The Shattering dimension of sets of linear functionals

Shahar Mendelson*
Research School of Information Sciences and Engineering,
The Australian National University,
Canberra, ACT 0200, Australia
e-mail: shahar.mendelson@anu.edu.au

Gideon Schechtman†
Department of Mathematics,
The Weizmann Institute,
Rehovot, Israel
e-mail: gideon@wisdom.weizmann.ac.il

October 6, 2002

Abstract

We evaluate the shattering dimension of various classes of linear functionals on various symmetric convex sets. This is applied in two different directions. The first is the determination of whether or not several classes of linear functionals satisfy the uniform Central Limit Theorem or the uniform Law of Large Numbers. The second direction, which was the main motivation of this study, is the estimation of error bounds for certain kernel machines used in Statistical Learning Theory. The proofs here relay mostly on methods from the Local Theory of Normed Spaces and include volume estimates, factorization techniques and tail estimates of norms, viewed as random variables on Euclidean spheres.

1 Introduction

Combinatorial dimensions, such as the Vapnik-Chervonenkis dimension and the shattering dimension are parameters which measure the richness of a given class of functions. The Vapnik-Chervonenkis dimension (VC dimension) of a class of $\{0,1\}$-valued functions is the largest dimension of a combinatorial cube that can be found in a coordinate projection of

*Supported by the ARC
†Supported by the ISF

the class, that is, in a restriction of the class to a finite subset of the domain. In this article we focus on a real valued analog of the VC dimension, called the shattering dimension; it is a scale sensitive parameter that measures the largest dimension of a "cube" of a given side length that can be found in a coordinate projection of the class.

Definition 1.1 *For every $\varepsilon > 0$, a set $\sigma = \{x_1, ..., x_n\} \subset \Omega$ is said to be ε-shattered by F if there is some function $s : \sigma \to \mathbb{R}$, such that for every $I \subset \{1, ..., n\}$ there is some $f_I \in F$ for which $f_I(x_i) \geq s(x_i) + \varepsilon$ if $i \in I$, and $f_I(x_i) \leq s(x_i) - \varepsilon$ if $i \notin I$. The shattering dimension of F is the function*

$$\mathrm{VC}(\varepsilon, F, \Omega) = \sup\Big\{|\sigma| \ \Big| \sigma \subset \Omega, \ \sigma \text{ is } \varepsilon\text{−shattered by } F\Big\}.$$

f_I is called the shattering function of the set I and the set $\{s(x_i)|x_i \in \sigma\}$ is called a witness to the ε-shattering. In cases where the underlying space is clear we denote the shattering dimension by $\mathrm{VC}(\varepsilon, F)$.

Combinatorial dimensions have been frequently used in the theory of empirical processes, mostly in the context of the uniform Law of Large Numbers and the uniform Central Limit Theorem. Recall the definition of the uniform law of large numbers, also known as the uniform Glivenko-Cantelli condition.

Definition 1.2 *Let F be a class of functions. We say that F is a uniform Glivenko Cantelli class (uGC class) if for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} \sup_{\mu} Pr\Big\{\sup_{f\in F} \big|\mathbb{E}_\mu f - \frac{1}{n}\sum_{i=1}^{n} f(X_i)\big| \geq \varepsilon\Big\} = 0, \tag{1.1}$$

where $(X_i)_{i=1}^{\infty}$ are independent random variables distributed according to μ.

Let us remark that in this article we ignore the question of measurability, since only mild assumption on the class are required to resolve this issue (see [7] for further details).

Vapnik and Chervonenkis proved that (under mild measurability assumptions) a class of Boolean functions is a uGC class if and only if it has a finite VC dimension [25], and this result was extended in [2] to the real-valued case, where it was shown that a class of uniformly bounded functions is a uGC class if and only if $\mathrm{VC}(\varepsilon, F)$ is finite for every $\varepsilon > 0$.

The shattering dimension can be used to obtain the tail bounds needed in (1.1), using the following line of argumentation. The starting point is a version of Talagrand's inequality (originally proved in [23]), due to Bousquet.

Theorem 1.3 *[5] Let F be a class of functions defined on a probability space (Ω, μ) such that $\sup_{f\in F} \|f\|_\infty \leq 1$. Let $(X_i)_{i=1}^{n}$ be independent random variables distributed according* *to* μ*, put* $\sigma^2 \geq \sup_{f\in F} \mathrm{Var}[f(X_1)]$ *and set* $Z = \sup_{f\in F} \left|\sum_{i=1}^n (f(X_i) - \mathbb{E}_\mu f)\right|$*. Then, for every* $x > 0$*,*

$$Pr\{Z \geq \mathbb{E}Z + x\} \leq \exp\left(-vh\left(\frac{x}{v}\right)\right), \tag{1.2}$$

where $v = n\sigma^2 + 2\mathbb{E}Z$ *and* $h(x) = (1+x)\log(1+x) - x$*. Moreover, for every* $x > 0$

$$Pr\{Z \geq \mathbb{E}Z + \sqrt{2xv} + \frac{x}{3}\} \leq e^{-x}. \tag{1.3}$$

In order to apply this result and obtain uniform deviation estimates one needs to bound $\mathbb{E}Z$. By symmetrization,

$$\mathbb{E}_\mu Z \leq 2\mathbb{E}_{\mu\times\varepsilon} \sup_{f\in F}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|, \tag{1.4}$$

where $(\varepsilon_i)_{i=1}^n$ are independent Rademacher random variables (i.e., take the values ± 1 with probability $1/2$ each). It turns out that the Rademacher averages in the left hand side of (1.4) can be estimated in terms of the *empirical covering numbers.*

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

Definition 1.4 *For every class* F *let the empirical covering numbers be*

$$N(\varepsilon, F, n) = \sup_{\mu_n} N(\varepsilon, F, L_2(\mu_n)),$$

where the supremum is taken with respect to all empirical measures $n^{-1}\sum_{i=1}^n \delta_{x_i}$ *supported on* n *points.* $\log N(\varepsilon, F) = \sup_n \log N_2(\varepsilon, F, n)$ *is called the uniform* L_2 *entropy of* F*.*

The following result shows that the uniform entropy can be bounded via the combinatorial parameters.

Theorem 1.5 *[16] There are absolute constants* K *and* c *such that for any class* F *which consists of functions bounded by* 1 *and every* $0 < \varepsilon < 1$*,*

$$N(\varepsilon, F) \leq \left(\frac{2}{\varepsilon}\right)^{K\cdot \mathrm{VC}(c\varepsilon, F)}.$$

Combining Theorem 1.5 with a chaining argument one can bound the Rademacher averages of (1.4) and thus $\mathbb{E}Z$ and obtain the necessary deviation estimates.

Theorem 1.6 *[17] Let* F *be a class of functions bounded by* 1*, and set* Z *to be as in Theorem 1.3. Assume that there are* $\gamma \geq 1$ *and* $0 < p < \infty$ *such that* $\mathrm{VC}(\varepsilon, F) \leq \gamma\varepsilon^{-p}$*. Then*

$$\mathbb{E}Z \leq C_p\gamma^{1/2}\begin{cases} \sqrt{n} & \text{if } 0 < p < 2, \\ \sqrt{n}\log^{3/2} n & \text{if } p = 2, \\ n^{1-1/p} & \text{if } p > 2, \end{cases}$$

where C_p *are constants which depend only on* p*.*

In Section 3 we give a necessary and sufficient condition for the unit ball B_{X^*} of a dual Banach space to be uGC class on B_X. In section 4 we use the considerations above to find that certain families of linear functionals acting on certain symmetric convex sets satisfy the uniform Central Limit Theorem.

We now turn to the description of the connection between bounds on the shattering dimension and error bounds used in the analysis of regression problems in nonparametric statistics and, more recently, in Machine Learning. In both applications, combinatorial parameters have played an important role. In the context of Machine Learning, they where used to estimate the size of a random sample needed to construct an almost optimal approximation of an unknown target function by an element in a fixed class of functions, where the give data are a sample $(X_i)_{i=1}^n$ and the values of the target on the sample [1, 17]. Such an error bound which is based on the shattering dimension is presented in the next theorem, which was adapted from [4].

Theorem 1.7 *Let (Ω, μ) be a probability space, set F be a class of measurable functions on Ω with ranges in $[-1,1]$ and assume that there is a constant $B \geq 1$ such that for every $f \in F$, $\mathbb{E}_\mu f^2 \leq B\mathbb{E}_\mu f$. If $(X_i)_{i=1}^n$ are independent random variables distributed according to μ, then for any $x > 0$ there is a set of probability larger than $1 - 2e^{-x}$, on which for any $f \in F$,*

$$\mathbb{E}_\mu f \leq \frac{2}{n}\sum_{i=1}^n f(X_i) + C\Big(\frac{I}{\sqrt{n}} + \frac{Bx}{n}\Big),$$

where C is an absolute constant and

$$I = \int_0^1 \sqrt{\mathrm{VC}(\varepsilon, F, \{X_1, ..., X_n\}) \log\Big(\frac{1}{\varepsilon}\Big)}\, d\varepsilon. \tag{1.5}$$

As an example, let G be a class of functions bounded by 1, and assume that $T \in G$. Set $F = \{(g-T)^2 | g \in G\}$, and note that by Theorem 1.7, for every $x > 0$ there is a set of probability larger than $1 - 2e^{-x}$ on which every function g which coincides with T on the sample $(X_1, ..., X_n)$ satisfies that $\mathbb{E}_\mu(g-T)^2 \leq C\big(\frac{I}{\sqrt{n}} + \frac{x}{n}\big)$. In particular, if I is small, any such g is an almost optimal approximation of T in G with respect to the $L_2(\mu)$ norm.

Let us mention that it is possible to obtain error bounds even in some cases when $I = \infty$ [15], and that in [4], error bounds with faster rates of convergence than $1/\sqrt{n}$ were established in the same setup.

In this article we investigate the shattering dimension of classes of linear functionals. The main reason for the focus on such classes comes from the growing interest in *kernel classes*, which take a central role in modern Machine Learning and Statistics (see e.g. [21, 6] and section 6.1). In a nutshell, kernel classes are associated with a positive definite, continuous function $K : \Omega \times \Omega \to \mathbb{R}$; a kernel class consists of functions $f = \sum_{i=1}^m \alpha_i K(x_i, \cdot)$ where $(x_i)_{i=1}^m \subset \Omega$ and $(\alpha_i)_{i=1}^m$ satisfy a certain constraint. Two examples of frequently used constraints are when $\sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) \leq 1$, in which case the class is the unit ball in the *reproducing kernel Hilbert space* associated with K, and when $\sum_i |\alpha_i|^p \leq 1$ for some $1 \leq p < \infty$.

An important property of kernel classes is that they can be represented as classes of linear functionals in the following sense: there is a Hilbert space H such that Ω can be embedded in a bounded subset of H via a mapping $x \to \Phi(x)$, and F is embedded in $H^* = H$ via the correspondence $f \to \Phi(f)$, such that for every x and f, $\langle \Phi(f), \Phi(x) \rangle = f(x)$. Hence, $\mathrm{VC}(\varepsilon, F, \Omega) = \mathrm{VC}(\varepsilon, \Phi(F), \Phi(\Omega))$, but the latter has an additional linear structure endowed by H.

