus, $H = B_2^n$, consisting of linear functionals on $\Omega = B_2^n$.

The proof of Theorem 3 requires two preliminary steps before one can use Theorem 1. First of all, one has to identify the critical level at which $P - VC(B_{\ell_2}, \varepsilon) \leq c(\delta)n$ for the constant $c(\delta)$ appearing in Theorem 1. Then, one has to estimate $N\left(\varepsilon, \left(B_2^n\right)^n, d_n\right)$.

Lemma 4. *There exists an absolute constant c such that for every $0 < \varepsilon < 1$,*

$$P - VC(\varepsilon, B_{\ell_2}) \leq \frac{c}{\varepsilon^2}.$$

Let us mention that the same estimate holds true for the combinatorial dimension (see, e.g [8]), and thus, for this class of functions, the two dimensions are equivalent.

The proof of Lemma 4 is based on Sudakov's minoration (see, for example, [5, 12]).

Lemma 5. *There exists an absolute constant c for which the following holds. If $T \subset \ell_2^n$ then*

$$c \sup_{\varepsilon > 0} \varepsilon \sqrt{\log N(\varepsilon, T, \ell_2^n)} \leq \mathbb{E} \sup_{t \in T} \left| \sum_{i=1}^n g_i t_i \right|,$$

where $(g_i)_{i=1}^n$ are independent, standard gaussian random variables and $t = (t_1, ..., t_n)$.

Note that if μ_n is the empirical measure on $\{1, ..., n\}$ and if one views each $t \in \ell_2^n$ as a function on $\{1, ..., n\}$ in the natural way, then $\|t\|_{\ell_2^n} = \sqrt{n}\|t\|_{L_2(\mu_n)}$. Thus, by Lemma 5,

$$\sup_{\varepsilon > 0} \varepsilon \sqrt{\log N\left(\varepsilon, T, L_2(\mu_n)\right)} \leq \frac{C}{\sqrt{n}} \mathbb{E} \sup_{t \in T} \left| \sum_{i=1}^n g_i t_i \right|. \tag{3.2}$$

Proof of Lemma 4. Assume that $\{x_1, ..., x_n\} \in B_{\ell_2}$ is ε P-shattered by B_{ℓ_2}. Then, there is a set $H' \subset H$, $|H'| \geq 2^{cn}$ which is $\varepsilon/4$-separated in $L_2(\mu_n)$, where μ_n is the empirical measure supported on $\{x_1, ..., x_n\}$. Indeed, each $h \in B_{\ell_2}$ can be associated with a point in $\{-1, 1\}^n$ according to whether $h(x_i) \in V_+$ or $h(x_i) \in V_-$. By a standard probabilistic argument, there is a subset of $\{-1, 1\}^n$ of cardinality 2^{cn} which is $n/4$ separated in the Hamming metric. Consider the elements in H that correspond to the separated set and let h, h' be two such elements. Thus, there is a set $I \subset \{1, ..., n\}$ of cardinality at least $n/4$ such that for every $i \in I$, if $h(x_i) \in V_+$ then $h'(x_i) \in V_-$ and vice-versa. Therefore, $\sum_{i=1}^n |h(x_i) - h'(x_i)| \geq \sum_{i \in I} |h(x_i) - h'(x_i)| \geq |I|\varepsilon$.

Let $(g_i)_{i=1}^n$ be standard independent gaussian variables. By (3.2), the fact that $\|x\|_{\ell_2} = \sup_{h\in B_{\ell_2}} h(x)$ and a standard estimate on $\mathbb{E}\|\sum_{i=1}^n g_i x_i\|_{\ell_2}$,

$$c\varepsilon\sqrt{n} \le c\sup_{\delta>0}\delta\sqrt{\log N\left(\delta, H, L_2(\mu_n)\right)} \le \frac{1}{\sqrt{n}}\mathbb{E}_g \sup_{h\in B_{\ell_2}} \sum_{i=1}^n g_i h(x_i)$$
$$\le \frac{1}{\sqrt{n}}\mathbb{E}_g \left\|\sum_{i=1}^n g_i x_i\right\|_{\ell_2} \le 1.$$

Therefore, $n \le c/\varepsilon^2$, as claimed. ∎

To conclude the proof of Theorem 3, it remains to bound $N(\varepsilon, \Omega^n, d_n)$, and, as we already mentioned, one can take $\Omega = B_2^n$. Note that the "easy" way to upper-bound $N(\varepsilon, (B_2^n)^n, d_n)$, as presented in (3.1), leads to the superfluous $\log n$ factor, and thus a different argument is required.

