PROCEEDINGS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 135, Number 5, May 2007, Pages 1455–1463
S 0002-9939(06)08620-5
Article electronically published on November 14, 2006

LIPSCHITZ REPRESENTATIONS OF SUBSETS OF THE CUBE

SHAHAR MENDELSON

(Communicated by N. Tomczak-Jaegermann)

ABSTRACT. We show that for any class of uniformly bounded functions H with a reasonable combinatorial dimension, the vast majority of small subsets of the n-dimensional combinatorial cube cannot 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.

1. INTRODUCTION

The aim of this article is to investigate the following question: consider a class of real-valued functions H, defined on a set Ω. We say that a set $A \subset \{-1,1\}^n$ is L-represented in H if there is a Lipschitz function $\phi : \mathbb{R} \to \mathbb{R}$ with a Lipschitz constant at most L and $t_1, ..., t_n \in \Omega$, such that for every $a \in A$ there is some $h_a \in H$ for which $\phi(h_a(t_j)) = a(j)$, where $a(j)$ is the j-th coordinate of a. Hence, A is a Lipschitz image with a constant at most L of a coordinate projection of a subset of H, and thus, the set A, when considered as a class of functions on $\{1, ..., n\}$, can be found in some sense in H. This notion of representation originated in Statistical Learning Theory (see, e.g., [1, 3, 5, 6] and references therein for more details on its significance).

A natural example which comes to mind is when Ω is the unit ball in the Hilbert space ℓ_2, H is the set of linear functionals of norm at most 1 on ℓ_2 and ϕ belongs to the margin family, that is, for every $\gamma > 0$, ϕ_γ is defined as

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

Observe that for every γ, ϕ_γ is a Lipschitz function with a constant of $1/\gamma$.

The question at hand in this restricted setup is as follows.

Question 1. Denote by B_{ℓ_2} the unit ball in ℓ_2, consider $A \subset \{-1,1\}^n$ of cardinality $|A| = N$ 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)?

Received by the editors April 29, 2005 and, in revised form, December 20, 2005.
2000 *Mathematics Subject Classification.* Primary 46B07, 60D05.
The author was supported in part by an Australian Research council Discovery grant.

From a geometric point of view, each point $a \in A$ defines a set $\sigma_a \subset \{1,...,n\}$ by $\sigma_a = \{i : a(i) = 1\}$, and one is looking for $x_1,...,x_N \in B_{\ell_2}$ such that for every $a \in A$,

$$d\left(\mathrm{conv}\left(x_i : i \in \sigma_a\right), \mathrm{conv}\left(x_i : i \notin \sigma_a\right)\right) \geq 2\gamma,$$

where the metric d is endowed with the ℓ_2 norm.

The first result in the direction of Question 1 is due to Ben-David, Eiron and Simon [1]. To formulate it, recall that for $A \subset \{-1,1\}^n$,

$$VC(A) = \sup\left\{|\sigma| \ : \ P_\sigma A = \{-1,1\}^{|\sigma|}\right\},$$

where $\sigma \subset \{1,...,n\}$ and $P_\sigma A = \left\{(a_i)_{i\in\sigma} : a \in A\right\}$ is the coordinate projection of A onto σ. In [1] it was shown that for every n and for a fixed d, only a vanishing fraction (at most $\sim 2^{-cn}$) of subsets of $\{-1,1\}^n$ with n elements and VC dimension at most d can be represented in B_{ℓ_2} using a function from the margin family with a Lipschitz constant smaller than n^α, $\alpha = 1/2 - 1/(2d) - 1/2^{d-1}$. It is easy to check that $\{-1,1\}^n$ itself is represented in B_{ℓ_2} using $\phi_{1/\sqrt{n}}$, which has a Lipschitz constant $\sqrt{n}$. Thus, most of the small subsets of $\{-1,1\}^n$ in the sense of VC theory are not an image of a coordinate projection of B_{ℓ_2} using some function ϕ_γ, unless $L = 1/\gamma$ is extremely large, i.e., close to the scale at which the entire cube is represented in B_{ℓ_2}.

Here, we take a different route; we show that it is impossible to represent in H with a small Lipschitz constant many subsets of small cardinality of $\{-1,1\}^n$, unless H itself contains a large "cubic" structure. The notion of a cubic structure we use here is a variant of the combinatorial dimension and was introduced by Pajor in [8].

Definition 1.1. We say that $\{t_1,...,t_n\} \in \Omega$ 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 $PVC(H,\varepsilon)$ the largest cardinality of a subset of Ω which is ε P-shattered by H.