The analysis of the shattering dimension of classes of linear functionals we present is based on methods in the local theory of normed spaces. We show that for such classes the shattering dimension is determined by the geometry of the class and the domain, which is expressed by the ability to factor a certain operator through ℓ_1^n. Firstly, we investigate the case when Ω is the unit ball of some Banach space and F is the dual unit ball. We show that if X is infinite dimensional and B_X is the unit ball of X, the shattering dimension $\mathrm{VC}(\varepsilon, B_{X^*}, B_X)$ is determined by the *Rademacher type* of X. In section 4 we use a volumetric argument and establish estimates on the shattering dimension when both the class and the domain are finite dimensional convex and symmetric sets. We then compute the shattering dimension of the unit ball in ℓ_p^n when considered as functions on the unit ball of ℓ_q^n, $1 \leq p, q \leq \infty$ and show that in many cases the volumetric approach yields sharp bounds. Our results are used to investigate the following problem: consider the unit ball of ℓ_q, denoted by B_q, as functions on the unit ball of ℓ_p. Does this class of functions satisfies the uniform Central Limit Theorem on this domain? In general, one can show that for any infinite dimensional Banach space X, $F = B_{X^*}$, does not satisfy the uniform Central Limit Theorem on the domain $\Omega = B_X$. We show that whenever $p < q$ $F = B_{\ell_{q'}}$ satisfies the uniform Central Limit Theorem on the domain $\Omega = B_p$ when q' is conjugate index to q.

Finally, we bound the shattering dimension of kernel classes. Our results are given in terms of eigenvalues associated with the kernel K; in Theorem 6.4 the bound is based on the spectrum of the integral operator $T_K f = \int K(x,y) f(y) d\nu(y)$ for a fixed measure ν, under an additional assumption that the eigenfunctions of the T_K are uniformly bounded, while in (6.1) we bound $\mathrm{VC}(\varepsilon, F, \{X_1, ..., X_n\})_{i=1}^n$ in terms of the eigenvalues of the Gram matrix $\big(K(X_i, X_j)\big)_{i,j=1}^n$. This result can be used to establish data-dependent bounds based solely on the values of the kernel on the given sample.

2 Preliminaries

Throughout, all absolute constants are denoted by c, C or K. Their values may change from line to line, or even within the same line. c_φ, C_φ denote constants which depend only on the parameter φ (which is usually a real number p or a couple of real numbers p, q), and $a \sim_\varphi b$ means that $c_\varphi b \le a \le C_\varphi b$. If the constants are absolute we use the notation $a \sim b$. Given a real Banach space X, let B_X or $B(X)$ be the unit ball of X. The dual of X, denoted by X^*, consists of all the bounded linear functionals on X, endowed with the norm $\|x^*\| = \sup_{\|x\|=1} |x^*(x)|$. For every integer n, we fix the Euclidean structure $\langle \ , \ \rangle$ on $\mathbb{R}^n$ with an orthonormal basis denoted by $(e_i)_{i=1}^n$.

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 $\mathrm{absconv}(K)$, is the convex hull of $K \cup -K$.

If $K \subset \mathbb{R}^n$ is bounded, convex and symmetric with a non empty interior, then K is a unit ball of a norm denoted by $\| \ \|_K$. It is possible to show that the *polar* of K, defined by

$$K^o = \{x \in \mathbb{R}^n \,|\, \sup_{k \in K} \langle k, x \rangle \le 1\}$$

is the unit ball of the dual space of $(\mathbb{R}^n, \| \ \|_K)$. In the sequel we shall abuse notation and denote by K the normed space whose unit ball is K. From here on, a ball will be a bounded, convex and symmetric subset of $\mathbb{R}^n$, with a nonempty interior.

If $1 \le p < \infty$, let ℓ_p^n be $\mathbb{R}^n$ endowed with the norm $\|\sum_{i=1}^n a_i e_i\|_p = (\sum_{i=1}^n |a_i|^p)^{1/p}$. ℓ_∞^n is $\mathbb{R}^n$ 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, and for every $1 \le p \le \infty$, $(B_p^n)^o = B_{p'}^n$, where $1/p + 1/p' = 1$. In this case, p' is called the conjugate index of p.

2.1 Volume estimates

As stated above, we can identify ℓ_2^n with $\mathbb{R}^n$. Hence, ℓ_2^n is endowed with the n-dimensional Lebesgue measure, denoted by $|\cdot|$. Let GL_n be the set of invertiable operators $T : \mathbb{R}^n \to \mathbb{R}^n$, and note that for every measurable set $A \subset \mathbb{R}^n$ and every $T \in GL_n$, $|TA| = |\det(T)|\,|A|$. We say that a set $A \subset \mathbb{R}^n$ is an ellipsoid if there is some $T \in GL_n$, such that $A = TB_2^n$.

It will be useful to determine the volume of the balls B_p^n and the volume of their sections. First, let us mention the following well known fact.

Theorem 2.1 *[20] There are absolute constants C and c such that for every integer n and every $1 \le p \le \infty$,*

$$cn^{-\frac{1}{p}} \le \left|B_p^n\right|^{\frac{1}{n}} \le Cn^{-\frac{1}{p}}.$$

Unlike the clear structure of sections of B_2^n, the geometry of sections of B_p^n is far less obvious. The following result, due to Meyer and Pajor [18], bounds the volume of k-dimensional sections of B_p^n.

Theorem 2.2 *For every k-dimensional subspace $E \subset \mathbb{R}^n$ and every $1 \le p \le q \le \infty$,*

$$\frac{|B_p^n \cap E|}{|B_p^k|} \le \frac{|B_q^n \cap E|}{|B_q^k|}.$$

By selecting $q = 2$ it follows that for $1 \le p \le 2$, the volume of any k-dimensional section of B_p^n is smaller than the volume of B_p^k. Similarly, by taking $p = 2$, the volume of any k-dimensional section of B_q^n for $2 \le q \le \infty$ is larger than the volume of B_q^k.

Remark 2.3 *A similar result holds in the infinite dimensional case. In particular, it follows that for any $1 \le p \le q \le \infty$ and for any n-dimensional subspace E,*

$$\Big(\frac{|B_p \cap E|}{|B_q \cap E|}\Big)^{\frac{1}{n}} \le C_{p,q} n^{\frac{1}{q} - \frac{1}{p}}. \tag{2.1}$$

An important fact about the volume of balls are the Santaló and inverse Santaló inequalities.

Theorem 2.4 *There is an absolute constant c such that for every integer n and every ball $K \subset \mathbb{R}^n$,*

$$c \le \left(\frac{|K|\,|K^o|}{|B_2^n|^2}\right)^{\frac{1}{n}} \le 1$$

The upper bound was established by Santaló, while the lower bound is due to Bourgain and Milman. The proof of both results can be found in [20].

One of the tools used in modern convex geometry is the notion of volume ratios. The idea is to compare the volume of a given ball with the "best" possible volume of an ellipsoid contained in it, since this may be used to understand "how close" the norm induced by the ball is to an Euclidean structure.

Definition 2.5 *For every ball $K \subset \mathbb{R}^n$, the volume ratio of K is*

$$\mathrm{vr}(K) = \sup\left(\frac{|K|}{|TB_2^n|}\right)^{\frac{1}{n}},$$

where the supremum is taken with respect to all $T \in GL_n$ such that $TB_2^n \subset K$.

The external volume ratio is defined as

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

where the infimum is with respect to all $T \in GL_n$ such that $K \subset TB_2^n$

It is possible to show [20] that both the supremum and the infimum in the definition above are uniquely attained. Hence, for every ball $K \subset \mathbb{R}^n$ there is an ellipsoid of maximal volume contained in K and an ellipsoid of minimal volume containing K. The ellipsoid of maximal volume contained in K is denoted by $\mathcal{E}_K$, and the ellipsoid of minimal volume containing K is denoted by $\tilde{\mathcal{E}}_K$. Note that for every ball K, $\mathcal{E}_K^o = \tilde{\mathcal{E}}_{K^o}$.

It follows from the definitions that if K is an ellipsoid then $\mathrm{vr}(K) = \mathrm{evr}(K) = 1$. Moreover, it is known that for every ball $K \subset \mathbb{R}^n$, $\mathrm{vr}(K) \leq \sqrt{n}$. More precisely, the volume ratio of ℓ_∞^n, which is of the order of $\sqrt{n}$, is the worst possible:

Theorem 2.6 *[3] For every integer n,*

$$\mathrm{vr}(K) \leq \mathrm{vr}(B_\infty^n) = \frac{4}{|B_2^n|^{1/n}}.$$

Another result we require is an estimate on the volume ratios of projections of ℓ_p.

Theorem 2.7 *[13] For every integer n,*

$$\sup_{E \subset \mathbb{R}^n} \mathrm{vr}(P_E B_p) \sim_p \begin{cases} 1 & 1 < p \leq 2 \\ n^{\frac{1}{2} - \frac{1}{p}} & 2 < p \leq \infty \end{cases}$$

where the supremum is taken with respect to all the projections onto n dimensional subspaces of ℓ_p.

A different notion of volume ratios is the *cubic ratios* which was introduced by K. Ball. For every ball $K \subset \mathbb{R}^n$ let

$$\mathrm{cr}(K) = \inf_{T \in GL_n,\ K \subset TB_\infty^n} \Big(\frac{|TB_\infty^n|}{|K|}\Big)^{\frac{1}{n}}.$$

Lemma 2.8 *[3] There are absolute constants c and C such that for every integer n and every ball $K \subset \mathbb{R}^n$,*

$$c\sqrt{n} \leq \mathrm{vr}(K)\mathrm{cr}(K) \leq C\sqrt{n}.$$

Finally, we can define the volume numbers of an operator. We follow the definition used by Gordon and Junge [12, 13].

Definition 2.9 *Given Banach spaces X and Y, an operator $T : X \to Y$ and an integer n, let the n-th volume number of T be*

$$v_n(T) = \sup\left\{ \left(\frac{|T(B_X \cap E)|}{|B_Y \cap F|}\right)^{\frac{1}{n}} \Big| E \subset X,\ T(E) \subset F \subset Y,\ \dim E = \dim F = n \right\}.$$

Note that if T is of rank smaller than n, $v_n(T) = 0$. Also, it is clear that the volume numbers are sub-multiplicative, i.e., $v_n(T_1T_2) \le v_n(T_1)v_n(T_2)$. If T is an operator between Hilbert spaces then the volume numbers may be calculated using the eigenvalues (λ_i) of $\sqrt{T^*T}$ (which are arranged in a non increasing order). In that case, for every integer n, $v_n(T) = (\prod_{i=1}^n \lambda_i)^{1/n}$.

Another example in which the volume numbers may be estimated is for the formal identity operator $id : \ell_q^m \to \ell_p^m$. By Theorem 2.2 it is evident that for every $n \le m$ and any $1 \le p \le q \le \infty$

$$v_n(id_{\ell_p^m \to \ell_q^m}) = \sup_{\dim E = n} \frac{\left|B_p^m \cap E\right|^{\frac{1}{n}}}{\left|B_q^m \cap E\right|^{\frac{1}{n}}} \le \frac{\left|B_p^n\right|^{\frac{1}{n}}}{\left|B_q^n\right|^{\frac{1}{n}}} \le C_{p,q} n^{\frac{1}{q}-\frac{1}{p}} \tag{2.2}$$

and clearly also

$$v_n(id_{\ell_p^m \to \ell_q^m}) \ge \frac{\left|B_p^n\right|^{\frac{1}{n}}}{\left|B_q^n\right|^{\frac{1}{n}}}. \tag{2.3}$$

In general, if $p \ge q$ then

$$v_n(id_{\ell_p^m \to \ell_q^m}) \le \|id\|_{\ell_p^m \to \ell_q^m} = m^{\frac{1}{q}-\frac{1}{p}}, \tag{2.4}$$

and this estimate is optimal, at least in cases where n divides m. To see this, let $k = m/n$ and for $j = 1, ...n$ let $v_j = \sum_{i=1}^k e_{j+k(i-1)}$. Note that for each r, $\text{span}\{v_1, ..., v_n\} \cap B_r^m = E \cap B_r^m$ has volume $(m/n)^{1/2-1/r}|B_r^n|$. Thus,

$$\frac{|B_p^m \cap E|^{1/n}}{|B_q^m \cap E|^{1/n}} = \left(\frac{m}{n}\right)^{\frac{1}{q}-\frac{1}{p}} \cdot \frac{|B_p^n|^{1/n}}{|B_q^n|^{1/n}} = C_{p,q} m^{\frac{1}{q}-\frac{1}{p}},$$

proving that the bound on the volume numbers is tight.