Theorem 4. *There exists an absolute constant c such that for any integer n and any $\varepsilon \ge c/\sqrt{n}$, $D(\varepsilon, (B_2^n)^n, d_n) \le 2^{n^2+1}$, and thus, $N(\varepsilon, (B_2^n)^n, d_n) \le 2^{n^2+1}$.*

Before presenting the proof let us introduce the following notation. For two sets $A, B \subset \mathbb{R}^m$ let $D(A,B)$ be the maximal number of points $a_i \in A$ such that the sets $a_i + B$ are disjoint. Observe that if B is a ball of radius ε with respect to a norm $\|\ \|_X$ then $D(A,B)$ is the maximal cardinality of an ε separated subset of A with respect to d_X.

Proof of Theorem 4. Since B_2^n consists of linear functionals, then for every $u, v \in (B_2^n)^n$, $d_n(u,v) = \frac{1}{n}\sup_{h\in B_2^n}\sum_{i=1}^n |h(u_i - v_i)|$. In fact, this metric is induced by a norm on the product space $\Pi_{i=1}^n \mathbb{R}^n$,

$$\|(x_1, ..., x_n)\| = \frac{1}{n}\sup_{h\in B_2^n}\sum_{i=1}^n |h(x_i)|.$$

Consider the unit ball of this norm, which we denote by $\mathcal{K}$. Fix $\varepsilon > 0$, and observe that our aim is to find the maximal number of disjoint translates of $\varepsilon\mathcal{K}$ that are centered at points in $\mathcal{B} = \Pi_{i=1}^n B_2^n$. To that end, we use a well known volumetric argument, which we present for the sake of completeness.

Let $\mathcal{U} = \varepsilon\mathcal{K}\cap\mathcal{B}$ which is also a convex, symmetric set, and clearly, $D(\mathcal{B}, \varepsilon\mathcal{K}) \le D(\mathcal{B},\mathcal{U})$. Let $y_1, ..., y_m$ be elements in $\mathcal{B}$ such that for every $i \ne j$, $y_i + \mathcal{U}$ and $y_j + \mathcal{U}$ are disjoint. Since $\mathcal{U} \subset \mathcal{B}$ then

$$\bigcup_{i=1}^m (y_i + \mathcal{U}) \subset 2\mathcal{B}.$$

Let $\mathrm{vol}(A)$ be the Lebesgue measure of $A \subset \Pi_{i=1}^n \mathbb{R}^n$. Since the sets $y_i + \mathcal{U}$ are disjoint, then $\sum_{i=1}^m \mathrm{vol}(y_i + \mathcal{U}) \le \mathrm{vol}(2\mathcal{B}) = 2^{n^2}\mathrm{vol}(\mathcal{B})$, and thus $m \le 2^{n^2}\mathrm{vol}(\mathcal{B})/\mathrm{vol}(\mathcal{U})$. To conclude the proof it is enough to show that as long as $\varepsilon \ge c/\sqrt{n}$, $\mathrm{vol}(\mathcal{B})/\mathrm{vol}(\mathcal{U}) \le 2$.

Let μ be the normalized volume measure on B_2^n, and set μ^n to be the product measure on $\mathcal{B}$. Therefore, if X is a random vector distributed according to μ, and if $X_1, ..., X_n$ are independent copies of X, then

$$\mathrm{vol}(\mathcal{U}) = \mathrm{vol}(\mathcal{B}) \cdot Pr\left((X_1, ..., X_n) \in \mathcal{U}\right).$$

Since $X_i \in B_2^n$, then

$$Pr\left((X_1, ..., X_n) \in \mathcal{U}\right) =$$
$$Pr\left((X_1, ..., X_n) \in \varepsilon\mathcal{K}\right) = Pr\left(\frac{1}{n} \sup_{h \in B_2^n} \sum_{i=1}^n |h(X_i)| \leq \varepsilon\right),$$

and to estimate this probability we can use the uniform law of large numbers.

Note that for every $h \in B_2^n$, $c_1\sqrt{n} \leq \mathbb{E}|h(X)| \leq c_2\sqrt{n}$ for suitable absolute constants c_1 and c_2; this could be verified by a tedious computation, or by a representation of the volume measure on B_2^n in terms of gaussian random variables (see, for example, [12, 10] for the basic facts and [1] for representations of the volume measure on the unit balls of other ℓ_p^n spaces).

