Journal of Machine Learning Research 2 (2001) 1-18 Submitted 2/01; Published 10/01

On the size of convex hulls of small sets

Shahar Mendelson SHAHAR@CSL.ANU.EDU.AU
Computer Sciences Laboratory
Research School of Information Sciences and Engineering
The Australian National University
Canberra 0200, Australia

Editor: Dana Ron

Abstract

We investigate two different notions of "size" which appear naturally in Statistical Learning Theory. We present quantitative estimates on the *fat-shattering dimension* and on the covering numbers of convex hulls of sets of functions, given the necessary data on the original sets. The proofs we present are relatively simple since they do not require extensive background in convex geometry.

Keywords: Convex hulls, fat-shattering dimension, covering numbers

1. Introduction

Convexity plays an important role in Machine Learning. Its significance can be seen in both sides of Learning Theory. Firstly, from the practitioner's point of view, minimizing empirical risks is much easier when the class is convex. Secondly, from the theoretical standpoint, the sample complexity needed for agnostic learning is considerably smaller for convex classes (Lee et al., 1998, Mendelson, 2001b). On the other hand, if the original class one is interested in happens to be non-convex, taking the convex hull increases the size of the class and it is not clear whether this is worth the effort, since the benefits of learning from a convex class might be negligible compared to the price one has to pay for using the much larger class. Thus, it is natural to ask "how large" can a convex hull of a given class be? We shall present an answer with respect to two important parameters which measure the size of a class: the covering numbers and the fat-shattering dimension. The estimates on the covering numbers we present are not new and have recently appeared in Carl et al. (1999). The main reason we chose to present an alternative proof is because the one in Carl et al. (1999) uses very deep results in the local theory of Banach spaces, hence it is less accessible to the non expert reader. The proof we present here is self contained and (almost) does not assume any prior knowledge.

The upper bound on the fat-shattering dimension of the convex hull of a class uses a notion originating from Banach spaces theory called *type*. This is a property of some Banach spaces which appears naturally when trying to compute the fat-shattering dimension of linear functionals (Gurvits, 2001, Mendelson, 2001a). The path we take is as follows: we begin by improving the known upper bounds on the fat-shattering dimension of the functionals $\{x^* | \|x^*\| \leq 1\}$ (defined below) when considered as functions on the unit ball of the space. It turns out that this linear fat-shattering dimension is determined by the *type*

of the Banach space. We use this fact to prove the results concerning the fat-shattering dimension of a convex hull of a general class. To that end, we embed both the class and the shattered set into two finite dimensional Banach spaces, where the dimension of the spaces is the size of the shattered set. The unit ball of the first space is the symmetric convex hull of the n-tuples $\big(f(\omega_i)\big)_{i=1}^n$ where $f \in F$ and $(\omega_i)_{i=1}^n$ is the shattered set. The second Banach space is the dual space to the first. It is possible to show that the fat-shattering dimension of a given class may be controlled by the fat-shattering dimension of the embedded class, only now, its members are considered as linear functionals on the images of $(\omega_i)_{i=1}^n$. Thus, it is possible to bound the fat-shattering dimension using "linear" methods.

The article is divided into two main sections. In section 2 we prove the covering numbers estimate. Section 3 is devoted to the investigation of the linear fat-shattering dimension and the fat-shattering of convex hulls.

1.1 Basic results, definitions and notation

Given a Banach space X, the *dual* of X, denoted by X^*, consists of all the bounded linear functionals on X, endowed with the norm

$$\|x^*\|_{X^*} = \sup_{\|x\|_X=1} |x^*(x)| .$$

We denote the unit ball of X by $B(X)$ and the dual unit ball by $B(X^*)$. If $1 \leq p < \infty$, let ℓ_p^n be $\mathbb{R}^n$ equipped with the norm $\|x\|_p = \big(\sum_{i=1}^n |x_i|^p\big)^{1/p}$ and set ℓ_∞^n to be $\mathbb{R}^n$ with respect to the sup norm.

Given a set A, let A^c be its complement, set $|A|$ to be its cardinality and denote its characteristic function by χ_A. Thus, $\chi_A(x) = 1$ if $x \in A$ and 0 otherwise. If A and B are sets, let $A + B = \{a + b | a \in A, b \in B\}$.

For any probability measure μ on a measurable space (Ω, Σ), let $\mathbb{E}_\mu$ denote the expectation with respect to μ. $L_p(\mu)$ is the set of functions which satisfy $\mathbb{E}_\mu |f|^p < \infty$ and set $\|f\|_{L_p(\mu)} = (\mathbb{E}_\mu |f|^p)^{1/p}$. Given $I \subset \Omega$, $L_\infty(I)$ is the space of bounded functions on I, with respect to the norm $\|f\|_\infty = \sup_{\omega \in I} |f(\omega)|$. For every $\omega \in \Omega$ let δ_ω be the point evaluation functional, that is, for every function f on Ω, $\delta_\omega(f) = f(\omega)$. We denote by μ_n an empirical measure supported on a set of n points, hence, $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{\omega_i}$. If $|I| = n$ and μ_n is the empirical measure supported on I, we denote $L_\infty(I)$ by $L_\infty(\mu_n)$.

Throughout this paper, all absolute constants are denoted by C or c. Their values may change from line to line or even within the same line.

The following are the definitions of well known combinatorial parameters which are often used in Learning Theory.

Definition 1.1 *Let F be a class of $\{0,1\}$-valued functions on a space Ω. We say that F shatters $\{\omega_1, ..., \omega_n\} \subset \Omega$, if for every $I \subset \{1, ..., n\}$ there is a function $f_I \in F$ for which $f_I(\omega_i) = 1$ if $i \in I$ and $f_I(\omega_i) = 0$ if $i \notin I$. Let*

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

$VC(F, \Omega)$ is called the VC dimension of F.

There is a parametric version of the VC dimension, called the fat-shattering dimension:

Definition 1.2 *For every $\varepsilon > 0$, a set $A = \{\omega_1, ..., \omega_n\} \subset \Omega$ is said to be ε–shattered by F if there is some function $s : A \to \mathbb{R}$, such that for every $I \subset \{1, ..., n\}$ there is some $f_I \in F$ for which $f_I(\omega_i) \geq s(\omega_i) + \varepsilon$ if $i \in I$, and $f_I(\omega_i) \leq s(\omega_i) - \varepsilon$ if $i \notin I$. Let*

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

The set $\big(s_i\big) = \big(s(\omega_i)\big)$ is called a witness to the shattering and for every $I \subset \{1, ..., n\}$ we call f_I the shattering function of the set I. In cases where the set Ω is clear, we will denote the fat-shattering dimension by $\mathrm{fat}_\varepsilon(F)$.