2.2 Uniform Central Limit Theorem

The fact that the shattering dimension can be used to bound the uniform entropy will enable us to show that some classes of functionals satisfy the *uniform CLT*.

Definition 2.10 *[7] Let $F \subset B\big(L_\infty(\Omega)\big)$, set P to be a probability measure on Ω and assume G_P to be a gaussian process indexed by F, which has mean 0 and covariance*

$$\mathbb{E} G_P(f) G_P(g) = \int fg dP - \int f dP \int g dP.$$

F is called a universal Donsker class if for any probability measure P the law G_P is tight in $\ell_\infty(F)$ and $\nu_n^P = n^{1/2}(P_n - P) \in \ell_\infty(F)$ converges in law to G_P in $\ell_\infty(F)$, where P_n a random empirical measure selected according to P.

Stronger than the universal Donsker property is the uniform Donsker property, which is the uniform version of the central limit theorem. For such classes, ν_n^P converges to G_P uniformly in P in some sense (see [7, 24] for more details). The following result of Giné and Zinn [9] is a relatively simple characterization of uniform Donsker classes.

For every probability measure P on Ω, let $\rho_P^2(f,g) = \mathbb{E}_P(f-g)^2 - \big(\mathbb{E}_P(f-g)\big)^2$, and for every $\delta > 0$, set $F_\delta = \{f - g | f, g \in F,\ \rho_P(f,g) \leq \delta\}$.

Theorem 2.11 *[9] F is a uniform Donsker class if and only if the following holds: for every probability measure P on Ω, G_P has a version with bounded, ρ_P-uniformly continuous sample paths, and for these versions,*

$$\sup_P \mathbb{E} \sup_{f \in F} |G_P(f)| < \infty, \qquad \lim_{\delta \to 0} \sup_P \mathbb{E} \sup_{h \in F_\delta} |G_P(h)| = 0.$$

The main tool in the analysis of uniform Donsker classes is the Koltchinskii-Pollard entropy integral.

Theorem 2.12 *[7] If $F \subset B\big(L_\infty(\Omega)\big)$ satisfies that*

$$\int_0^\infty \sup_n \sup_{\mu_n} \sqrt{\log N\big(\varepsilon, F, L_2(\mu_n)\big)}\, d\varepsilon < \infty,$$

then it is a uniform Donsker class.

3 Shattering by B_{X^*}

The goal of this section is to bound the shattering dimension of the dual unit ball of a given Banach space. To that end, we present the geometric interpretation of the shattering dimension when $\Omega \subset X$ and $F = B_{X^*}$.

Lemma 3.1 *Let X be a Banach space. Assume that $\{x_1, ..., x_n\}$ is ε-shattered by B_{X^*} and set $E = \operatorname{span}\{x_1, ..., x_n\}$. If A is the symmetric convex hull of $\{x_1, ..., x_n\}$ then $\varepsilon(B_X \cap E) \subset A$.*

Proof. Let $\{x_1, ..., x_n\}$ be ε-shattered by B_{X^*} and let $\{s_1, ..., s_n\}$ to be a witness of the shattering. Put $(a_i)_{i=1}^n \subset \mathbb{R}$, set $I = \{i | a_i \geq 0\}$ and let x_I^* be the functional shattering the set I. For every such I and every $i \in I$,

$$x_I^*(x_i) - x_{I^c}^*(x_i) \geq s_i + \varepsilon - (s_i - \varepsilon) = 2\varepsilon,$$

and if $i \notin I$,

$$x_I^*(x_i) - x_{I^c}^*(x_i) \leq s_i - \varepsilon - (s_i + \varepsilon) = -2\varepsilon.$$

Thus,

$$\left\|\sum_{i=1}^{n} a_i x_i\right\| = \sup_{x^*\in B_{X^*}} \left|x^*\Big(\sum_{i=1}^{n} a_i x_i\Big)\right|$$
$$\geq \frac{1}{2}\sup_{x^*,\tilde{x}^*\in B_{X^*}} \left|x^*\Big(\sum_{i=1}^{n} a_i x_i\Big) - \tilde{x}^*\Big(\sum_{i=1}^{n} a_i x_i\Big)\right| = (*)$$

Selecting $x^* = x_I^*$ and $\tilde{x}^* = x_{I^c}^*$,

$$(*) \geq \frac{1}{2}\left|x_I^*\Big(\sum_{i\in I} a_i x_i + \sum_{i\in I^c} a_i x_i\Big) - x_{I^c}^*\Big(\sum_{i\in I} a_i x_i + \sum_{i\in I^c} a_i x_i\Big)\right|$$
$$= \frac{1}{2}\left|\sum_{i\in I} a_i\big(x_I^*(x_i) - x_{I^c}^*(x_i)\big) + \sum_{i\in I^c}(-a_i)\big(x_{I^c}^*(x_i) - x_I^*(x_i)\big)\right|$$
$$\geq \varepsilon\sum_{i=1}^{n}|a_i|\,.$$

Since every point x on the boundary of A is given by $x = \sum_{i=1}^n a_i x_i$, where $\sum_{i=1}^n |a_i| = 1$, then $\|x\| = \|\sum_{i=1}^n a_i x_i\| \geq \varepsilon$, which proves our claim. ■

Corollary 3.2 *$\{x_1,...,x_n\} \subset B_X$ is ε-shattered by B_{X^*} if and only if $(x_i)_{i=1}^n$ are linearly independent and ε-dominate the ℓ_1^n unit-vector basis; i.e., $\varepsilon\sum_{i=1}^n |a_i| \leq \|\sum_{i=1}^n a_i x_i\|$ for every $a_1,...,a_n \in \mathbb{R}$.*

Proof. Let $E = \mathrm{span}\{x_1,...,x_n\}$ for some linearly independent elements of B_X and define $T : \ell_1^n \to \ell_2^n$ by $Te_i = x_i$. For every $I \subset \{1,...,n\}$ there is some $v \in B_\infty^n$ such that $\langle v, e_i\rangle = 1$ if $i \in I$ and $\langle v, e_j\rangle = -1$ otherwise. Note that $\langle v, e_i\rangle = \langle v, T^{-1}Te_i\rangle = \langle {T^{-1}}^* v, Te_i\rangle$ and that $A^o = (TB_1^n)^o = {T^{-1}}^* B_\infty^n$, implying that ${T^{-1}}^* v \in A^o$. If $\{x_1 \ldots, x_n\}$ ε-dominate the ℓ_1^n unit-vector basis then $\varepsilon(B_X\cap E) \subset A$ and $A^o \subset \varepsilon^{-1}(B_X\cap E)^o = \varepsilon^{-1}P_E B_{X^*}$, where P_E is the orthogonal projection onto E. Thus, there is some $x^* \in B_{X^*}$ such that ${T^{-1}}^* v = tP_E x^*$ for some $0 < t \leq \varepsilon^{-1}$. Hence, $\langle x^*, x_i\rangle = \langle x^*, Te_i\rangle = \langle P_E x^*, Te_i\rangle = t^{-1}\langle {T^{-1}}^* v, Te_i\rangle \geq \varepsilon$ if $i \in I$. By a similar argument, $\langle x^*, Te_j\rangle \leq -\varepsilon$ if $j \notin I$, which shows that $\{x_1,...,x_n\}$ is ε-shattered by B_{X^*}.

Conversely, if $\{x_1,...,x_n\} \subset B_X$ is ε-shattered then for every $a_1,...,a_n \in \mathbb{R}$,

$$\varepsilon\sum_{i=1}^{n}|a_i| \leq \|\sum_{i=1}^{n} a_i x_i\|,$$

hence $(x_i)_{i=1}^n$ are independent and ε-dominate the ℓ_1^n unit-vector basis. ■

This result enables us to estimate the shattering dimension of the dual unit ball of an infinite dimensional Banach space X when considered as a class of functions on B_X. It turns out that the shattering dimension is determined by the *type* of X.

Definition 3.3 *A Banach space X has type p if there is some constant C such that for every integer n and every $x_1, ..., x_n \in X$,*

$$\mathbb{E}\left\|\sum_{i=1}^{n} \varepsilon_i x_i\right\| \leq C\Big(\sum_{i=1}^{n} \|x_i\|^p\Big)^{1/p} \tag{3.1}$$

where $(\varepsilon_i)_{i=1}^n$ are independent Rademacher random variables. The smallest constant for which (3.1) *holds is called the type p constant of X and is denoted by $T_p(X)$.*

The basic facts concerning the concept of type may be found, for example, in [19]. Clearly, for every Banach space (3.1) holds in the case $p = 1$ with $T_1(X) = 1$. If $p^* = \sup\{p|\text{X has type } p\}$ then $1 \leq p^* \leq 2$. If $p^* = 1$ then X is said to have a trivial type.

Recall that the distance between two isomorphic Banach spaces X and Y is defined as $d(X, Y) = \inf \|T\| \cdot \|T^{-1}\|$ where the infimum is take with respect to all isomorphisms between X and Y. It is easy to see that if X,Y and Z are isomorphic, then $d(X, Z) \leq d(X, Y) \cdot d(Y, Z)$.

Theorem 3.4 *Let X be an infinite dimensional Banach space. Then $\mathrm{VC}(\varepsilon, B_{X^*}, B_X)$ is finite for every $\varepsilon > 0$ if and only if X has a nontrivial type. If X has type p then*

$$\Big(\frac{1}{\varepsilon}\Big)^{\frac{p^*}{p^*-1}} - 1 \leq \mathrm{VC}(\varepsilon, B_{X^*}, B_X) \leq \Big(\frac{T_p(X)}{\varepsilon}\Big)^{\frac{p}{p-1}} + 1.$$

The lower bound and a weaker version of the upper one were established in [14]. We repeat the proof of the lower bound for the sake of completeness.

Proof. If $\{x_1, ..., x_n\}$ is ε-shattered then it ε-dominates the ℓ_1^n unit-vector basis. By selecting $a_i = \varepsilon_i$, $\varepsilon n \leq \|\sum_{i=1}^n \varepsilon_i x_i\|$. On the other hand, taking the expectation with respect to the Rademacher variables

$$\mathbb{E}\left\|\sum_{i=1}^{n} \varepsilon_i x_i\right\|_X \leq T_p(X)\Big(\sum_{i=1}^{n} \|x_i\|_X^p\Big)^{\frac{1}{p}} \leq T_p(X) n^{\frac{1}{p}}.$$

Thus, there is a realization $(\varepsilon_i)_{i=1}^n$ such that $\|\sum_{i=1}^n \varepsilon_i x_i\| \leq T_p(X) n^{1/p}$. Combining the two inequalities, $n \leq (T_p(X)/\varepsilon)^{p/(p-1)}$.

Conversely, for every $\lambda > 0$ and every integer n, there is a subspace $X_n \subset X$ such that $\dim X_n = n$ and $d(\ell_{p^*}^n, X_n) \leq 1 + \lambda$ (see [19]). Recall that $d(\ell_1^n, \ell_{p^*}^n) = n^{1-1/p^*}$ (see [22]), hence, $d(X_n, \ell_1^n) \leq (1+\lambda)n^{1-1/p^*}$, and in particular, there are $x_1, ..., x_n \subset B_X$ such that for every $(a_i)_{i=1}^n \subset \mathbb{R}$,

$$\frac{1}{(1+\lambda)n^{1-1/p^*}} \sum_{i=1}^n |a_i| \leq \| \sum_{i=1}^n a_i x_i \|.$$

Therefore, $\{x_1, ..., x_n\}$ is $n^{(1-p^*)/p^*}(1+\lambda)^{-1}$-shattered by B_{X^*}. The claim follows by taking λ to 0.