Thus, as long as $\varepsilon \geq c/\sqrt{n}$ for an appropriate $c > 0$, it suffices to estimate

$$Pr\left(\frac{1}{n} \sup_{h \in B_2^n} \left|\sum_{i=1}^n |h(X_i)| - \mathbb{E}|h(X)|\right| \geq \frac{c'}{\sqrt{n}}\right)$$

and to show that for a large enough c', this probability is smaller than $1/2$. And indeed, by a symmetrization argument and the contraction principle for the absolute value function (see, e.g. [5]),

$$\mathbb{E} \sup_{h \in B_2^n} \left|\frac{1}{n}\sum_{i=1}^n |h(X_i)| - \mathbb{E}|h(X)|\right| \leq \frac{2}{n}\mathbb{E} \sup_{h \in B_2^n} \left|\sum_{i=1}^n \varepsilon_i h(X_i)\right| = \frac{2}{n}\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i X_i\right\|_{\ell_2}$$
$$\leq \frac{2}{n}\mathbb{E}\left(\sum_{i=1}^n \|X_i\|_2^2\right)^{1/2} \leq \frac{2}{\sqrt{n}},$$

and the claim follows from Chebyshev's inequality. ∎

4 Concluding Remarks

Despite several attempts to apply the method developed here to the case studied in [2], as of yet we were not able to find a unified approach to resolve both questions. The reason for that stems in the different subsets of the cube that one considers. Here, we focus on random sets with $c(\delta)n$ elements, while in [2] the authors considered subsets of cardinality n but with VC dimension at most d. Since a "typical" subset of $\{-1, 1\}^n$ with n elements has VC dimension of the order $\log n$, the sets studied in [2] can not be reached using the counting

measure we use. One possible way forward could be to find a representation of the counting probability measure on the set of subsets of $\{-1,1\}^n$ with VC dimension at most d, and to combined it with Theorem 2, though finding such a representation seems highly nontrivial.

The main point of this note belongs to the "*no free lunch*" philosophy. It is natural to assume that there are no simple, universal classes of functions, which is what could be seen here. The only way one can reach most of the small subsets of the cube is if the base class itself is so rich, that it is pointless to use it in a representation.

Of course, the result presented here, much like that ones it followed, does not imply that embedding/represetation type methods are useless, as real world problems most likely do not correspond to "typical" subsets of the combinatorial cube. They have a special symmetry or a geometric structure that makes them easier to handle. Our goal should be to find the significant symmetries and to exploit them by matching the correct H to the given learning problem. The main objective in writing this article was to point out to this important, yet under-studied problem: find out which classification problems (subsets of the cube) can be represented, and by which base classes H; that is, for each H, find the subsets of the cube that have a compatible structure with that of H, and vice-versa.

References

1. F. Barthe, O. Guédon, S. Mendelson, A. Naor, A probabilistic approach to the geometry of the ℓ_p^n ball, Annals of Probability, 33 (2) 480-513, 2005.
2. 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.
3. B. Bollobás, *Extremal graph theory*, Academic Press, 1978.
4. J. Forster, N. Schmitt, and H.U. Simon, Estimating the optimal margins of embeddings in Euclidean halfspaces, in *Proceedings of the 14th Annual Conference on Computational Learning Theory, 2001*, LNCS volume 2111, 402-415. Springer, Berlin, 2001.
5. M. Ledoux, M. Talagrand, *Probability in Banach spaces*, Springer, 1991.
6. N. Linial, S. Mendelson, G. Schechtman, A. Shraibman, Complexity measures of sign matrices, preprint.
7. S. Mendelson, R. Vershynin, Entropy and the combinatorial dimension, Inventiones Mathematicae, 152(1), 37-55, 2003.
8. S. Mendelson, G. Schechtman, The shattering dimension of sets of linear functionals, Annals of Probability, 32 (3A), 1746-1770, 2004.
9. S. Mendelson, Embedding with a Lipschitz function, Random Structures and Algorithms, to appear (available on the journal's web-page).
10. V.D. Milman, G. Schechtman, *Asymptotic theory of finite dimensional normed spaces*, Lecture Notes in Mathematics 1200, Springer, 1986.
11. A. Pajor, *Sous espaces ℓ_1^n des espaces de Banach*, Hermann, Paris, 1985.
12. G. Pisier, *The volume of convex bodies and Banach space geometry*, Cambridge University Press, 1989.
13. A. W. Van der Vaart, J. A. Wellner, *Weak convergence and Empirical Processes*, Springer-Verlag, 1996.