If (X, d) is a metric space and if $F \subset X$, let $N(\varepsilon, F, d)$ be the minimal number of open balls with radius $\varepsilon > 0$ (with respect to the metric d) needed to cover F. The numbers $N(\varepsilon, F, d)$ are called the covering numbers of F. A set $A \subset X$ is said to be an ε-cover of F if the union of open balls $\bigcup_{a \in A} B(a, \varepsilon)$ contains F. In cases where the subset F is obvious, we denote the covering numbers by $N(\varepsilon, d)$. In cases where the metric is clear we denote the covering numbers by $N(\varepsilon, F)$. The logarithm of the covering numbers of a set is sometimes called the *entropy* of the set.

A set is called ε-separated if the distance between any two elements of the set is larger than ε. Set $D(\varepsilon, F)$ to be the maximal cardinality of an ε-separated set in F. It is easy to see that $N(\varepsilon, F) \leq D(\varepsilon, F) \leq N(\varepsilon/2, F)$.

The following result, which is due to Alon, Ben-David, Cesa-Bianchi, and Haussler (1997), enables one to estimate the $L_\infty(\mu_n)$ covering numbers of classes in terms of the fat-shattering dimension.

Theorem 1.3 *Let F be a class of functions from Ω into $[0, 1]$ and set $d = \mathrm{fat}_{\varepsilon/4}(F)$. Then, for every empirical measure μ_n on Ω,*

$$D\big(\varepsilon, F, L_\infty(\mu_n)\big) \leq 2\Big(\frac{4n}{\varepsilon^2}\Big)^{d \log\left(en/(d\varepsilon)\right)}.$$

1.2 The main results

We end this introduction with a summary of the main results presented in the sequel which are relevant in the context of Machine Learning.

In the next section, we present a general estimate on the L_2 covering numbers of convex hulls of a class, given the covering numbers of the class. As a corollary of that result we obtain the following:

Theorem 1.4 *Let F be a class of functions on Ω.*

1. *There is an absolute constant C such that for every VC class of functions F, every probability measure μ on Ω and every $\varepsilon > 0$,*

$$\log N\big(\varepsilon, \mathrm{conv}(F), L_2(\mu)\big) \leq Cd\Big(\frac{1}{\varepsilon}\Big)^{\frac{2d}{d+1}},$$

where $\mathrm{VC}(F) = d$.

2. *If F maps Ω into $[0,1]$ and $\mathrm{fat}_\varepsilon(F) \leq \gamma\varepsilon^{-p}$ for some $p<2$ and $\gamma>0$, then for every $p<p'<2$ there is a constant $C=C(p,p',\gamma)$, such that for every probability measure on Ω and every $\varepsilon>0$,*

$$\log N\big(\varepsilon, \mathrm{conv}(F), L_2(\mu)\big) \leq \frac{C}{\varepsilon^2}\log^{1-\frac{2}{p'}}\frac{1}{\varepsilon}.$$

In the final section we investigate the fat-shattering dimension of convex hulls. In particular, we prove the following:

Theorem 1.5 *There is an absolute constant C, such that for every class F which consists of functions which map Ω into $[0,1]$ and every $\varepsilon>0$,*

$$\mathrm{fat}_\varepsilon\big(\mathrm{conv}(F)\big) \leq C\frac{\mathrm{fat}_{\frac{\varepsilon}{4}}(F)}{\varepsilon^2}\log^2\Big(\frac{2\mathrm{fat}_{\frac{\varepsilon}{4}}(F)}{\varepsilon}\Big).$$

2. Entropy of convex hulls of classes

In this section we provide estimates on the entropy of convex hulls of classes when considered as subsets of L_2 spaces. We divide our discussion into two parts. First, we deal with classes of $\{0,1\}$-valued functions which have a finite VC dimension. Then, we investigate classes of functions with a uniformly bounded range for which the fat-shattering dimension is polynomial in ε^{-1}.

The path we take is rather general. We estimate the covering numbers of a convex hull of a set, given that the covering numbers of the set itself are polynomial in ε^{-1}. We combine this general result with well known bounds on the covering numbers of classes using the fat-shattering dimension, thus obtaining the desired entropy estimate.

2.1 General Estimates

We shall investigate two generic cases. The first is when $N\big(\varepsilon,F,L_2(\mu)\big)=O(\varepsilon^{-p})$ for some $p>0$, and the second, (which is more difficult), when $\log N\big(\varepsilon,F,L_2(\mu)\big)=O(\varepsilon^{-p})$ for $0<p<2$. The first case was investigated by Dudley (1987). He showed that for every $\delta>0$, the log-covering numbers of the convex hull of F are polynomial in ε^{-1} with exponent $\delta+2p(p+2)^{-1}$. This result was improved independently by van der Vaart and Wellner (1996) and by Carl (1997) who removed the superfluous δ. Those results indicate that if $N\big(\varepsilon,F,L_2(\mu)\big)\leq\gamma\varepsilon^{-p}$ for some $\gamma>0$ and $p>0$, and if K is the symmetric convex hull of F, then the entropy integral $\int_0^1\log^{1/2}N\big(\varepsilon,K,L_2(\mu)\big)d\varepsilon$ converges. This fact is very significant, since this integral measures in some sense how "large" the class is. For example, classes with bounded entropy integrals satisfy the *uniform central limit theorem* (see Dudley, 1999).

On the other hand, from the quantitative point of view, there were no estimates on the constant $C=C(\gamma,p)$ for which

$$\log N\big(\varepsilon,K,L_2(\mu)\big)\leq C\Big(\frac{1}{\varepsilon}\Big)^{\frac{2p}{p+2}}.$$

Note that from the Machine Learning point of view, the constant is significant, since for VC classes the exponent p is half the VC dimension of the class. Hence, it is only natural to

try and find the way in which the constant depends on p. Another natural question which Dudley's assertion raises is whether similar results may be obtained even when the covering numbers are considerably larger. For example, does the entropy integral of the symmetric convex hull of F converge if $\log N\big(\varepsilon, F, L_2(\mu)\big) = O(\varepsilon^{-p})$ for some $0 < p < 2$. We present partial answers to both questions.

Theorem 2.1 *Let μ be a probability measure on Ω. Assume that F is a subset of the unit ball of $L_2(\mu)$ and set K to be the symmetric convex hull of F.*

1. *If there are $\gamma, p > 0$ such that $N\big(\varepsilon, F, L_2(\mu)\big) \leq \gamma\varepsilon^{-p}$ for every $\varepsilon > 0$, then there is an absolute constant C such that for every $\varepsilon > 0$,*

$$\log N\big(\varepsilon, K, L_2(\mu)\big) \leq C\gamma^{\frac{2}{p}} p\Big(\frac{1}{\varepsilon}\Big)^{\frac{2p}{2+p}}.$$

2. *If there are $\gamma > 0$ and $0 < p < 2$ such that $\log N\big(\varepsilon, F, L_2(\mu)\big) \leq \gamma\varepsilon^{-p}$ for every $\varepsilon > 0$, then there is some constant $C(p,\gamma)$ (which depends only on p and γ), such that*

$$\log N\big(\varepsilon, K, L_2(\mu)\big) \leq C(p,\gamma)\frac{1}{\varepsilon^2}\log^{1-\frac{2}{p}}\frac{1}{\varepsilon}.$$

In particular, the entropy integral of K converges for $p < 2/3$.