The assertion in the case $p^* = 1$ follows in a similar manner. ■

The uGC part of the next corollary was first proved in [8]; the proof presented below is new.

Corollary 3.5 *Let X be an infinite dimensional Banach space. Then, $F = B_{X^*}$ is a uGC class on $\Omega = B_X$ if and only if X has a nontrivial type. Also, for any infinite dimensional X, F is not a uniform Donsker class on Ω.*

Proof. The fact that the pair is a uGC class if and only if X has a nontrivial type follows from Theorem 3.4 and the characterization of uGC classes as classes with a finite shattering dimension at every scale ε (see [2]).

As for the second part, in [9] example 3.3, it was shown that if $X = \ell_2$ then $F = B_{\ell_2}$ is not a uniform Donsker class on $\Omega = B_{\ell_2}$. Moreover, an easy modification of the proof reveals the following: if there is a constant C such that for every integer n there are spaces $X_n \subset X$ of dimension n for which $d(X_n, \ell_2^n) \leq C$ then $F = B_{X^*}$ is not a uniform Donsker class on $\Omega = B_X$. By Dvoretzky's Theorem [19], every infinite dimensional Banach space has such subspaces X_n (with a constant C arbitrarily close to 1). ■

4 The shattering dimension of finite dimensional bodies

Unlike the infinite dimensional case, in which the growth of $\mathrm{VC}(\varepsilon, B_{X^*}, B_X)$ is determined by the type of X, it is not clear whether the same holds for finite dimensional spaces; indeed, the lower bound in Theorem 3.4 is based on the fact that X contains spaces which are arbitrarily close to $\ell_{p^*}^n$ for every integer n, which is only true for infinite dimensional spaces.

As we show in the sequel, some applications require that the set of functionals is not the dual of the domain but some other convex symmetric set; thus, in the finite dimensional context it is natural to investigate the following question.

Question 4.1 *Let K and L be two convex symmetric bodies in $\mathbb{R}^d$ and view the elements of L° as functions on K using the fixed inner product in $\mathbb{R}^d$. What is $\mathrm{VC}(\varepsilon, L^\circ, K)$?*

We have shown that $\mathrm{VC}(\varepsilon, L^\circ, K) = n$ if and only if n is the largest such that there are n points $\{x_1, ..., x_n\} \subset K$ for which $\varepsilon(L \cap E) \subset \mathrm{absconv}(x_1, ..., x_n)$, where $E = \mathrm{span}\{x_1, ..., x_n\}$.

The next theorem provides a general upper bound on $\mathrm{VC}(\varepsilon, L^\circ, K)$ based on a volumetric argument. The result is presented for finite dimensional bodies but can be easily extended to the infinite dimensional case.

Theorem 4.2 *There is an absolute constant C such that for every integers $n \leq m$ and every convex and symmetric sets $K, L \subset \mathbb{R}^m$ the following holds: if $\{x_1, ..., x_n\} \subset K$ is ε-shattered by L^o then*

$$\sqrt{n} \leq \frac{C}{\varepsilon} \mathrm{vr}\big((K \cap E)^o\big) \frac{|K \cap E|^{\frac{1}{n}}}{|L \cap E|^{\frac{1}{n}}},$$

where $E = \mathrm{span}\{x_1, ..., x_n\}$.

Proof: Assume that $\{x_1, ..., x_n\} \subset K$ is ε-shattered by L^o. By Lemma 3.1, $\varepsilon(L \cap E) \subset A \subset K \cap E$, where A is the symmetric convex hull of $\{x_1, ..., x_n\}$, and thus, $(K \cap E)^o \subset A^o$. By Lemma 2.8

$$\begin{aligned} c\sqrt{n} \leq & \mathrm{vr}\big((K \cap E)^o\big) \mathrm{cr}\big((K \cap E)^o\big) \leq \mathrm{vr}\big((K \cap E)^o\big) \left(\frac{|A^o|}{|(K \cap E)^o|} \right)^{\frac{1}{n}} \\ \leq & \frac{1}{\varepsilon} \mathrm{vr}\big((K \cap E)^o\big) \left(\frac{|(L \cap E)^o|}{|(K \cap E)^o|} \right)^{\frac{1}{n}} \leq \frac{C}{\varepsilon} \mathrm{vr}\big((K \cap E)^o\big) \left(\frac{|K \cap E|}{|L \cap E|} \right)^{\frac{1}{n}}, \end{aligned}$$

where the last inequality follows from the Santalò and inverse Santalò inequalities. ■

Combining this theorem with Remark 2.3 on the ratio $|B_p \cap E| / |B_q \cap E|$ and Theorem 2.7 on the volume ratio of projections of ℓ_p, the following is evident:

Corollary 4.3 *For every $1 \leq p \leq q < \infty$ there is a constant $C_{p,q}$ for which the following holds: if $\{x_1, ... x_n\} \subset B_p$ is ε-shattered by $B_{q'}$ then*

$$\varepsilon \leq C_{p,q} \begin{cases} n^{1/p - 1/2} & \text{if } 1 \leq p \leq 2 \\ n^{1/q - 1/p - 1/2} & \text{if } 2 < p < \infty. \end{cases}$$

In the sequel we will show that this estimate is sharp. Since a similar argument is used in the proof of Theorem 4.10, we shall not present the proof of the corollary here.

Let us mention the following observations; firstly, using Santalò's inequality, $\mathrm{vr}\big((K \cap E)^o\big) |K \cap E|^{\frac{1}{n}} \leq |\tilde{\mathcal{E}}_{K \cap E}|^{\frac{1}{n}}$. Therefore, from the volumetric point of view, all that matters is the ratio between the volume of the ellipsoid of minimal volume containing the section of K spanned by $\{x_1, ..., x_n\}$ and the volume of $L \cap E$.

Secondly, estimating the shattering dimension is equivalent to understanding the behavior of its formal inverse, which, for a given linearly independent set $\{x_1, ..., x_n\} \subset K$ is the largest $\varepsilon > 0$ for which $\varepsilon(L \cap E) \subset \text{absconv}(x_1, ..., x_n)$, where $E = \text{span}\{x_1, ..., x_n\}$. Thus, one can take $K = TB_1^n$, where $T : \ell_1^n \to \ell_2$ is defined by $Te_i = x_i$, and the volume of the ellipsoid of minimal volume containing TB_1^n is the significant quantity.

If $(\lambda_i)_{i=1}^n$ are the singular values of the operator T, that is, the eigenvalues of $\sqrt{T^*T}$, then $|\tilde{\mathcal{E}}_{TB_1^n}|^{1/n}$ is equivalent to $n^{-1/2}\left(\prod_{i=1}^n \lambda_i\right)^{1/n}$. Also, $\mathcal{E} \subset \ell_2$ is an ellipsoid and if $TB_1^n \subset \mathcal{E}$ then $|\tilde{\mathcal{E}}_{TB_1^n}| \leq \prod_{i=1}^n a_i$, where $(a_i)_{i=1}^\infty$ are the lengths of the principle axes of $\mathcal{E}$ arranged in a non-increasing order. Indeed, the n-dimensional section of $\mathcal{E}$ spanned by TB_1^n is an ellipsoid containing TB_1^n, therefore, its volume must be larger than that of $\tilde{\mathcal{E}}_{TB_1^n}$.

4.1 Shattering and factorization through ℓ_1^n

An alternative way to formulate the problem of estimating the shattering dimension is as a factorization problem.

Definition 4.4 *For every two balls K and L in $\mathbb{R}^m$ and every integer $n \leq m$, let*

$$\Gamma_n(K, L) = \inf \|A\| \, \|B\| \, ;$$

the infimum is taken with respect to all subspaces of $E \subset \mathbb{R}^m$ of dimension n, and all operators $B : (E, \| \, \|_{L\cap E}) \to \ell_1^n$, $A : \ell_1^n \to (E, \| \, \|_{K\cap E})$ such that $AB = id : L\cap E \to K\cap E$.

The following lemma shows that $1/\Gamma_n(K, L)$ is the formal inverse of the shattering dimension.

Lemma 4.5 *For every integer n and any balls K and L,*

$$\frac{1}{\Gamma_n(K, L)} = \sup\{\varepsilon | \exists \{x_1, ..., x_n\} \subset K, \ \varepsilon(L \cap E) \subset \text{absconv}(x_1, ..., x_n)\} \tag{4.1}$$

$$= \sup\{\varepsilon | \text{VC}(\varepsilon, L^\circ, K) \geq n\} \tag{4.2}$$

where $E = \text{span}\{x_1, ..., x_n\}$.

Proof. If the identity admits an optimal factorization $id = AB$, set $A' = A/\|A\|_{\ell_1^n \to K\cap E}$ and observe that the set $A'e_1, ..., A'e_n \subset K \cap E$ satisfies that for any $a_1, ..., a_n \in \mathbb{R}$

$$\|A\|_{\ell_1^n \to L\cap E} \cdot \|B\|_{K\cap E \to \ell_1^n} \cdot \| \sum_{i=1}^n a_i A'e_i \|_L \geq \|B(\sum_{i=1}^n a_i A e_i)\|_{\ell_1^n} \geq \| \sum_{i=1}^n a_i e_i \|_{\ell_1^n} = \sum_{i=1}^n |a_i|.$$

Hence, $\text{absconv}(A'e_1, ..., A'e_n) \subset K \cap E$ contains $(\|A\|\|B\|)^{-1}(L \cap E)$ and

$$\frac{1}{\Gamma_n(K, L)} \leq \sup\{\varepsilon | \exists \{x_1, ..., x_n\} \subset K, \ \varepsilon(L \cap E) \subset \text{absconv}(x_1, ..., x_n)\}.$$

For the reverse inequality, if $\{x_1, ..., x_n\} \subset K$ are such that $\varepsilon(L \cap E) \subset \text{absconv}(x_1, ..., x_n)$, define $T : \mathbb{R}^n \to \mathbb{R}^n$ by $Te_i = x_i$. Clearly, $\|T\|_{\ell_1^n \to K \cap E} \leq 1$ and

$$\|T^{-1}\|_{L \cap E \to \ell_1^n} = \sup_{x \in L \cap E} \|T^{-1}x\|_{\ell_1^n} = \sup_{x \in L \cap E} \|x\|_{T\ell_1^n} \leq \frac{1}{\varepsilon}.$$

Thus, $\|T\|_{\ell_1^n \to K \cap E} \cdot \|T^{-1}\|_{L \cap E \to \ell_1^n} \leq 1/\varepsilon$ and $1/\Gamma_n(K, L) \geq \varepsilon$. ■

Combining Theorem 4.2 and the previous lemma, we obtain

Corollary 4.6 *There is an absolute constant $c > 0$ such that for any two integers $n \leq m$ and any two balls $K, L \subset \mathbb{R}^m$,*

$$\Gamma_n(K, L) \geq c \, \frac{\sqrt{n}}{v_n(id : K \to L) \sup_E \text{vr}(P_E K^\circ)},$$

where $\dim(E) = n$.

4.2 Factorization constants of ℓ_p^m

The goal of the next section is to investigate the shattering dimension of the class of linear functionals $F = B_{q'}^m$ on $\Omega = B_p^m$ for $1 \leq p, q \leq \infty$. Firstly, In Theorems 4.9 and 4.10 below we present a tight estimate on the factorization constant of $id : \ell_q^n \to \ell_p^n$ through ℓ_1^n. Then, we use this result to estimate $\Gamma_n(B_p^m, B_q^m)$ and thus bound $\text{VC}(\varepsilon, B_{q'}^m, , B_p^m)$; finally, we show that if $1 \leq p < q \leq \infty$ then $F = B_{q'}$ is a uniform Donsker class on $\Omega = B_p$.

We begin with two lemmas needed for the proof of Theorem 4.9.