The notion of representation we focus on here is a weaker, "isomorphic" version of Lipschitz representations, rather that the "isometric" one defined above.

Definition 1.2. Let H be a class of real-valued functions on Ω and set $1/2 < \delta \leq 1$. If $A \subset \{-1,1\}^n$, $|A| = N$, we say that A can be (L,δ) 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) $\|\phi\|_{\mathrm{lip}} \leq L$, 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 $j \in J_i$, $\phi\left(h_i\left(x_j\right)\right) = a_i(j)$.

In other words, rather than controlling $\phi(h_i(x_j))$ for every i,j (which is the "isometric" case), here, for every i one only controls the values of h_i on a subset of proportional cardinality of $\{x_1,...,x_n\}$, and this set can change with i.

Before formulating our main result (Theorem 1.4) we require several preliminary definitions.

Let Ω^n be the n product of Ω endowed with the pseudometric $d_n(u,v) = \frac{1}{n}\sup_{h\in H}\sum_{i=1}^n |h(u_i) - h(v_i)|$, where $u = (u_i)_{i=1}^n$, $v = (v_i)_{i=1}^n$ and $u_i, v_i \in \Omega$. For the sake of simplicity we assume that d_n is a metric (which is the case, for example, if H separates points in Ω), though all our claims remain true without this assumption.

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.

Definition 1.3. 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) of radius ε 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 easy to verify that for every $\varepsilon > 0$, $N(\varepsilon, K, d) \leq D(\varepsilon, K, d) \leq N(\varepsilon/2, K, d)$.

Theorem 1.4. *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 $n \geq n_0(\delta)$, $PVC(H, k/L) \leq c(\delta)n$ 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) represented in H.

Observe that if $PVC(H,\varepsilon) = n$, then the entire cube $\{-1,1\}^n$ is $(2/\varepsilon, 1)$ represented in H. Theorem 1.4 implies that if the P-combinatorial dimension at scale ε is slightly smaller than n, the vast majority of "small" subsets of $\{-1,1\}^n$ are not c/ε represented in H, even in the weak sense.

Example 1.5. Assume that (Ω, d) is a metric space and that H consists of Lipschitz functions with constant at most 1. Clearly, one can trivially bound $d_n(u,v) \leq \max_{1\leq i\leq n} d(u_i, v_i)$ and thus, $\log N(\varepsilon, \Omega^n, d_n) \leq n \log N(\varepsilon, \Omega, d)$. Therefore, by Theorem 1.4, if $N \geq c(\delta) \log N(c'(\delta)/L, \Omega, d)$ and as long as H is not "too large" in the PVC sense, most subsets of $\{-1,1\}^n$ of cardinality N are not (L,δ) represented in H. This is the case, for example, if Ω is a unit ball in an n-dimensional Banach space and H is the unit ball in the dual space. Indeed, a standard volumetric estimate [9] shows that $\log N(c'(\delta)/L, \Omega^n, d_n) \leq n \log(L/c'(\delta))$, though this estimate is often suboptimal and leads to a logarithmic looseness (see Theorem 2.7).

The novelty in Theorem 1.4, compared with results of a similar flavor (see, for example, [1, 3, 5, 6]), is in its nonlinear nature. All previous results (with the exception of [6]) deal only with the case of $H = B_{\ell_2}$ and assume that ϕ is taken from the margin family. The more general notion of representation was introduced in [6], and although the proof of Theorem 1.4 uses essentially the same ideas as in [6], what we do here goes beyond the situation of H being the dual unit ball of an n-dimensional normed space, which was the main theme there. In addition, using the metric d_n allows us to improve the best known estimates in what is arguably the most important case—when $H = B_{\ell_2}$.

In what follows, we will present a detailed survey of the known estimates in the case $H = B_{\ell_2}$, but for now, let us formulate

Theorem 1.6. *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 $(c'(\delta)\sqrt{n}, \delta)$ represented in H.*

To put Theorem 1.6 in the right perspective, $\{-1,1\}^n$ itself is $(\sqrt{n},1)$ represented in B_{ℓ_2}, and, in fact, one can use the margin function $\phi_{1/\sqrt{n}}$ for the representation. However, by Theorem 1.6, 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 1.4 and Theorem 1.6.

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. Proofs

The first step in the proof of Theorem 1.4 is a covering argument. Here, one shows that it suffices to control a fine enough net in (Ω^n, d_n) and a finite set of Lipschitz functions.

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) representation. Since H consists of functions which are bounded by 1, it is enough to consider Lipschitz functions that map $[-1,1]$ to $[-1,1]$.