An estimate on the growth rate of the entropy numbers recently appeared in Carl et al. (1999), using very deep results in the local theory of Banach spaces. The significance of the proof we present here is the fact that it does not use the powerful machinery of convex geometry, hence, it is more accessible.

We present a complete proof of the second claim. The proof of the first assertion follows the same path and some of the details are omitted. Both proofs are based on an idea which was used in van der Vaart and Wellner (1996, pg. 142), and in fact, the first claim follows from a careful analysis of the proof in van der Vaart and Wellner (1996).

We shall require three preliminary results. The first result is due to B. Maurey, appeared in (Pisier, 1981), and is interesting by itself.

Lemma 2.2 *Let $F \subset L_2(\mu)$ be a set of n functions and denote its diameter by* $\mathrm{diam}(F)$. *Then, for every $\varepsilon > 0$,*

$$N\big(\varepsilon\mathrm{diam}(F), \mathrm{conv}(F), L_2(\mu)\big) \leq (e + en\varepsilon^2)^{\frac{2}{\varepsilon^2}}.$$

Let $(\delta_i)_{i=1}^{\infty}$ be a positive sequence decreasing to 0 and set F_i to be an increasing family of sets, such that for every i, F_i is a δ_i-separated set in F which is maximal with respect to inclusion (that is, each F_i is δ_i-separated and if $F_i \subset A \subset F$ then A is not δ_i-separated). Note that in our case, each one of the sets F_i is finite.

For every $i > j$ and every $x \in F_i$, let $P_j x$ be a member of F_j which is nearest to x and put $G_{ij} = \{x - P_j x \mid x \in F_i\}$.

Lemma 2.3 *For every $i > j \geq 1$ and every $\varepsilon, \varepsilon' > 0$,*

$$\mathrm{conv}(F_i) \subset \mathrm{conv}(F_j) + \mathrm{conv}(G_{ij} \cup \{0\}),$$

and

$$N(\varepsilon, \mathrm{conv}(F_i)) \leq \Big(1+\frac{4\varepsilon'}{\varepsilon}\Big)^{|F_j|} N\big(\varepsilon', \mathrm{conv}(F_j)\big)\Big(e + e\,|F_i|\,\frac{\varepsilon^2}{16\delta_j^2}\Big)^{32\delta_j^2/\varepsilon^2}.$$

Proof: Clearly, the cardinality $|G_{ij}\cup\{0\}| \leq |F_i| - |F_j| + 1$ and, since F_j is a maximal δ_j-separated set in F, then for every $x \in F$, $\|x - P_j x\| \leq \delta_j$.

Set $F_i = \{x_1, ..., x_{|F_i|}\}$ and $F_j = \{x_1, ..., x_{|F_j|}\} \subset F_i$. If $z \in \mathrm{conv}(F_i)$, then there are $\lambda_i \geq 0$ such that $\sum_{i=1}^{|F_i|}\lambda_k = 1$ and

$$z = \sum_{k=1}^{|F_i|}\lambda_k x_k = \sum_{k=1}^{|F_j|}\lambda_k x_k + \sum_{k=|F_j|+1}^{|F_i|}\lambda_k P_j x_k + \sum_{k=|F_j|+1}^{|F_i|}\lambda_k(x_j - P_j x_k) = z_1 + z_2$$

where $z_1 \in \mathrm{conv}(F_j)$ and $z_2 \in \mathrm{conv}(G_{ij}\cup\{0\})$. Hence,

$$\mathrm{conv}(F_i) \subset \mathrm{conv}(F_j) + \mathrm{conv}(G_{ij}\cup\{0\})$$

and the first assertion is verified.

It is routine to see that if $A, B, C \subset L_2(\mu)$ are such that $A \subset B + C$, then for every $\varepsilon_1, \varepsilon_2 > 0$,

$$N(\varepsilon_1+\varepsilon_2, A) \leq N(\varepsilon_1, B)\cdot N(\varepsilon_2, C).$$

Thus, by the first claim it follows that

$$N(\varepsilon, \mathrm{conv}(F_i)) \leq N\big(\frac{\varepsilon}{2}, \mathrm{conv}(F_j)\big)\cdot N\big(\frac{\varepsilon}{2}, \mathrm{conv}(G_{ij}\cup\{0\})\big). \tag{2.1}$$

To estimate the first term, note that $\mathrm{span}(F_j)$ can be isometrically embedded in $\ell_2^{|F_j|}$. Therefore, the covering numbers of F_j in $L_2(\mu)$ and in $\ell_2^{|F_j|}$ are the same. Let $\mathcal{B}$ be the unit ball in $\ell_2^{|F_j|}$. Using a standard volume estimate (see Pisier, 1989), one can show that for every $\varepsilon, \varepsilon'$,

$$N(\frac{\varepsilon}{2}, \varepsilon'\mathcal{B}) \leq \Big(1+\frac{4\varepsilon'}{\varepsilon}\Big)^{|F_j|}.$$

Hence,

$$\begin{aligned} N\big(\frac{\varepsilon}{2}, \mathrm{conv}(F_j)\big) &\leq N\big(\frac{\varepsilon}{2}, \varepsilon'\mathcal{B}\big)\cdot N\big(\varepsilon', \mathrm{conv}(F_j)\big) \\ &\leq \Big(1+\frac{4\varepsilon'}{\varepsilon}\Big)^{|F_j|}\cdot N\big(\varepsilon', \mathrm{conv}(F_j)\big). \end{aligned}$$

As for the second term in (2.1), since $\mathrm{diam}(G_{ij}\cup\{0\}) \leq 2\delta_j$ and $|G_{ij}\cup\{0\}| \leq |F_i|$, then by Lemma 2.2

$$\begin{aligned} N\Big(\frac{\varepsilon}{2}, \mathrm{conv}\big(G_{ij}\cup\{0\}\big)\Big) &= N\Big(\frac{\varepsilon}{4\delta_j}2\delta_j, \mathrm{conv}\big(G_{ij}\cup\{0\}\big)\Big) \\ &\leq \Big(e + e\,|F_i|\,\frac{\varepsilon^2}{16\delta_j^2}\Big)^{32\delta_j^2/\varepsilon^2}. \end{aligned}$$

■

Lemma 2.3 reveals our strategy which is similar in nature to chaining. Here, we create an increasing family of sets which form an increasingly finer approximation of F. It is possible to estimate the covering numbers of the convex hulls of the "finer" sets using an estimate on the "coarser" sets. The rates by which the mesh of the classes F_i decreases will be dictated by the growth rates of the covering numbers in the class F. In the next two technical results we select an appropriate rate of decay for the "mesh" sequence (δ_n).