Lemma 4.7 *Let μ be the Haar measure on the n dimensional sphere S^{n-1}. Set K and L to be balls in $\mathbb{R}^m$ and put α to be such that*

$$\mu\Big(x \in S^{n-1} \big| \|x\|_K > \frac{1}{\alpha}\Big) < \frac{1}{2n}.$$

If ε satisfies that

$$\mu\Big(x \in S^{n-1} \big| \|x\|_{L^\circ} > \frac{\alpha}{\varepsilon\sqrt{n}}\Big) < 2^{-(n+1)}$$

then $\Gamma_n(K, L) \leq 1/\varepsilon$.

Proof. Denote by O_n the orthogonal group and let P_{O_n} be the Haar measure on O_n. Set $U \in O_n$ and define $x_i = \alpha U e_i$. Using the standard connection between P_{O_n} and μ on S^{n-1},

$$P_{O_n}\big(x_i \in K\big) = \mu\Big(x \in S^{n-1} \big| \|x\|_K \leq \frac{1}{\alpha}\Big),$$

hence

$$P_{O_n}\Big(x_i \in K \text{ for all } i\Big) \geq 1 - n\mu\Big(x \in S^{n-1}\big|\|x\|_K > \frac{1}{\alpha}\Big) > \frac{1}{2}.$$

Moreover,

$$P_{O_n}\big(\operatorname{conv}(\pm\alpha U e_i) \supset \varepsilon L\big) = P_{O_n}\Big(\sup_{(\sigma_i)_{i=1}^n \in \{-1,1\}^n} \|\frac{1}{\alpha}\sum_{i=1}^n \sigma_i U e_i\|_{L^\circ} \leq \frac{1}{\varepsilon}\Big).$$

For every vector $(\sigma_1, ..., \sigma_n) \in \{-1,1\}^n$,

$$\begin{aligned} P_{O_n}\Big(\|\frac{1}{\alpha}\sum_{i=1}^n \sigma_i U e_i\|_{L^\circ} > \frac{1}{\varepsilon}\Big) &= P_{O_n}\Big(\|U(\frac{1}{\sqrt{n}}\sum_{i=1}^n e_i)\|_{L^\circ} > \frac{\alpha}{\varepsilon\sqrt{n}}\Big) \\ &= \mu\Big(x \in S^{n-1}\big|\|x\|_{L^\circ} > \frac{\alpha}{\varepsilon\sqrt{n}}\Big). \end{aligned}$$

Thus,

$$P_{O_n}\big(\operatorname{conv}(\pm\alpha U e_i) \supset \varepsilon L\big) \geq 1 - 2^n\mu\big(x \in S^{n-1}\big|\|x\|_{L^\circ} \geq \frac{\alpha}{\varepsilon\sqrt{n}}\big) > \frac{1}{2},$$

and there is some orthogonal operator which belongs to both events. The operator $T = \alpha U$ satisfies that $\|T\|_{\ell_1^n \to K\cap E} \leq 1$ and $\|T^{-1}\|_{L\cap E \to \ell_1^n} \leq 1/\varepsilon$, as claimed. ■

Lemma 4.8 *There are constants C_p for which the following holds: for every integer n,*

$$\mu\Big(x \in S^{n-1} \ \big|\|x\|_{\ell_p^n} \geq C_p n^{\frac{1}{p}-\frac{1}{2}}\Big) \leq 2^{-(n+1)},$$

if $1 \leq p \leq 2$, and if $2 \leq p < \infty$ then

$$\mu\Big(x \in S^{n-1} \ \big|\|x\|_{\ell_p^n} \geq C_p n^{\frac{1}{p}-\frac{1}{2}}\Big) \leq e^{-n^{2/p}}.$$

Proof. denote by $M(B_p^n)$ the median of $\|x\|_p$ on S^{n-1}. By Levy's inequality [19],

$$\mu\Big(x \in S^{n-1}\big| \ \|x\|_p \geq (1+t)M(B_p^n)\Big) \leq e^{-\frac{t^2 n M^2(B_p^n)}{2\|id\|^2_{\ell_2^n \to \ell_p^n}}}.$$

Recall that $M(B_p^n) \sim_p n^{1/p-1/2}$ (see, e.g. [19]) and that $\|id\|_{\ell_2^n \to \ell_p^n} = \max\{n^{1/p-1/2}, 1\}$. It follows that for $1 \leq p \leq 2$ and C large enough, depending only on p,

$$\mu\Big(x \in S^{n-1} \ \big|\|x\|_{\ell_p^n} \geq Cn^{\frac{1}{p}-\frac{1}{2}}\Big) \leq e^{-c_p C^2 n} \leq 2^{-(n+1)},$$

while for $2 \leq p < \infty$ and C depending only on p,

$$\mu\Big(x \in S^{n-1} \ \big|\|x\|_{\ell_p^n} \geq Cn^{\frac{1}{2}-\frac{1}{p}}\Big) \leq e^{-n^{2/p}}.$$

■

The above results will play an important role in the proof of the following theorem, in which we construct factorizations of $id : \ell_q^n \to \ell_p^n$ through ℓ_1^n.

Theorem 4.9 *Let $K = B_p^n$ and $L = B_q^n$. Then, $\Gamma_n(K,L)$ satisfies that*

$$\Gamma_n(K,L) \le C_{p,q} \begin{cases} n^{\frac{1}{2}+\frac{1}{p}-\frac{1}{q}} & \text{if } 2 \le p,q \le \infty, \\ n^{1-\frac{1}{q}} & \text{if } 1 \le p \le 2, \\ n^{1-\frac{1}{q}} & \text{if } 1 \le q \le 2 \le p \le \infty \text{ and } p' > q, \\ n^{\frac{1}{p}} & \text{if } 1 \le q \le 2 \le p \le \infty \text{ and } p' \le q. \end{cases}$$

Proof. First, assume that $2 \le p < \infty$ and $2 \le q \le \infty$. By Lemma 4.8 and Lemma 4.7 it suffices to choose $1/\alpha = C_p n^{1/p-1/2}$ and select ε which satisfies that $C_q n^{1/q'-1/2} = C_q n^{1/2-1/q} = C_q \alpha/\varepsilon\sqrt{n}$, i.e., $\frac{1}{\varepsilon} \sim_q \frac{n^{1-\frac{1}{q}}}{\alpha}$. Therefore $\Gamma_n(B_p^n, B_q^n) \le C_{p,q} n^{1/2+1/p-1/q}$.

Next, if $1 \le p \le 2$ then

$$\Gamma_n(K,L) \le \|id\|_{\ell_q^n \to \ell_1^n} \|id\|_{\ell_1^n \to \ell_p^n} = n^{1-\frac{1}{q}}.$$

If $2 \le p < \infty$ and $1 \le q \le 2$, one has to treat two cases; if $1 \le q \le p'$, then using the identity operator as above, $\Gamma_n(B_p^n, B_q^n) \le n^{1-1/q}$. On the other hand, if $p' \le q \le 2$ then by the first part of our claim,

$$\Gamma_n(B_p^n, B_q^n) \le \|id\|_{\ell_q^n \to \ell_2^n} \Gamma_n(B_p^n, B_2^n) \le C_p n^{\frac{1}{p}}.$$

Finally, one has to address the situation when p is infinity. If $p = q = \infty$ then $\Gamma_n(B_\infty^n, B_\infty^n) = d(\ell_1^n, \ell_\infty^n) \le C n^{1/2}$ [22].

For $p = \infty$ we first examine the case $q = 2$. Let $\ell_p^n(\mathbb{C})$ to be $\mathbb{C}^n$ endowed with the ℓ_p norm and set $T = \left(n^{-1/2} e^{2\pi i jk/n}\right)_{j,k=1}^n$. It is easy to check that $\|T\|_{\ell_1^n(\mathbb{C}) \to \ell_\infty^n(\mathbb{C})} \le n^{-1/2}$ and that $\|T\|_{\ell_\infty^n(\mathbb{C}) \to \ell_2^n(\mathbb{C})} \le n^{1/2}$. For our purpose, $\ell_p^n(\mathbb{C})$ can be considered as the ℓ_p^n sum of 2-dimensional Euclidean spaces, ℓ_2^2, over the reals. Since for any $1 \le p \le \infty$, $\|id\|_{\ell_p^n(\mathbb{C}) \to \ell_p^{2n}} \cdot \|id\|_{\ell_p^n \to \ell_p^n(\mathbb{C})} \le \sqrt{2}$, then $\Gamma_{2n}(\ell_2^{2n}, \ell_\infty^{2n}) \le 2$. The case where n is odd is easily reduced to the even case.

Finally, for a general q,

$$\Gamma_n(B_\infty^n, B_q^n) \le \|id\|_{\ell_q^n \to \ell_2^n} \Gamma_n(B_\infty^n, B_2^n) \le C \|id\|_{\ell_q^n \to \ell_2^n},$$

as claimed. ■

The next step in our analysis is to show that the bounds in Theorem 4.9 are tight. The proof uses the notion of r-summing operators. Recall that an operator $T : X \to Y$ is r-summing for $1 \le r < \infty$, if there is a $C < \infty$ such that

$$\sum_{i=1}^{n} \|Tx_i\|^r \le C^r \sup_{x^* \in B_{X^*}} \sum_{i=1}^{n} |x^*(x_i)|^r \tag{4.3}$$

for all integers n and all $x_1, \ldots, x_n \in X$. The smallest C for which (4.3) holds is denoted by $\pi_r(T)$.

Theorem 4.10 *Let $K = B_p^n$ and $L = B_q^n$. Then, $\Gamma_n(K,L)$ satisfies that*

$$\Gamma_n(K,L) \ge c_{p,q} \begin{cases} n^{\frac{1}{2}+\frac{1}{p}-\frac{1}{q}} & \text{if } 2 \le p,q \le \infty, \\ \\ n^{1-\frac{1}{q}} & \text{if } 1 \le p \le 2. \end{cases}$$

Also, if $1 < q \le 2 \le p < \infty$ and any $r > \max\{p, q'\}$,

$$\Gamma_n(K,L) \ge c_{p,q,r} n^{\frac{1}{r}}.$$

Proof. The first two cases follow from the volumetric estimate as in Corollary 4.3. Indeed, if $\{x_1, ..., x_n\} \subset K$ is ε-shattered by L° then

$$\varepsilon\sqrt{n} \le C \mathrm{vr}(K^\circ) \left(\frac{|K|}{|L|} \right)^{\frac{1}{n}}$$

for some absolute constant C. Since $K = B_p^n$ and $L = B_q^n$ then $(|K|/|L|)^{1/n} \sim_p n^{1/q-1/p}$ and $\mathrm{vr}(K^\circ) \sim n^{1/p-1/2}$ for $1 \le p \le 2$ and ~ 1 for $2 \le p \le \infty$. Hence,

$$\varepsilon \le C_{p,q} \begin{cases} n^{\frac{1}{q}-\frac{1}{p}-\frac{1}{2}} & \text{if } 2 \le p \le \infty, \\ \\ n^{\frac{1}{q}-1} & \text{if } 1 \le p \le 2, \end{cases}$$

and the lower estimate on Γ_n is evident from Lemma 4.5.

For $1 \le q < 2 < p \le \infty$ we can get a better estimate: than what the volumetric estimates provide. We first investigate $id : \ell_q^n \to \ell_{q'}^n$. Observe that if AB is a factorization of the identity through ℓ_1^n then B^*A^* is a factorization of id through ℓ_∞^n. A theorem of Maurey (see [22], Theorem 21.4(ii)) asserts that, for every $r > q'$, B^* is r-summing with $\pi_r(B^*) \le C_{q,r}\|B^*\|$ and thus, by the properties of π_r, $\pi_r(id) \le \|A^*\|\pi_r(B^*) \le C_{q,r}\Gamma_n(B_{q'}, B_q)$.