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 $[-1,1]$ into 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 a decomposition with at most cL intervals Y_i for some absolute constant c. Now, for every ϕ, define ϕ' as follows. If Y_i intersects $W_+ + (-\gamma/100, \gamma/100)$, then $\phi' = 1$ on Y_i, and if it intersects $W_- + (-\gamma/100, \gamma/100)$ it is -1 on Y_i. On the complement, which is a finite union of intervals, define ϕ' as the linear interpolation of the boundary values at each interval. Clearly, $W_+ \subset \{\phi' = 1\}$, $W_- \subset \{\phi' = -1\}$, $\|\phi'\|_{\text{lip}} < cL$, 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 2.1. *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') represent A.*

Proof. Let $0 < \rho < 1/1000$ be a constant which will be determined later, set ϕ' to be as above, and select 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 are used in the representation of A. It is evident that for every i, there is a set J_i, $|J_i| \geq \delta n$ such that $\phi(h_i(x_j)) = a_i(j)$. Recall that $\delta > 1/2$ and set $\delta' = 1/2 + (\delta - 1/2)/2$. It follows that 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. $\square$

By Lemma 2.1 it suffices to show that A is not (L,δ) represented using any $(\phi',y) \in \Phi' \times D_n(k'(\delta)/L)$, and there are at most $3^{kL} \cdot N(k'(\delta)/L, \Omega^n, d_n)$ such pairs.

The next step in the proof is to deal with the difficulty that stems from 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 $1 \leq i \leq N$.

To that end, consider the vectors $(h_i(x_j))_{j=1}^n$ as rows of a matrix. If in each row one can control δn of the entries for $\delta > 1/2$ and if there are enough rows in the matrix, then one can find a large "rectangle", or submatrix on which one controls all the entries. The exact formulation of this claim is:

Lemma 2.2. *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 a submatrix of 1s with s rows and t columns, for $s \geq 2^{\beta n}$, $t \geq \alpha n$, and $\alpha + \beta \geq 1 + \Delta/2$.*

The proof of this statement can be found in [6] and follows easily from an estimate on the "problem of Zarankiewicz" [2].

Using this combinatorial lemma, we can show that for any ϕ which has a Lipschitz constant at most L and $x \in \Omega^n$, there is a relatively small set B which contains all the sets $A \subset \{-1,1\}^n$ that are (L,δ) represented using (ϕ, x).

Theorem 2.3. *For every $1/2 < \delta \leq 1$ there are constants $c(\delta)$ and $n_0(\delta)$ depending only on δ, for which the following holds. Fix $n \geq n_0$ and $L > 0$, assume that $PVC(H, 2/L) \leq c(\delta)n$ and set $\Delta = \frac{1}{2}(\delta - \frac{1}{2})(1 - \log_2(3-2\delta))$. Let $x = (x_1, ..., x_n) \in \Omega^n$ and set ϕ to be a Lipschitz function with constant at most L. Then, there is a set $B \subset \{-1,1\}^n$, $|B| \leq 2^{n(1-\Delta)}$, such that if $A \subset \{-1,1\}^n$, and if x and ϕ can be used to (L,δ) 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 2.2 and assume that $PVC(H, 2/L) \leq c(\delta)n$. Note that $v \in \{-1,1\}^n$ can be (L,δ) represented using x and ϕ if and only if there are $h_v \in H$ 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) 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 2.2 (and using its notation), T contains a submatrix of 1s with s rows and t columns, 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 $h_v \in H$ which satisfies that for every $j \in J$, $\phi(h_v(x_j)) = v_j$.

Consider the coordinate P_J projection 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 (see, e.g. [4]), there is a subset $J_1 \subset J$ of cardinality $|J_1| \geq c(\delta)n$, for which $P_{J_1} B' = P_{J_1} P_J B = \{-1,1\}^{|J_1|}$.

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

Proof of Theorem 1.4. By Lemma 2.1 (and using its notation), it suffices to show that "most" subsets of the cube are not (kL, δ') 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, $PVC(H, 2/(kL)) \leq c(\delta')n$, where $c(\delta')$ is selected as in Theorem 2.3, 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 2.3 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') 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) represented in H. □

2.1. **Application: $H = B_{\ell_2}$.** The case of $H = B_{\ell_2}$ has been studied, in one form or another, by several authors. A careful examination of the proof in [1] shows that only a vanishing fraction of the subsets of $\{-1, 1\}^n$ with N elements is $(L, 1)$ represented in B_{ℓ_2} using a function from the margin family with a Lipschitz constant smaller than $c\sqrt{n/\log N}$ and as long as $N/n^2 \to \infty$.