Lemma 2.4 *Let F be a subset of the unit ball in $L_2(\mu)$ and assume that there are constants $\gamma, p > 0$, such that for every $\varepsilon > 0$,*

$$N\big(\varepsilon, F, L_2(\mu)\big) \leq \gamma\Big(\frac{1}{\varepsilon}\Big)^p.$$

Let $\delta_n = \gamma^{1/p} n^{-1/p}$ and set F_n to be as in Lemma 2.3. Then, there are bounded sequences (A_k) and (B_k) such that for every n and k,

$$\log N\big(A_k n^{\frac{1}{2}} \delta_n, \operatorname{conv}(F_{nk^{3p}})\big) \leq B_k n. \tag{2.2}$$

Moreover, there are absolute constants A and B such that for every k, $A_k \leq \gamma^{1/p} A$ and $B_k \leq Bp$.

Lemma 2.5 *Let F be a subset of the unit ball of $L_2(\mu)$ and assume that there are $\gamma > 0$ and $0 < p < 2$ such that for every $\varepsilon > 0$,*

$$\log N\big(\varepsilon, F, L_2(\mu)\big) \leq \gamma \varepsilon^{-p}.$$

Set $\delta_n = 2\gamma^{1/p} \log^{-1/p} n$, $\varepsilon_n = n^{-1} \log^{1/p} n$ and let F_n be as in Lemma 2.3. Then, there are sequences (A_k) and (B_k) which depend on p, such that for all integers $n \geq 2$ and $k \geq 1$

$$\log N\big(A_k \varepsilon_n, \operatorname{conv}(F_{[n^{k^\alpha}]})\big) \leq B_k \frac{n^2}{\log^{\frac{4}{p}-1} n},$$

where $\alpha = 4p/(2-p)$ and $[x]$ denotes the integer value of x. Moreover, $\sup A_k \leq A'_p$ and $\sup B_k \leq B'_p$ for some constants A'_p and B'_p which depend only on p.

We present a complete proof of Lemma 2.5. The proof of Lemma 2.4 follows from a similar argument. The idea behind the proof of Lemma 2.4 is due to van der Vaart and Wellner (1996). The quantitative estimate on the constants does not appear in that text, but may be derived by a close analysis of the proof the authors present.

Proof of Lemma 2.5: We use a nested induction argument. First, we prove our claim for $k = 1$ using induction on n. We then prove the claim for a general k for every fixed n.

Recall that for every integer n, F_n is maximal δ_n separated in F. Thus

$$|F_n| \leq N(\frac{\delta_n}{2}, F) \leq e^{\gamma(2/\delta_n)^p} = n.$$

Let $n_0 \geq 4$ be an integer such that for every $n \geq n_0$,

$$\frac{\lceil \frac{n}{2} \rceil^2}{\log^{\frac{4}{p}-1} \lceil \frac{n}{2} \rceil} \leq \frac{3}{4} \frac{n^2}{\log^{\frac{4}{p}-1} n}.$$

Let $k = 1$ and $2 \leq n \leq 2n_0$ and set $A_1 = 2n_0 \log^{-1/p} 2n_0$. Since for such a value of n, $A_1\varepsilon_n \geq 1$, only a single ball is required to cover F and our claim follows.

Next, let $n > 2n_0$ and assume that for every $2 \leq m < n$,

$$\log N(A_1\varepsilon_m, F_m) \leq B_1 m^2 \log^{1-4/p} m,$$

where B_1 is to be specified later. We can apply Lemma 2.2 with $i = n$, $j = \lceil n/2 \rceil$, $\varepsilon = A_1\varepsilon_n$ and $\varepsilon' = A_1\varepsilon_{\lceil n/2 \rceil}$. It follows that

$$\begin{aligned} &N\big(A_1\varepsilon, \operatorname{conv} F_n\big) \\ &\leq \Big(1 + \frac{4\varepsilon_{\lceil \frac{n}{2} \rceil}}{\varepsilon_n}\Big)^{\lceil \frac{n}{2} \rceil} \cdot N\big(A_1\varepsilon_{\lceil \frac{n}{2} \rceil}, \operatorname{conv} F_{\lceil \frac{n}{2} \rceil}\big) \cdot \Big(e + e\,|F_n|\,\frac{\varepsilon^2}{16\delta_j^2}\Big)^{32\delta_j^2/\varepsilon^2}. \end{aligned} \tag{2.3}$$

A straightforward calculation shows that there is an absolute constant C such that for every integer n,

$$\frac{\varepsilon_{\lceil n/2 \rceil}}{\varepsilon_n} \leq C, \qquad \frac{\delta_j}{\varepsilon_n} \leq C\gamma^{\frac{1}{p}} \frac{n}{\log^{\frac{2}{p}} n}.$$

Applying the induction hypothesis and (2.3) there is a constant $C = C(p)$ such that

$$\log N\big(A_1\varepsilon_n, \operatorname{conv} F_n\big) \leq Cn + B_1 \frac{\lceil \frac{n}{2} \rceil^2}{\log^{\frac{4}{p}-1} \lceil \frac{n}{2} \rceil} + C\frac{\gamma^{\frac{2}{p}} n^2}{\log^{\frac{4}{p}} n}.$$

By the selection of n_0,

$$B_1 \lceil \frac{n}{2} \rceil^2 \log^{1-\frac{4}{p}} \lceil \frac{n}{2} \rceil \leq \frac{3}{4} B_1 n^2 \log^{1-\frac{4}{p}} n,$$

and thus, there is a constant $C(p, \gamma)$ such that if $B_1 = C(p, \gamma)$ then

$$N(A_1\varepsilon_n, \operatorname{conv} F_n) \leq B_1 n^2 \log^{1-\frac{4}{p}} n$$

as claimed.