The behavior of the π_r norm of the identity between ℓ_p^n and other spaces was investigated in the two papers [10] and [11]. In particular, in the range we are interested in, it is proved in [10] that $\pi_r(id : \ell_q^n \to \ell_{q'}^n) \geq c_{q,r} n^{1/r}$. (For the interested reader, we found that the best way to understand this is to apply Theorem 1 there to our setup. This is rather easy, as is the proof of Theorem 1.)

This settles the case $p = q'$. Turning to the general case, assume first that $2 \leq q' \leq p < r < \infty$. For any factorization $AB = id_{q\to p}$, $id_{p\to q'}AB$ is a factorization of $id_{q\to q'}$. Therefore, for any $s > q'$,

$$C_{q,r} n^{\frac{1}{s}} \leq \|B\| \|id_{p\to q'}A\| \leq \|A\| \|B\| \|id_{p\to q'}\| \leq \|A\| \|B\| n^{\frac{1}{q'}-\frac{1}{p}},$$

hence $\|A\|\|B\| \geq C_{q,r} n^{\frac{1}{p}+\frac{1}{s}-\frac{1}{q'}}$. Choosing s such that $\frac{1}{r} = \frac{1}{p} + \frac{1}{s} - \frac{1}{q'}$ gives the result in this case.

A similar argument may be used to handle the case $q' > p$. ■

Next we estimate $\Gamma_n(B_p^m, B_q^m)$ when $n \leq m$. Note that the results we obtain are not for the full range of p and q.

Theorem 4.11 *For every integers $n \leq m$ the following holds:*

1. *If $2 \leq q \leq p < \infty$ then $\Gamma_n(B_p^m, B_q^m) \sim_{p,q} n^{1/2} m^{1/p-1/q}$.*
2. *If $q \leq p \leq 2$ then $\Gamma_n(B_p^m, B_q^m) \sim_{p,q} n^{1-1/p} m^{1/p-1/q}$.*
3. *If $p \leq q$ and $1 \leq p \leq 2$ then $\Gamma_n(B_p^m, B_q^m) \sim_{p,q} n^{1-1/q}$.*
4. *If $p \leq q$ and $2 < p < \infty$ then $\Gamma_n(B_p^m, B_q^m) \sim_{p,q} n^{1/2+1/p-1/q}$.*

Proof. In all cases, the lower bound follows from Corollary 4.6 combined with the estimate on the volume numbers of $id_{p\to q}$ in (2.2) and (2.4), and the volume ratios of quotients of ℓ_p^n from Theorem 2.7.

As for the upper bound, the optimal choice in (1) and (2) (at least when n divides m) is the section E spanned by $v_j = \sum_{i=1}^k e_{j+k(i-1)}$, $j = 1, ..., n$. Then $B_p^m \cap E = (m/n)^{1/2-1/p} B_p^n$. Clearly $\Gamma_n(B_p^m, B_q^m) \leq \Gamma_n(B_p^m \cap E, B_q^m \cap E)$ and when $2 \leq q \leq p$ the latter can be approximated using the probabilistic argument from Lemma 4.8 and Lemma 4.7. Indeed, a straightforward computation shows that one can take $\alpha = m^{1/2-1/p}$ and that ε needs to satisfy that $m^{1/2-1/q} = \alpha/n^{1/2}\varepsilon = m^{1/2-1/p}/n^{1/2}\varepsilon$. Thus, $1/\varepsilon \leq n^{1/2} m^{1/p-1/q}$, which proves the bound is tight.

When $q \leq p \leq 2$ one uses the identity operator as the factorizing operator between $(m/n)^{1/2-1/q} B_q^n$ and $(m/n)^{1/2-1/p} B_p^n$ to obtain the required result.

The upper bound in (3) is obtained by taking the canonical section $\text{span}\{e_1, ..., e_n\}$ and applying Theorem 4.9. ■

The information one can obtain from these estimates is summarized in the following

Theorem 4.12 *Let $1 \le p < q \le \infty$, set $F = B_{\ell_{q'}}$ and $\Omega = B_{\ell_p}$. Then,*

1.
$$\mathrm{VC}(\varepsilon, F, \Omega) \sim_{p,q} \begin{cases} \varepsilon^{-\frac{q}{q-1}} & \text{if } 1 \le p \le 2, \\ \varepsilon^{-\frac{1}{\frac{1}{2}+\frac{1}{p}-\frac{1}{q}}} & \text{if } 2 < p \le \infty. \end{cases}$$

2. *F is a uniform Donsker class on Ω.*

3. *For any probability measure μ on B_p, every integer n and every $t > 0$,*
$$Pr\Big\{ \sup_{x^* \in B_{q'}} \big| \mathbb{E}_\mu x^* - \frac{1}{n}\sum_{i=1}^n x^*(X_i) \big| \ge C_{p,q}\Big(\frac{1}{\sqrt{n}} + \frac{t}{n}\Big) \Big\} \le e^{-t},$$
where $(X_i)_{i=1}^n$ are independent and distributed according to μ.

Before presenting the proof, we require an additional lemma which follows from Theorem 1.5. Although the first equality is not needed in the sequel, it might be useful in other applications.

Lemma 4.13 *For any $1 \le p < q \le \infty$ there is a constant $C_{p,q}$ for which the following holds: if $x_1, ..., x_n \in B_p$ and $T : \ell_{q'} \to \ell_2^n$ is given by $Tx^* = n^{-1/2}\sum_{i=1}^n x^*(x_i)e_i$, then, for every $\varepsilon > 0$,*
$$\log N(\varepsilon, TB_{q'}, \ell_2^n) = \log N(\varepsilon, B_{q'}, L_2(\mu_n)) \le C_{p,q}\mathrm{VC}(\varepsilon, B_{q'}, B_p) \cdot \log \frac{2}{\varepsilon},$$
where μ_n is the empirical measure supported on $\{x_1, ..., x_n\}$.

Proof (of Theorem 4.12). The first part of the claim follows from Corollary 4.3 which yields the upper bound, while the lower one follows immediately from Theorem 4.11.

The second part is evident because, by Lemma 4.13, the class has a converging entropy integral, which by Theorem 2.12 suffices to ensure that F is a uniform Donsker class.

Finally, the last part follows from the first, combined with Talagrand's inequality (Theorem 1.3) and the estimate on the expected deviation in terms of the shattering dimension (Theorem 1.6). ∎

5 The shattering dimension of Images of B_1^m

Although the volumetric approach yields sharp results in some cases, and in particular, for $\Gamma_n(B_p^n, B_q^n)$ for certain range of p and q, an exact estimate on the factorization constant $\Gamma_n(TB_1^n, B_q^n)$ does not follow from the volumetric argument. The reason for this is that the position of B_1^n is significant, and not only the volume of the ellipsoid of minimal volume containing TB_1^n. Indeed, we show that spectral information does not suffice for sharp estimates on the shattering dimension. To demonstrate this, given a set of singular values (arranged in a non-increasing order) $\Lambda = (\lambda_1, ..., \lambda_n)$, let $\mathbb{T}_\Lambda$ be the subset of GL_n consisting of the matrices which have Λ as singular values.

Theorem 5.1 *For every set Λ of singular values,*

$$\sup_{T \in \mathbb{T}_\Lambda} \Gamma_n(TB_1^n, B_q^n) = \frac{1}{\lambda_n} \begin{cases} n^{1-\frac{1}{q}} & \text{if } q \geq 2 \\ \\ n^{\frac{1}{2}} & \text{if } q < 2 \end{cases}$$

and

$$\inf_{T \in \mathbb{T}_\Lambda} \Gamma_n(TB_1^n, B_q^n) \sim_q \Big(\sum_{i=1}^n |\lambda_i|^{-2}\Big)^{\frac{1}{2}} \begin{cases} 1 & \text{if } q \geq 2 \\ \\ n^{\frac{1}{2}-\frac{1}{q}} & \text{if } q < 2. \end{cases}$$

To compare this result to the one obtained via the volumetric approach (Theorem 4.2), take $q = 2$, and recall that Theorem 4.2 implies that

$$\Gamma_n(TB_1^n, B_2^n) \geq n^{\frac{1}{2}} (\prod_{i=1}^n \lambda_i^{-2})^{\frac{1}{2n}},$$

which, by the means inequality, is weaker than the conclusion of Theorem 5.1.
Proof. By Lemma 3.1, for every $T \in GL_n$,

$$\Gamma_n(TB_1^n, B_q^n) = \big(\sup\{\varepsilon | \varepsilon B_q^n \subset TB_1^n\}\big)^{-1} = \max_{\|x\|_q = 1} \|x\|_{TB_1^n}.$$

Since $(TB_1^n)^o = T^{-1^*} B_\infty^n$, then for every x,

$$\|x\|_{TB_1^n} = \sup_{y \in (TB_1^n)^o} \langle x, y \rangle = \sup_{y \in B_\infty^n} \langle x, T^{-1^*} y \rangle = \sup_{(\varepsilon_i)_{i=1}^n \in \{-1,1\}^n} \langle T^{-1} x, \sum_{i=1}^n \varepsilon_i e_i \rangle,$$

and

$$\max_{\|x\|_q = 1} \|x\|_{TB_1^n} = \sup_{(\varepsilon_i)_{i=1}^n \in \{-1,1\}^n} \sup_{\|x\|_q = 1} \langle T^{-1} x, \sum_{i=1}^n \varepsilon_i e_i \rangle.$$

By the polar decomposition, $T^{-1} = ODU$ where U and O are orthogonal and D is the diagonal matrix with eigenvalues λ_i^{-1}. Thus

$$\inf_{T\in\mathbb{T}_\Lambda} \max_{\|x\|_q=1} \|x\|_{TB_1^n} = \inf_{O,V\in O_n} \sup_{(\varepsilon_i)_{i=1}^n\in\{-1,1\}^n} \sup_{\|x\|_q=1} \big\langle ODVx, \sum_{i=1}^n \varepsilon_i e_i\big\rangle$$

and

$$\sup_{T\in\mathbb{T}_\Lambda} \max_{\|x\|_q=1} \|x\|_{TB_1^n} = \sup_{O,V\in O_n} \sup_{(\varepsilon_i)_{i=1}^n\in\{-1,1\}^n} \sup_{\|x\|_q=1} \big\langle ODVx, \sum_{i=1}^n \varepsilon_i e_i\big\rangle,$$

where O_n denotes the set of orthogonal matrices on $\mathbb{R}^n$. Set $(\mu_i)_{i=1}^n$ to be the eigenvalues of D arranged in a non-increasing order, that is, $\mu_1 = \lambda_n^{-1} \geq \cdots \geq \mu_n = \lambda_1^{-1}$.

Let $f(O,V) = \max_{(\varepsilon_i)_{i=1}^n\in\{-1,1\}^n} \max_{\|x\|_q=1}\big\langle ODVx, \sum_{i=1}^n \varepsilon_i e_i\big\rangle$, and observe that

$$\begin{aligned} f(O,V) &= \max_{\|x\|_q=1} \max_{(\varepsilon_i)_{i=1}^n} \big\langle x, \sum_{k=1}^n \big(\sum_{j=1}^n V_{jk}\mu_j \sum_{i=1}^n \varepsilon_i O_{ij}\big) e_k\big\rangle \\ &= \max_{(\varepsilon_i)_{i=1}^n} \Big(\sum_{k=1}^n \big|\sum_{j=1}^n \mu_j \big(\sum_{i=1}^n \varepsilon_i O_{ij}\big) V_{jk}\big|^{q'}\Big)^{\frac{1}{q'}}. \end{aligned}$$

Clearly,

$$\begin{aligned} \max_{O,V} f(O,V) &= \max_{O,V} \max_{(\varepsilon_i)_{i=1}^n} \Big(\sum_{k=1}^n \big|\sum_{j=1}^n \mu_j \big(\sum_{i=1}^n \varepsilon_i O_{ij}\big) V_{jk}\big|^{q'}\Big)^{\frac{1}{q'}} \\ &= \max_{V} \max_{\|z\|_2=\sqrt{n}} \Big(\sum_{k=1}^n \big|\sum_{j=1}^n \mu_j z_j V_{jk}\big|^{q'}\Big)^{\frac{1}{q'}} \\ &= \max_{V} \max_{\|x\|_2=\mu_1\sqrt{n}} \|xV\|_{q'} = \mu_1\sqrt{n} \max_{x\neq 0} \frac{\|x\|_{q'}}{\|x\|_2}, \end{aligned}$$

from which the first part of the claim follows.