A different approach, based on operator ideal theory, was used in [5] to prove that if $N \geq cn$, then with probability at least $1 - \exp(-cN)$, a subset of $\{-1, 1\}^n$ with N elements can only be represented in B_{ℓ_2} with the trivial constant of $c_1\sqrt{n}$, again, under the assumption that ϕ belongs to the margin family. In other words, it improves [1] in the way N depends on n and because the restriction on L is the optimal one, namely, $L \leq c\sqrt{n}$. However, this estimate too only applies when the Lipschitz function is taken from the margin family, and was not stated with a "weak" analog.

These two results are limited since they are completely Hilbertian in nature. They do not extend to the case where H is the dual unit ball of a non-Hilbert space X, let alone to cases where H is not a class of linear functionals.

In [6], the method of proof (which is essentially the same as the proof of Theorem 1.4) enables one to deal with weak representations by an arbitrary Lipschitz function and to treat the case where H is the dual unit ball of 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) represented in B_{ℓ_2} if $L \leq c''(\delta)\sqrt{n}$. The price paid for this extension, that is, going from a function in the margin family to an arbitrary Lipschitz function, was that N was no longer linear in n. Our next goal is to show how to remove this parasitic logarithmic factor.

Let us recall the result we wish to prove here:

Theorem 2.4. *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 subset of $\{-1, 1\}^n$ with N elements is not (L, δ) represented in B_{ℓ_2}.*

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, $H = B_2^n$, when viewed as a set of linear functionals on $\Omega = B_2^n$.

The proof of Theorem 2.4 requires two preliminary steps before one can use Theorem 1.4. First of all, one has to identify the critical level at which $PVC(B_{\ell_2}, \varepsilon) \leq$

$c(\delta)n$ for the constant $c(\delta)$ appearing in Theorem 1.4. Then, one has to estimate $N\left(\varepsilon,(B_2^n)^n,d_n\right)$ at an appropriate level. The proof of Theorem 2.4 follows immediately from Theorem 1.4 and these two estimates.

Lemma 2.5. *For every $0<\varepsilon<1$, $PVC(\varepsilon,B_{\ell_2})\leq c/\varepsilon^2$, where c is an absolute constant.*

The proof of Lemma 2.5 is based on Sudakov's inequality (see, for example, [4, 9]).

Lemma 2.6. *There exists an absolute constant c for which the following holds. If $T\subset\ell_2^n$, then $c\sup_{\varepsilon>0}\varepsilon\sqrt{\log N(\varepsilon,T,\ell_2^n)}\leq\mathbb{E}\sup_{t\in T}|\sum_{i=1}^n g_it_i|$, 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 2.6, $\sup_{\varepsilon>0}\varepsilon\sqrt{\log N\left(\varepsilon,T,L_2(\mu_n)\right)}\leq\frac{C}{\sqrt{n}}\mathbb{E}\sup_{t\in T}|\sum_{i=1}^n g_it_i|$.

Proof of Lemma 2.5. 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 $h\in B_{\ell_2}$ taking part in the shattering 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 elements in that 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, $\sum_{i=1}^n|h(x_i)-h'(x_i)|^2\geq\sum_{i\in I}|h(x_i)-h'(x_i)|^2\geq|I|\varepsilon^2$.

Let $(g_i)_{i=1}^n$ be standard independent Gaussian variables. Using that $\|x\|_{\ell_2}=\sup_{h\in B_{\ell_2}}h(x)$ and by a standard estimate on $\mathbb{E}\|\sum_{i=1}^n g_ix_i\|_{\ell_2}$, it follows that

$$c\varepsilon\sqrt{n}\leq\sup_{\delta>0}\delta\sqrt{\log N\left(\delta,H,L_2(\mu_n)\right)}\leq\frac{c'}{\sqrt{n}}\mathbb{E}_g\left\|\sum_{i=1}^n g_ix_i\right\|_{\ell_2}\leq c'.$$

Therefore, $n\leq c/\varepsilon^2$, as claimed. □

To conclude the proof of Theorem 2.4, it remains to bound $\log N(\varepsilon,\Omega^n,d_n)$, and, as we already mentioned, one can take $\Omega=B_2^n$. Note that the "easy" way to upper-bound $N(\varepsilon,(B_2^n)^n,d_n)$, using the n product of an ε net in B_2^n with respect to the Euclidean norm, leads to the superfluous $\log n$ factor for $\varepsilon=c/L\sim 1/\sqrt{n}$, which is precisely the looseness we wish to remove. Thus, a different argument is required.