Now, we fix some $n \geq 2$ and use induction with respect to k. Let $i = [n^{k^\alpha}]$ and $j = [n^{(k-1)^\alpha}]$. Note that $|G_{ij} \cup \{0\}| \leq [n^{k^\alpha}]$ and that

$$\operatorname{diam}(G_{ij} \cup \{0\}) \leq 2\delta_j = 4\gamma^{1/p} \log^{-1/p} [n^{(k-1)^\alpha}].$$

By Lemma 2.2,

$$\begin{aligned} N\Big(\frac{\varepsilon_n}{k^2}, \operatorname{conv}(G_{ij} \cup \{0\})\Big) &\leq N\Big(\frac{\varepsilon_n}{2\delta_j k^2} \operatorname{diam}(G_{ij} \cup \{0\}), \operatorname{conv}(G_{ij} \cup \{0\})\Big) \leq \\ &\leq \Big(e + e\,|G_{ij} \cup \{0\}|\,\frac{\varepsilon_n^2}{4k^2\delta_j^2}\Big)^{8k^2\delta_j^2/\varepsilon_n^2}. \end{aligned}$$

It is straightforward to see that there is some constant $C = C(p)$ such that

$$\frac{k^2\delta_j^2}{\varepsilon_n^2} = \frac{\gamma^{\frac{2}{p}} k^2}{\log^{\frac{2}{p}} [n^{(k-1)^\alpha}]} \cdot \frac{n^2}{\log^{\frac{2}{p}} n} \leq C \frac{\gamma^{\frac{2}{p}} k^2}{(k-1)^{\frac{2\alpha}{p}}} \cdot \frac{n^2}{\log^{\frac{4}{p}} n}.$$

Also,

$$\frac{|G_{ij}\cup\{0\}|\,\varepsilon_n^2}{k^2\delta_j^2}\le C\frac{n^{k^\alpha}}{k^{2-\frac{2\alpha}{p}}}\cdot\frac{\log^{\frac{4}{p}}n}{n^2}\le C\frac{n^{k^\alpha}}{k^{2-\frac{2\alpha}{p}}}.$$

Using the above estimates and the definition of α it follows that

$$\begin{aligned}&\log N\Big(\frac{\varepsilon_n}{k^2},\mathrm{conv}(G_{ij}\cup\{0\})\Big)\\ \le& C(p,\gamma)\frac{n^2}{\log^{\frac{4}{p}-1}n}\cdot\frac{k^{2+\alpha}\log k}{(k-1)^{\frac{2\alpha}{p}}}=C(p,\gamma)\frac{n^2\log k}{k^2\log^{\frac{4}{p}-1}n}.\end{aligned}$$

On the other hand, by the induction hypothesis,

$$\log\Big(A_{k-1}\varepsilon_n,\mathrm{conv}\,F_{[n^{(k-1)^\alpha}]}\Big)\le B_{k-1}\frac{n^2}{\log^{\frac{4}{p}-1}n}.$$

Applying Lemma 2.3 and combining the two covers, we obtain an $A_k\varepsilon_n=(A_{k-1}+1/k^2)\varepsilon_n$ cover of $\mathrm{conv}\,F_{[n^{k^\alpha}]}$. Hence,

$$\log\big(A_k\varepsilon_n,\mathrm{conv}\,F_{[n^{k^\alpha}]}\big)\le\Big(B_{k-1}+C(p,\gamma)\frac{\log k}{k^2}\Big)\frac{n^2}{\log^{\frac{4}{p}-1}n}.$$

And our claim follows. It is important to note that the sequences (A_k) and (B_k) are bounded by some constant $C=C(p,\gamma)$.

■

Proof of Theorem 2.1: We begin with the proof of the first part of our theorem. Fix some integer $n\ge 2$ and let $\varepsilon_n=\gamma^{1/p}n^{-1/2-1/p}$. By Lemma 2.4 it follows that for every integer k,

$$\log N\big(A_k\varepsilon_n,\mathrm{conv}(F_{nk^{3p}})\big)\le B_kn\le Bpn=Bp\Big(\frac{1}{\varepsilon_n}\Big)^{\frac{2p}{2+p}}.$$

Since for every k, $A\ge A_k$, then

$$\log N\big(A\varepsilon_n,\mathrm{conv}(F_{nk^{3p}})\big)\le\log N\big(A_k\varepsilon_n,\mathrm{conv}(F_{nk^{3p}})\big)\le Bp\Big(\frac{1}{\varepsilon_n}\Big)^{\frac{2p}{2+p}}.$$

Taking k to infinity,

$$\log N\big(A\varepsilon_n,\mathrm{conv}(F)\big)\le Bp\Big(\frac{1}{\varepsilon_n}\Big)^{\frac{2p}{2+p}}.$$

The claim for a general ε follows since $\varepsilon_n/\varepsilon_{n+1}$ is a bounded sequence.

Turning to the second assertion, according to Lemma 2.5 and by the same argument as above for every $n\ge 2$,

$$\log N\big(A_p'\varepsilon_n,\mathrm{conv}(F)\big)\le B_p'\frac{n^2}{\log^{\frac{4}{p}-1}n}.$$

Since $\varepsilon_n=n^{-1}\log^{1/p}n$ then $n=\varepsilon_n^{-1}\log^{\frac{1}{p}}n$ and $n\ge\frac{C}{\varepsilon_n}\log^{\frac{1}{p}}\frac{1}{\varepsilon_n}$. Hence,

$$\log N\big(A_p'\varepsilon_n,\mathrm{conv}(F)\big)\le B_p'\frac{1}{\varepsilon_n^2}\log^{1-\frac{2}{p}}\frac{1}{\varepsilon_n}.$$

Again, the claim follows since $\varepsilon_n/\varepsilon_{n+1}$ is a bounded sequence.

■

2.2 Applications

We shall present several examples which are interesting from the Machine Learning perspective.

One of the basic results regarding VC classes is that the covering numbers of such classes in any $L_2(\mu)$ are polynomial in $1/\varepsilon$ (Haussler, 1995, van der Vaart and Wellner, 1996):

Theorem 2.6 *Let F be a class of $\{0,1\}$-valued functions such that $VC(F) = d$. Then, there is an absolute constant C such that for every probability measure μ on Ω, $N(\varepsilon, F, L_2(\mu)) \leq Cd(4e)^d \varepsilon^{-2d}$.*

Corollary 2.7 *Assume that F is a $\{0,1\}$-class such that $VC(F) = d$. By Theorem 2.1 and Theorem 2.6 for $p = 2d$ and $\gamma = Cd(4e)^d$, and since $\gamma^{2/p}$ is bounded by some absolute constant, it follows that*

$$\log N\big(\varepsilon, \mathrm{conv}(F), L_2(\mu)\big) \leq Cd\Big(\frac{1}{\varepsilon}\Big)^{\frac{2d}{1+d}},$$

where C is an absolute constant.

Next, we establish a similar estimate in the case where the fat-shattering dimension satisfies $\mathrm{fat}_\varepsilon(F) = O(\varepsilon^{-p})$ for some $0 < p < 2$. It is possible to connect the fat-shattering dimension and the $L_2(\mu)$ covering numbers (Mendelson, 2001c).