To prove the second part, note that

$$\begin{aligned} \min_{O,V}\big(f(O,V)\big)^{q'} &\geq \min_{O,V} \mathbb{E}_\varepsilon \sum_{k=1}^n \big|\sum_{j=1}^n \mu_j \big(\sum_{i=1}^n \varepsilon_i O_{ij}\big) V_{jk}\big|^{q'} \\ &= \min_{O,V} \sum_{k=1}^n \mathbb{E}_\varepsilon \big|\sum_{i=1}^n \varepsilon_i \sum_{j=1}^n \mu_j V_{jk} O_{ij}\big|^{q'} = (*). \end{aligned}$$

where $(\varepsilon_i)_{i=1}^n$ are independent Rademacher random variables. Therefore, By Khintchine's inequality,

$$(*) \geq \min_{O,V} C_q \sum_{k=1}^n \Big(\sum_{i=1}^n \big(\sum_{j=1}^n \mu_j V_{jk} O_{ij}\big)^2\Big)^{\frac{q'}{2}}.$$

Denoting $h_k = (\mu_j V_{jk})_{j=1}^n$,

$$\Big(\sum_{i=1}^n \big(\sum_{j=1}^n \mu_j V_{jk} O_{ij}\big)^2\Big)^{\frac{1}{2}} = \|h_k O\|_2 = \|h_k\|_2 = \Big(\sum_{j=1}^n \mu_j^2 V_{jk}^2\Big)^{\frac{1}{2}},$$

and applying Khintchine's inequality again,

$$\Big(\sum_{j=1}^n \mu_j^2 V_{jk}^2\Big)^{\frac{q'}{2}} \geq C_q \mathbb{E}_\varepsilon \Big|\sum_{j=1}^n \varepsilon_j \mu_j V_{jk}\Big|^{\frac{q'}{2}}.$$

By Jensen's inequality and since the matrix $(\varepsilon_j V_{jk})_{j,k=1}^n$ is also orthogonal for any realization of the Rademacher variables,

$$\begin{aligned}\min_{O,V} f(O,V) &\geq C_q \min_V \Big(\mathbb{E}_\varepsilon \sum_{k=1}^n \Big|\sum_{j=1}^n \varepsilon_j \mu_j V_{jk}\Big|^{q'}\Big)^{1/q'} \geq C_q \mathbb{E}_\varepsilon \min_V \Big(\sum_{k=1}^n \Big|\sum_{j=1}^n \varepsilon_j \mu_j V_{jk}\Big|^{q'}\Big)^{1/q'} \\ &= C_q \min_V \|\mu V\|_{q'},\end{aligned}$$

and

$$\min_V \|\mu V\|_{q'} = \|\mu\|_2 \begin{cases} 1 & \text{if } q' \leq 2 \\ n^{1/q'-1/2} & \text{if } q' > 2. \end{cases}$$

Finally, to see that the lower bound is tight, set $O = id$, and thus

$$f(id, V) = \max_{(\varepsilon_i)_{i=1}^n} \Big\|\sum_{i=1}^n \varepsilon_i (V^*\mu)_i e_i\Big\|_{q'} = \|V^*\mu\|_{q'}.$$

The sharpness is evident by optimizing with respect to V.

■

6 Application - kernel machines

Thus far, we have investigated the shattering dimension of classes of linear functionals. Here, we use the methods developed in previous sections to bound the shattering dimension of classes that can be represented as linear functionals. In principle, our analysis holds for any class of functions that can be embedded in a "nice" Banach space X, such that the set of point evaluation functionals $\{\delta_\omega | \omega \in \Omega\}$ is a bounded subset of X^*. We chose to present the case of kernel classes because of the central role these classes have in the modern theory of Machine Learning, though similar results can be derived for many other classes (e.g. unit balls in certain Sobolev spaces).

To define the function class we are interested in, assume that Ω is a compact set and let $K : \Omega \times \Omega \to \mathbb{R}$ be a positive definite, continuous function, which we assume to be bounded by 1. Let ν be a probability measure on Ω, and consider the integral operator $T_K : L_2(\nu) \to L_2(\nu)$ given by $T_K f = \int K(x,y) f(y) d\nu(y)$. By the Spectral Theorem, T_K has a diagonal representation, that is, there exists a complete, orthonormal basis of $L_2(\nu)$, which is denoted by $\big(\phi_n(x)\big)_{n=1}^{\infty}$, and a non-increasing non-negative sequence of eigenvalues $(\lambda_n)_{i=1}^{\infty}$ which satisfy that for every sequence $(a_n) \in \ell_2$, $T_K(\sum_{n=1}^{\infty} a_n \phi_n) = \sum_{n=1}^{\infty} a_n \lambda_n \phi_n$. Mercer's Theorem implies that $\nu \times \nu$-almost surely, $K(x,y) = \sum_{n=1}^{\infty} \lambda_n \phi_n(x) \phi_n(y)$. Thus, under a mild assumption on ν, this representation holds for every $x, y \in \Omega$.

Let F_K be the class consisting of all the functions $\sum_{i=1}^{\infty} a_i K(x_i, \cdot)$, where $x_i \in \Omega$ and $a_i \in \mathbb{R}$ such that $\sum_{i,j=1}^{\infty} a_i a_j K(x_i, x_j) \leq 1$.

One can show that F_K is the unit ball of a Hilbert space associated with the kernel, called the reproducing kernel Hilbert space, and we denote it by $\mathcal{H}$. We refer the reader to [6, 21] for the basic facts about reproducing kernel Hilbert spaces and their application to Learning Theory.

The inner product in $\mathcal{H}$ satisfies that for every $f \in \mathcal{H}$, $\langle f, K(x, \cdot) \rangle_{\mathcal{H}} = f(x)$. Hence, every function can be viewed as a linear functional on $\mathcal{H}$, or more precisely, on $\{K(x, \cdot) | x \in \Omega\} \subset \mathcal{H}$.

An alternative way to define the reproducing kernel Hilbert space which makes this interpretation clearer is via the *feature map*. Define $\Phi : \Omega \to \ell_2$ by $\Phi(x) = \big(\sqrt{\lambda_i}\phi_i(x)\big)_{i=1}^{\infty}$. Then,

$$F_K = \big\{ f(\cdot) = \langle \beta, \Phi(\cdot) \rangle_{\mathcal{H}} \big| \|\beta\|_2 \leq 1 \big\}.$$

In other words, the feature map is a way of embedding the space Ω in the reproducing kernel Hilbert space, and then the class F_K is a class of linear functionals on the image of the Ω via the feature map, denoted by $\Phi(\Omega)$.

This observation is significant because the shattering dimension of F_K as a class of functions on Ω is equal to the shattering dimension of B_2, the unit ball of ℓ_2, when considered as a set of functionals on $\Phi(\Omega)$. Moreover, when restricted to $\{x_1, ..., x_n\}$, the shattering dimension of F_K is $\mathrm{VC}(\varepsilon, B_2, TB_1^n)$, where $T\ell_1^n \to \ell_2$ and $Te_i = \Phi(x_i)$.

Recall that we have obtained bounds on the shattering dimension in terms of the eigenvalues of T, or in terms of the volume of the ellipsoid of minimal volume containing TB_1^n. Since $\langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} = K(x,y)$ for every $x, y \in \Omega$, one can control the shattering dimension using the eigenvalues of $\sqrt{T^*T}$, which is the square-root of the Gram matrix $\big(K(x_i, x_j)\big)_{i,j=1}^{n}$.

Another family of classes of functions which will be of interest to us are the so-called ℓ_p machines

$$F^p(y_1, ..., y_m) = \big\{ f(x) = \sum_{i=1}^m \alpha_i K(y_i, x) \big| \sum_{i=1}^m |\alpha_i|^p \leq 1, \ \|f\|_\infty \leq 1 \big\}.$$

The case $p = 1$ is particularly interesting in practical applications because, by convexity, the L_∞ constraint on the functions holds automatically.

6.1 The shattering dimension of kernel machines

In this section we apply the volumetric approach and investigate the shattering dimension of kernel machines. In most cases, the bounds we present are given in terms of the geometric mean of the eigenvalues of Gram matrices. The difficulty arises since the bounds will be in a "worst case scenario" - the supremum with respect to all Gram matrices of a given dimension, making it difficult to compute. Therefore, we present some bounds which depend on the eigenvalues of the integral operator, under an additional assumption on the kernel. Throughout this section, we denote by ℓ_2 the reproducing kernel Hilbert space.

6.1.1 The Shattering Dimension of F_K

Assume that $S = \{x_1, ..., x_n\} \subset \Omega$. We wish to estimate $\mathrm{VC}(\varepsilon, F_K, S)$ in terms of the eigenvalues of the Gram matrix $\big(K(x_i, x_j)\big)_{i,j=1}^n$.

For every $S = \{x_1, ..., x_n\}$ let P_S be the orthogonal projection in the reproducing kernel Hilbert space onto the space spanned by $\{\Phi(x_1), ..., \Phi(x_n)\}$. Clearly, S is ε-shattered by F_K if and only if $\Phi(S)$ is ε-shattered by $P_S(B_2) = B_2^n$, and thus, $\mathrm{VC}(\varepsilon, F_K, S) = \mathrm{VC}\big(\varepsilon, B_2^n, \Phi(S)\big)$. Let $T_S : \ell_1^n \to \ell_2^n$ be given by $T_S e_i = \Phi(x_i)$. Using the volumetric approach (Theorem 4.2), we may take $K = TB_1^n$ and $L^o = B_2^n$. Hence, if $\{x_1, ..., x_n\}$ is ε-shattered by F_K then $n \leq C\varepsilon^{-2}(\prod_{i=1}^n \lambda_i^2)^{1/n}$, where $(\lambda_i)_{i=1}^n$ are the singular values of T_S. Therefore,

$$\mathrm{VC}(\varepsilon, F_K, \{X_1, ..., X_n\}) \leq C\varepsilon^{-2}(\prod_{i=1}^n \lambda_i^2)^{1/n}, \tag{6.1}$$

and

$$\mathrm{VC}(\varepsilon, F_K, \Omega) \leq C \sup_{T_S} \varepsilon^{-2}(\prod_{i=1}^n \lambda_i^2)^{1/n},$$

where the supremum is taken with respect to all the injective linear operators which map $(e_i)_{i=1}^n$, into $\Phi(\Omega)$.

However, in this case one can use the more direct approach, which yields the optimal bound in terms of the singular values of the operator T_S.

The following is evident from Theorem 5.1.

Corollary 6.1 *There is an absolute constant C such that for any kernel K and every $\varepsilon > 0$,*

$$\mathrm{VC}(\varepsilon, F_K, \Omega) \leq C \sup\Big\{ n | \exists s_n = \{x_1, ..., x_n\} \ \Big(\frac{1}{\sum_{i=1}^n \theta_i^{-2}(s_n)}\Big)^{\frac{1}{2}} \geq \varepsilon \Big\},$$

where $(\theta_i(s_n))_{i=1}^n$ are the eigenvalues of the Gram matrix $\big(K(x_i, x_j)\big)_{i,j=1}^n$.

6.2 The shattering dimension of ℓ_p machines

Next, we provide an estimate on the shattering dimension of ℓ_p machines. Unlike the previous case, we will see that the number of the basis functions which form the class influences the bounds we have.