Theorem 2.7. *There exists an absolute constant c such that for any integer n and any $\varepsilon\geq c/\sqrt{n}$, $D(\varepsilon,(B_2^n)^n,d_n)\leq 2^{n^2+1}$.*

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 2.7. Clearly, $d_n(u,v) = n^{-1}\sup_{h\in B_2^n}\sum_{i=1}^n |h(u_i - v_i)|$. Thus, it is induced by a norm on the product space $\prod_{i=1}^n \mathbb{R}^n$, defined as $\|(x_1,...,x_n)\| = n^{-1}\sup_{h\in B_2^n}\sum_{i=1}^n |h(x_i)|$. Consider the unit ball of this norm, which we denote by $\mathcal{K}$. Fix $\varepsilon > 0$, and observe that our aim is to upper bound the maximal number of disjoint translates of $\varepsilon\mathcal{K}$ that are centered at points in $\mathcal{B} = \prod_{i=1}^n B_2^n$. To that end, we use a well-known volumetric argument, which is presented for the sake of completeness.

Let $\mathcal{U} = \varepsilon\mathcal{K}\cap\mathcal{B}$, which is also a convex, symmetric set, and note that $D(\mathcal{B},\varepsilon\mathcal{K}) \leq D(\mathcal{B},\mathcal{U})$. Let $y_1,...,y_m$ be elements in $\mathcal{B}$ such that for every $i\neq j$, $y_i+\mathcal{U}$ 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 $A\subset\prod_{i=1}^n\mathbb{R}^n$. Since the sets $y_i+\mathcal{U}$ are disjoint, then $\sum_{i=1}^m \mathrm{vol}(y_i+\mathcal{U}) \leq \mathrm{vol}(2\mathcal{B}) = 2^{n^2}\mathrm{vol}(\mathcal{B})$, and thus $m \leq 2^{n^2}\mathrm{vol}(\mathcal{B})/\mathrm{vol}(\mathcal{U})$. To conclude the proof it is enough to show that $\mathrm{vol}(\mathcal{B})/\mathrm{vol}(\mathcal{U}) \leq 2$ as long as $\varepsilon \geq c/\sqrt{n}$.

Let μ be the normalized volume measure on B_2^n. Therefore, if X is a random vector distributed according to μ and $\mathbb{X} = (X_i)_{i=1}^n$, where (X_i) are i.i.d. copies of X, then

$$\mathrm{vol}(\mathcal{U}) = \mathrm{vol}(\mathcal{B})\cdot Pr\left(\mathbb{X}\in\mathcal{U}\right).$$

Since $X_i\in B_2^n$, then

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

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 (see, for example, [9, 7]).

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

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

and to show that for a large enough c', this probability is smaller than $1/2$. Indeed, let $(\varepsilon_i)_{i=1}^n$ be independent, symmetric $\{-1,1\}$-valued random variables. By a symmetrization argument and the contraction principle for the absolute value function (see, e.g., [4]),

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

and the claim follows from Chebyshev's inequality. □

Proof of Theorem 2.4. By Lemma 2.5, the first assumption of Theorem 1.4 holds as long as $L\leq c_1(\delta)\sqrt{n}$, and by Theorem 2.7, $\log N(c/\sqrt{n}, B_2^n, d_n) \leq c'n^2$. Hence, the second condition in Theorem 1.4, for $L\leq c_2(\delta)\sqrt{n}$, yields that one can take $N\geq c_3(\delta)n$, as claimed. □

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. MR1984025 (2004f:68048)
2. B. Bollobás, *Extremal graph theory*, Academic Press, 1978. MR0506522 (80a:05120)
3. 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. MR2042049
4. M. Ledoux, M. Talagrand, *Probability in Banach spaces*, Springer, 1991. MR1102015 (93c:60001)
5. N. Linial, S. Mendelson, G. Schechtman, A. Shraibman, Complexity measures of sign matrices, Combinatorics, to appear.
6. S. Mendelson, Embedding with a Lipschitz function, Random Structures and Algorithms, 27(1) 25-45, 2005. MR2149294
7. V.D. Milman, G. Schechtman, *Asymptotic theory of finite dimensional normed spaces*, Lecture Notes in Mathematics 1200, Springer, 1986. MR0856576 (87m:46038)
8. A. Pajor, *Sous espaces ℓ_1^n des espaces de Banach*, Hermann, Paris, 1985. MR0903247 (88h:46028)
9. G. Pisier, *The volume of convex bodies and Banach space geometry*, Cambridge University Press, 1989. MR1036275 (91d:52005)

Centre for Mathematics and its Applications, The Australian National University, Canberra, ACT 0200, Australia

E-mail address: shahar.mendelson@anu.edu.au