Theorem 2.8 *Let F be a class of functions into $[0,1]$. Then, there is an absolute constant C such that for every probability measure μ,*

$$\log N\big(\varepsilon, F, L_2(\mu)\big) \leq C\mathrm{fat}_{\frac{\varepsilon}{32}}(F) \log^2\Big(\frac{2\mathrm{fat}_{\frac{\varepsilon}{32}}(F)}{\varepsilon}\Big).$$

Corollary 2.9 *Let F be a class of functions from Ω into $[0,1]$ and assume that there are $\gamma > 0$ and $0 < p < 2$ such that $\mathrm{fat}_\varepsilon(F) \leq \gamma\varepsilon^{-p}$. Then, for every $p' > p$ there is some constant $C = C(p, p', \gamma)$ such that for every probability measure μ,*

$$\log N\big(\varepsilon, \mathrm{conv}(F), L_2(\mu)\big) \leq C\frac{1}{\varepsilon^2}\log^{1-\frac{2}{p'}}\frac{1}{\varepsilon}.$$

3. Type and the fat-shattering dimension

In this section, we tackle the problem of estimating the fat-shattering dimension of convex hulls of classes. To that end, we first deal with "linear" fat-shattering dimension. By this we mean the fat-shattering dimension of a set of linear functionals, for example, the unit ball of the dual space of some Banach space, when considered to be a class of functions on the unit ball of the space. As it turns out, given a Banach space X, the fat-shattering dimension of the set $B(X^*)$ when considered as functions on $B(X)$ is completely determined by a property of X called *type* (defined below).

The results we present aid our goal since one can apply an embedding argument to show that, in some sense, the fat-shattering dimension of a given class may be controlled by the fat-shattering dimension of a class of linear functionals.

In order to bound the fat-shattering dimension of convex hulls we take the following course of action: first, we apply the embedding argument, which reduces the problem to a "linear" one. Then, we show that the symmetric convex hull of the embedded class can be approximated by a class which has "almost" the same fat-shattering dimension, but also a well behaved type structure. This is done by using an estimate on the covering numbers of the class with respect to the L_∞ norm. The fat-shattering of the approximating class may be bounded via the bound on $\mathrm{fat}_\varepsilon\big(B(X^*), B(X)\big)$ mentioned above.

Before we continue, we require additional definitions, originating in the theory of Banach spaces. For the basic definitions we refer the reader to (Pisier, 1989) or (Tomczak-Jaegermann, 1989).

Let K be a bounded, convex symmetric subset of $\mathbb{R}^n$ which has a nonempty interior. One can define a norm on $\mathbb{R}^n$ whose unit ball is K. This is done using the Minkowski functional on K, which is denoted by $\| \ \|_K$, and given by

$$\|x\|_K = \inf\{t > 0 | t^{-1}x \in K\}.$$

It is possible to show that if $K \subset \ell_2^n$ is convex and symmetric with a nonempty interior then $\| \ \|_K$ is indeed a norm and K is its unit ball. Set $\| \ \|_{K^*}$ to be the dual norm to $\| \ \|_K$.

Definition 3.1 *If F is a bounded subset of ℓ_2^n, let*

$$F^{\mathrm{o}} = \{x \in \ell_2^n | \sup_{f \in F} \big|\langle f, x\rangle\big| \leq 1\},$$

where $\langle -, - \rangle$ is the inner product in ℓ_2^n. The set F^{o} is called the polar of F.

For any set F, let $\mathrm{absconv}(F)$ be its symmetric convex hull. Formally,

$$\mathrm{absconv}(F) = \Big\{ \sum_{i=1}^n a_i f_i | n \in \mathbb{N}, f_i \in F, \sum_{i=1}^n |a_i| = 1 \Big\}.$$

It is easy to see that $F^o = \big(\mathrm{absconv}(F)\big)^o$ and that if $G \subset F$ then $F^o \subset G^o$. Note that F^o is the unit ball of the norm $\| \ \|_{K^*}$, where $K = \mathrm{absconv}(F)$. Hence, for every $x \in \ell_2^n$,

$$\|x\|_{F^o} = \sup_{y \in \mathrm{absconv(F)}} \langle y, x\rangle = \sup_{y \in F} \big|\langle y, x\rangle\big|.$$

In particular, if $F = \{f_1, ..., f_m\}$ then $\|x\|_{F^o} = \max_{1 \leq i \leq m} \big|\langle f_i, x\rangle\big|$.

Given a class F and an empirical measure μ_n supported on $\{\omega_1, ..., \omega_n\}$, we endow $\mathbb{R}^n$ with the Euclidean structure of $L_2(\mu_n)$, which is isometric to ℓ_2^n. Therefore, for every $f \in L_2(\mu_n)$,

$$\|f\|_{L_2(\mu_n)} = \Big(\frac{1}{n} \sum_{i=1}^n f^2(\omega_i)\Big)^{\frac{1}{2}}.$$

Let F/μ_n be the image of F in $L_2(\mu_n)$ under the inclusion operator, that is,

$$F/\mu_n = \Big\{ \sum_{i=1}^n f(\omega_i)\chi_{\{\omega_i\}} | f \in F \Big\}.$$

Since $(n^{1/2}\chi_{\{\omega_i\}})_{i=1}^n$ is an orthonormal basis of $L_2(\mu_n)$, then

$$F/\mu_n = \Big\{ n^{-\frac{1}{2}} \sum_{i=1}^n f(\omega_i) e_i | f \in F \Big\}.$$

Throughout this section, given an empirical measure μ_n, we denote by $(e_i)_{i=1}^n$ the orthonormal basis of $L_2(\mu_n)$ given by $(n^{1/2}\chi_{\{\omega_i\}})_{i=1}^n$.

The next lemma is straightforward and its proof is omitted.

Lemma 3.2 *Let $S = \{\omega_1, ..., \omega_n\}$ be a sample and let μ_n be the empirical measure supported on S.*

1. *If S is shattered by F then the set $\{\sqrt{n}e_1, ..., \sqrt{n}e_n\} \subset (F/\mu_n)^o$ is shattered by F/μ_n.*
2. *If S is ε-shattered by F then $\{\sqrt{n}e_1, ..., \sqrt{n}e_n\} \subset (F/\mu_n)^o$ is ε-shattered by F/μ_n.*

3.1 The fat-shattering dimension of linear functionals

Recall the definition of the Rademacher type of a Banach space:

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

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

where (ε_i) are independent Rademacher random variables (i.e., symmetric $\{-1, 1\}$-valued). The best constant for which (3.1) holds is called the Rademacher p-type constant of X and is denoted by $T_p(X)$.

Theorem 3.4 *Let X be a Banach space which has type p for some $1 < p \leq 2$ with a type constant $T_p(X)$. Then*

$$\mathrm{fat}_\varepsilon \big(B(X^*), B(X) \big) \leq \Big(\frac{T_p(X)}{\varepsilon} \Big)^{\frac{p}{p-1}}.$$

Note that a similar result to this was demonstrated by Gurvits (2001), though in his result the bound is on the level fat-shattering dimension (that is, the witness $(s_i)_{i=1}^n$ to the shattering is a constant set: there is some a such that $s_i = a$ for every $1 \leq i \leq n$).