Theorem 6.2 *There is an absolute constant C for which the following holds. Let K be a kernel and let $1 \leq p \leq 2$. If $\{x_1, ..., x_n\}$ is ε-shattered by $F^p(y_1, ..., y_m)$ then $n \leq C\varepsilon^{-p}\big(\prod_{i=1}^n \theta_i^{(1)}\theta_i^{(2)}\big)^{p/2n}$, where $(\theta_i^{(1)})_{i=1}^n$ are the eigenvalues of $\big(K(x_i, x_j)\big)_{i,j=1}^n$ and $(\theta_i^{(2)})_{i=1}^m$ are the n largest eigenvalues of $\big(K(y_i, y_j)\big)_{i,j=1}^m$.*

Proof: Fix an integer m and a set $\{y_1, ..., y_m\}$. By the reproducing kernel property, if $s_n = \{x_1, ..., x_n\}$ is ε-shattered by $\mathcal{H} = F^p(y_1, ..., y_m)$ then the same holds for $\{\Phi(x_1), ..., \Phi(x_n)\}$. Define $T_1 : \ell_1^n \to \ell_2$ by $T_1 e_i = \Phi(x_i)$ and $T_2 : \ell_2^m \to \ell_2$ by $T_2 e_i = \Phi(y_i)$. Set $Q = T_2 id_{\ell_p^m \to \ell_2^m}$ and note that $\mathcal{H} \subset QB_p^m$. Therefore,

$$\mathrm{VC}\big(\varepsilon, \mathcal{H}, s_n\big) \leq \mathrm{VC}\big(\varepsilon, QB_p^m, \{T_1e_1, ..., T_1e_n\}\big) = \mathrm{VC}\big(\varepsilon, B_p^m, \{Q^*T_1e_1, ..., Q^*T_1e_n\}\big).$$

Let $T = Q^*T_1 : \ell_1^n \to \ell_{p'}^m$ and set $E = T\ell_1^n$. Using the notation of Theorem 4.2, $K = TB_1^n$ and $L^o = B_p^m$. Thus,

$$\sqrt{n} \leq C\varepsilon^{-1}\big(|\tilde{\mathcal{E}}_K| / |L \cap E|\big)^{1/n}.$$

If Q_E^* is the restriction of Q^* to E and since $\tilde{\mathcal{E}}_K = Q_E^* T_1 B_2^n$ then

$$|\tilde{\mathcal{E}}_K|^{1/n} \leq cn^{-1/2}\big(\prod_{i=1}^n \lambda_i^1 \lambda_i^2\big)^{1/n},$$

where $(\lambda_i^j)_{i=1}^n$, $j = 1, 2$ are the singular values of T_j. By the Meyer-Pajor Theorem, $|B_{p'}^m \cap E| \geq |B_{p'}^n|$. Hence,

$$n \leq C\varepsilon^{-p}\big(\prod_{i=1}^n \theta_i^{(1)}\theta_i^{(2)}\big)^{\frac{p}{2n}},$$

where $(\theta_i^{(1)})_{i=1}^n$ are the eigenvalues of $\big(K(x_i, x_j)\big)_{i,j=1}^n$ and $(\theta_i^{(2)})_{i=1}^n$ are the n largest eigenvalues of $\big(K(y_i, y_j)\big)_{i,j=1}^m$, as claimed. ∎

6.3 Global Assumptions

As we mentioned in previous sections, one important aspect in the volumetric approach is to find a bound on the volume of the ellipsoid of minimal volume containing TB_1^n, where $T : \ell_1^n \to \ell_2$ maps e_i to $\Phi(x_i)$.

In general, there is no apriori bound on this ellipsoid. For example, if Ω is a finite set $\{x_1, ..., x_n\}$ and $K(x,y) = \delta_{x,y}$ then $F_K = \tilde{\mathcal{E}}_{\Phi(\Omega)} = B_2^n$. In many interesting examples, the spectrum of K as an integral operator decays rapidly, and there is some control on the ellipsoid containing $\Phi(\Omega)$. To that end, we assume that the eigenfunctions of the integral operator are uniformly bounded, which implies that there is a relatively small ellipsoid $\mathcal{E} \subset \ell_2$ such that $\Phi(\Omega) \subset \mathcal{E}$.

Lemma 6.3 *[6, 26] Assume that $K(x,y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(y)$ for every $x, y \in \Omega$ and that the eigenvalues $(\phi_i)_{i=1}^{\infty}$ of T_K are all bounded by 1. Set $(a_n)_{n=1}^{\infty} \in \ell_2$ to be such that $(b_n)_{n=1}^{\infty} = (\sqrt{\lambda_n}/a_n)_{n=1}^{\infty} \in \ell_2$ and put $R = \|(b_n)\|_{\ell_2}$. If $A : \ell_2 \to \ell_2$ is defined by $Ae_i = Ra_i e_i$, and if $\mathcal{E} = A\big(B(\ell_2)\big)$, then $\Phi(\Omega) \subset \mathcal{E}$.*

We can apply our results (using the notation of the previous results) and obtain the following estimates on the shattering dimension

Theorem 6.4 *There is an absolute constant C such that the following holds: for every kernel which satisfies the assumption of Lemma 6.3 and any sample s_n which is ε-shattered by F_K, $\varepsilon \leq C\big(\prod_{i=1}^n Ra_i\big)^{1/n}/\sqrt{n}$. If s_n is ε-shattered by $F^1(y_1, ..., y_m)$ then $\varepsilon \leq C\big(\prod_{i=1}^n Ra_i\big)^{2/n}/\sqrt{n}$.*

Proof: In both cases we shall use Theorem 4.2. Let $\mathcal{E} \subset \ell_2$ be the ellipsoid containing $\Phi(\Omega)$ with $(Ra_i)_{i=1}^{\infty}$ as axes. Then, $\mathrm{VC}(\varepsilon, F_K, \Omega) \leq \mathrm{VC}(\varepsilon, B_2, \mathcal{E})$ for every $\varepsilon > 0$. Take $K = \mathcal{E}$ and $L^o = B_2$. Since K is an ellipsoid then for every subspace E, $K \cap E$ is also an ellipsoid and thus $\mathrm{vr}\big((K \cap E)^o\big) = 1$. Moreover, if $\dim(E) = n$ then

$$|\mathcal{E} \cap E|^{1/n} \leq Cn^{-1/2}\big(\prod_{i=1}^{n} Ra_i\big)^{1/n}$$

and the first part of our claim follows. As for the second, by convexity $F^1(y_1, ..., y_m) \subset \mathrm{conv}\big(\Phi(\Omega)\big) \subset \mathcal{E}$. Hence, $\mathrm{VC}(\varepsilon, F_1, \Omega) \leq \mathrm{VC}(\varepsilon, \mathcal{E}, \mathcal{E})$. As an ellipsoid,

$$|(\mathcal{E}^o \cap E)|^{-\frac{1}{n}} = \frac{|P_E \mathcal{E}|^{\frac{1}{n}}}{|B_2^n|^{\frac{2}{n}}},$$

and the latter is smaller than $C\sqrt{n}\big(\prod_{i=1}^n Ra_i\big)^{1/n}$ for some absolute constant C. Our claim now follows from Theorem 4.2. ■

One natural case in which the global assumptions can be used is for translation invariant kernels on compact abelian groups. We will present a very simple example, which can be easily extended.

Let $\Omega = S^1$ and put K to be a continuous, translation invariant kernel (i.e. $K(x,y) = k(x-y)$) which is positive definite. Set ν to be the Haar measure on Ω and let T_K be the integral operator associated with ν and K. Thus, by the properties of the Haar measure, the diagonal representation for K holds for any pair (x,y) and by translation invariance, the eigenfunctions of the integral operator are bounded by 1. In particular, it is easy to check that if there are $B > 0$ and $\alpha > 1$ such that $\lambda_n \leq B/n^{\alpha}$, then F_K is a universal Donsker class.

References

[1] M. Anthony, P.L. Bartlett *Neural Network Learning, Theoretical Foundations*, Cambridge University Press, 1999.

[2] N. Alon, S. Ben–David, N. Cesa–Bianchi, D. Haussler: Scale sensitive dimensions, uniform convergence and learnability, J. of ACM 44 (4) 615–631, 1997.

[3] K. Ball: Volumes of sections of cubes and related problems, in *Lecture Notes in Mathematics 1470*, 36-47, Springer-Berlin.

[4] P.L. Bartlett, O. Bousquet, S. Mendelson: Localized Rademacher averages, preprint.

[5] O. Bousquet: A Bennett concentration inequality and its application to suprema of empirical processes, C. R. Acad. Sci. Paris, Ser. I, 334, 495-500, 2002. .

[6] P. Cucker, S. Smale: On the mathematical foundations of learning, Bull. AMS 39(1) 1–49, 2002.

[7] R.M. Dudley: *Uniform Central Limit Theorems*, Cambridge Studies in Advanced Mathematics 63, Cambridge University Press 1999.

[8] R.M. Dudley, E. Giné, J. Zinn: Uniform and universal Glivenko–Cantelli classes, J. Theoret. Prob. 4, 485–510, 1991.

[9] E. Giné, J. Zinn: Gaussian characterization of uniform Donsker classes of functions, Ann. Probab. 19(2), 758–782, 1991.

[10] E.D. Gluskin: Estimates of the norms of certain p-absolutely summing operators, (Russian) Funktsional. Anal. i Prilozhen. 12(2) 24–31, 1978. Translated in Funct. Anal. Appl. 12, 94–101, 1978.

[11] E.D. Gluskin, A. Pietsch, J. Puhl: A generalization of Khintchine's inequality and its application in the theory of operator ideals. Studia Math. 67(2), 149–155 (1980).

[12] Y. Gordon, M. Junge: Volume formulas in L_p spaces, Positivity 1, 7–43, 1997.

[13] Y. Gordon, M. Junge: Volume ratios in L_p spaces, Studia Math. 136(2), 147–182, 1999.

[14] S. Mendelson: Learnability in Hilbert spaces with Reproducing Kernels, Journal of Complexity, 18(1), 152-170, 2002.

[15] S. Mendelson: Rademacher averages and phase transitions in Glivenko-Cantelli class, IEEE transactions on information theory, 48(1), 251-263, 2002.

[16] S. Mendelson, R. Vershynin, Entropy and the combinatorial dimension, Invent. Math. to appear.

[17] S. Mendelson: A few notes on Statistical Learning Theory in *Proceedings of the Machine Learning Summer School, Canberra 2002, S. Mendeslon and A.J. Smola (Eds.)*, to appear in Lecture notes in computer sciences, Springer 2003.

[18] M. Meyer, A. Pajor: Sections of the unit ball of ℓ_p^n, J. Func. Anal. 80 (1), 109-123, 1988.

[19] V.D. Milman, G. Schechtman: *Asymptotic theory of finite dimensional normed spaces*, Lecture Notes in Mathematics 1200, Springer 1986.

[20] G. Pisier: *The volume of convex bodies and Banach space geometry*, Cambridge University Press, 1989.

[21] B. Schölkopf, A.J. Smola: *Learning with kernels*, MIT press, 2001.

[22] N. Tomczak–Jaegermann: *Banach–Mazur distance and finite–dimensional operator Ideals*, Pitman monographs and surveys in pure and applied Mathematics 38, 1989

[23] M. Talagrand: Sharper bounds for Gaussian and empirical processes, Ann. Probab. 22(1), 28-76, 1994.

[24] A.W. Van der Vaart, J.A. Wellner: *Weak convergence and Empirical Processes*, Springer-Verlag, 1996.

[25] V. Vapnik, A. Chervonenkis: Necessary and sufficient conditions for uniform convergence of means to mathematical expectations, Theory Prob. Applic. 26(3), 532-553, 1971.

[26] R.C. Williamson, B. Scheolkopf, A. Smola: Generalization performance of regularization networks and support vector machines via entropy numbers of compact networks, IEEE trans. Info. Theory 47(6), 2516-2532, 2001.