Proof: Assume that the set $\{x_1, ..., x_n\} \subset B(X)$ is ε-shattered by $B(X^*)$ and set $\{s_1, ..., s_n\}$ to be a witness to the shattering. Let $I \subset \{1, ..., n\}$ and put x_I^* to be the functional shattering the set I. Note that if $i \in I$ then

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

and if $i \in I^c$,

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

Thus,

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

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

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

Now, we can use the type property to establish an upper bound for an appropriate subset I which will be selected randomly. Indeed, since X has type p and since $\|x_i\|\leq 1$ then

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

Thus, there is a realization of the random variables (ε_i) for which this inequality holds. Let $I=\{i|\varepsilon_i=1\}$. Then,

$$\left\|\sum_{i\in I}x_i-\sum_{i\in I^c}x_i\right\| \leq T_p(X)n^{\frac{1}{p}}.$$

Thus, $n\varepsilon \leq T_p(X)n^{1/p}$ and our claim follows.

■

Remark 3.5 *It is possible to show that this bound is tight (see Mendelson, 2001a). Indeed, one can show that if* $p^* = \sup\{p|\text{X has type } p\}$ *then for every* $\varepsilon>0$,

$$\Big(\frac{1}{\varepsilon}\Big)^{\frac{p^*}{p^*-1}}-1 \leq \mathrm{fat}_\varepsilon\big(B(X^*),B(X)\big).$$

This implies that if X *is a Hilbert space then for every* $\varepsilon>0$,

$$\Big(\frac{1}{\varepsilon^2}\Big)-1 \leq \mathrm{fat}_\varepsilon\big(B(X^*),B(X)\big) \leq \Big(\frac{1}{\varepsilon^2}\Big).$$

3.2 Covering numbers and type constants

Here, we show that if the class F is well behaved, (either a VC class or a class with a finite fat-shattering dimension for every ε), then the Rademacher type-2 constant of $(F/\mu_n)^o$ does not grow too rapidly as a function of n.

In order to bound the type-2 constant of $(F/\mu_n)^o$, we use a combination of two facts. The first, which is the only non elementary fact we require, is an estimate on $T_2(\ell_\infty^n)$. It is possible to show (Tomczak-Jaegermann, 1989) that there are absolute constants c and C such that for every integer n,

$$c(1+\log n)^{1/2} \leq T_2(\ell_\infty^n) \leq C(1+\log n)^{1/2}. \tag{3.2}$$

The second fact we use is that if the cardinality of F is small then F^o can be isometrically embedded into ℓ_∞^n for a relatively small n.

Lemma 3.6 *Let $F \subset \ell_2^n$ be a finite set. Then F^o can be isometrically embedded into $\ell_\infty^{|F|}$.*

Proof: Let $F = \{f_1, ..., f_m\}$ and define $T : (\mathbb{R}^n, \| \ \|_{F^o}) \to \ell_\infty^{|F|}$ by $Tx^* = \big(x^*(f_i)\big)_{i=1}^m$. Then, for every $x^* \in \mathbb{R}^n$

$$\|Tx^*\|_{\ell_\infty^{|F|}} = \sup_{1\leq i\leq m} |x^*(f_i)| = \|x^*\|_{F^o},$$

implying that T is an isometry.

■

This fact is very useful from our point of view, since F/μ_n are relatively small sets. In the real valued case, the $L_\infty(\mu_n)$ covering numbers of F may be bounded using $\mathrm{fat}_\varepsilon(F)$ (Theorem 1.3), whereas for VC classes, one may use the following version of Sauer's Lemma (van der Vaart and Wellner, 1996):

Lemma 3.7 *There is an absolute constant C such that if F is a class of $\{0,1\}$-valued functions on Ω with $VC(F) = d$, then for every empirical measure μ_n, $|F/\mu_n| \leq Cn^d$.*

Now, we can bound $T_2\big((F/\mu_n)^o\big)$ for VC classes. In fact, we prove the following:

Theorem 3.8 *Let F be a class of $\{0,1\}$-valued functions. Then, F is a VC class if and only if there is some constant $C > 0$ such that for every integer n and every empirical measure μ_n, $T_2\big((F/\mu_n)^o\big) \leq C(1+\log n)^{1/2}$.*

Proof: Note that if $VC(F) = d$ then by Sauer's Lemma there is an absolute constant C such that $|F/\mu_n| \leq Cn^d$. Hence, $(F/\mu_n)^o$ can be isometrically embedded in $\ell_\infty^{Cn^d}$. It follows that $T_2(F^o) \leq C(1+d\log n)^{1/2} \leq Cd^{1/2}(1+\log n)^{1/2}$ as claimed.

Conversely, assume that there is a constant C such that for every integer n and every empirical measure μ_n, $T_2((F/\mu_n)^o) \leq C(1+\log n)^{1/2}$. Let $S = \{\omega_1, ..., \omega_n\}$ be shattered by F and set μ_n to be the empirical measure supported on S. By the first part of Lemma 3.2 and Theorem 3.4,

$$n \leq \mathrm{fat}_{\frac{1}{2}}((F/\mu_n)^o, F/\mu_n) \leq 4T_2^2\big((F/\mu_n)^o\big) \leq C(1+\log n),$$

implying that n can not be arbitrarily large, and thus $VC(F,\Omega) < \infty$.

■

Remark 3.9 *The proof of theorem 3.8 implies that if F is a $\{0,1\}$-valued class such that $\sup_{\mu_n} T_2\big((F/\mu_n)^o\big) = o(n)$ then F is a VC class.*

Note that the upper bound on $T_2\big((F/\mu_n)^o\big)$ can not be improved. Indeed, let F be the set of characteristic functions of intervals $[-1,a]$ for $a \in (-1,1]$. Thus, $VC(F) = 1$ and for every non degenerate empirical measure μ_n of $[-1,1]$,

$$F/\mu_n = \Big\{(1,0,...,0),(1,1,0,...,0),...,(1,1,...,1)\Big\}.$$

Therefore, $(F/\mu_n)^o$ is isometric to ℓ_∞^n and $T_2\big((F/\mu_n)^o\big) \geq C(1+\log n)^{1/2}$.

Also, Theorem 3.8 does not apply to classes which are not $\{0,1\}$-valued classes. For example, let $\Omega = B(\ell_p)$ for some $1 < p < 2$ and $F = B(\ell_q)$ where $p^{-1}+q^{-1} = 1$. For every i denote by e_i the i-th unit vector in ℓ_p, and given an integer n, let μ_n be the empirical measure supported on $\{e_1,...,e_n\} \subset B(\ell_p)$. Since $F/\mu_n = n^{-1/2}B(\ell_q^n)$ then $(F/\mu_n)^o$ is isometric to ℓ_p^n and $T_2\big((F/\mu_n)^o\big) \geq n^{\frac{1}{p}-\frac{1}{2}}$ (see Tomczak-Jaegermann, 1989). Hence, it is impossible to obtain an analogous result to Theorem 3.8 in the general case. We bypass this obstacle by replacing the set F/μ_n (which may be infinite) by an $L_\infty(\mu_n)$ cover of F.

The following lemma indicates that an ε-cover of a set in $L_\infty(\mu_n)$ has essentially the same fat-shattering dimension as the original set.

Lemma 3.10 *Let F be a class of functions into $[0,1]$ which δ-shatters $A = \{\omega_1,...,\omega_n\}$, and let μ_n be the empirical measure on A. Assume that $\delta > \varepsilon$ and that H is an ε-cover of F/μ_n in $L_\infty(\mu_n)$. Then, $A' = \{\sqrt{n}e_1,...,\sqrt{n}e_n\} \subset (F/\mu_n)^o$ is $(\delta-\varepsilon)$-shattered by H, where the elements of H are viewed as linear functionals on $(F/\mu_n)^o$.*

Proof: By Lemma 3.2, A' is δ-shattered by F/μ_n and set (s_i) to be a witness to the shattering. Fix $I \subset \{1,...,n\}$ and let $f_I \in F/\mu_n$ be the δ-shattering functional on I. Put $h_I \in H$ such that $\|f_I - h_I\|_{L_\infty(\mu_n)} < \varepsilon$. Since $\|\sqrt{n}e_i\|_{L_1(\mu_n)} = 1$ then for every $i \in I$

$$\begin{aligned}\big\langle h_I, \sqrt{n}e_i\big\rangle &= \big\langle f_I, \sqrt{n}e_i\big\rangle + \big\langle h_I - f_I, \sqrt{n}e_i\big\rangle \\ &\geq s_i + \delta - \|h_I - f_I\|_{L_\infty(\mu_n)} \big\|\sqrt{n}e_i\big\|_{L_1(\mu_n)} \\ &\geq s_i + \delta - \varepsilon,\end{aligned}$$

and in a similar fashion, if $j \in I^c$ then $\big\langle h_I, \sqrt{n}e_i\big\rangle \leq s_i - \delta + \varepsilon$, and our claim follows. ■

3.3 The fat-shattering dimension of convex hulls

We begin by formulating the main result, which enables one to estimate the fat-shattering dimension of the convex hull of a class F using the $L_\infty(\mu_n)$ covering numbers of F.

We introduce the following notation. Given a class F, an integer n and $\varepsilon > 0$, let

$$N_\infty(n,\varepsilon) = \sup_{\mu_n} N\big(\varepsilon, F, L_\infty(\mu_n)\big).$$

Thus, $N_\infty(n,\varepsilon)$ is the supremum of the ε-covering numbers of F in empirical L_∞ spaces, where the empirical measure is supported on at most n elements of Ω.

Theorem 3.11 *Let F be a class of functions whose range is a subset of $[0,1]$ and set K to be its symmetric convex hull. Then, for every $\varepsilon, \tau > 0$,*

$$\mathrm{fat}_{\varepsilon+\tau}\big(K,\Omega\big) \leq \max\Big\{n \big| n \leq C\big(1+\log N_\infty(n,2\varepsilon)\big)\frac{1}{\tau^2}\Big\},$$

where C is an absolute constant.

Proof: Fix $\varepsilon, \tau > 0$, assume that $A = \{\omega_1, ..., \omega_n\}$ is $(\varepsilon + \tau)$-shattered by $K = \mathrm{absconv}(F)$ and let μ_n be an empirical measure supported on A. Let $H \subset F/\mu_n$ be an ε-cover to F/μ_n in $L_\infty(\mu_n)$ and set G to be the symmetric convex hull of H. Clearly, $|H| \leq N\big(2\varepsilon, F/\mu_n, L_\infty(\mu_n)\big)$.

Since H is an ε-cover of F/μ_n in $L_\infty(\mu_n)$ then G is an ε-cover to K in $L_\infty(\mu_n)$. By Lemma 3.10, if the set A is $\varepsilon + \tau$ shattered by K then the set $A' = \{\sqrt{n}e_1, ..., \sqrt{n}e_n\}$ is τ-shattered by G, when members of G are viewed as linear functionals on $(F/\mu_n)^o$. Also, note that $A' \subset (F/\mu_n)^o$. Thus,

$$n \leq \mathrm{fat}_\tau\big(G, (\sqrt{n}e_i)_{i=1}^n\big) \leq \mathrm{fat}_\tau\big(G, (F/\mu_n)^o\big).$$

Since $H \subset F/\mu_n$, then by taking convex hulls $G \subset K$. Using the properties of the polar, $(F/\mu_n)^o = K^o \subset G^o$, thus $\mathrm{fat}_\tau\big(G, (F/\mu_n)^o\big) \leq \mathrm{fat}_\tau(G, G^o)$. By Lemma 3.6 there is an absolute constant C such that

$$T_2(G^o) \leq C(1 + \log|H|)^{\frac{1}{2}} \leq C\Big(1 + \log N\big(2\varepsilon, F/\mu_n, L_\infty(\mu_n)\big)\Big)^{\frac{1}{2}}.$$

Hence, by Theorem 3.4,

$$\mathrm{fat}_\tau(G, G^o) \leq C\Big(1 + \log N\big(\varepsilon, F/\mu_n, L_\infty(\mu_n)\big)\Big)\frac{1}{\tau^2}.$$

Therefore,

$$n \leq C\Big(1 + \log N\big(2\varepsilon, F/\mu_n, L_\infty(\mu_n)\big)\Big)\frac{1}{\tau^2}.$$

■

Applying the known bounds on the $L_\infty(\mu_n)$ covering numbers in terms of the VC or the fat-shattering dimension we can establish the following estimate on the fat-shattering dimension of convex hulls.

Corollary 3.12 *There is an absolute constant C such that for every $\{0,1\}$-class of functions and every $\varepsilon, \tau > 0$,*

$$\mathrm{fat}_{\varepsilon+\tau}(K, \Omega) \leq C\frac{d}{\tau^2}\log\frac{2d}{\tau},$$

where K is the symmetric convex hull of F and $VC(F) = d$.
In particular, there is an absolute constant C such that for every $\tau > 0$,

$$\mathrm{fat}_\tau(K, \Omega) \leq C\frac{d}{\tau^2}\log\frac{2d}{\tau}.$$

Proof: By Sauer's Lemma, there is an absolute constant C such that for every integer n, $\log N_\infty(n, \varepsilon) \leq Cd\log n$. Fix $\varepsilon, \tau > 0$ and let $\mathrm{fat}_{\varepsilon+\tau}(K, \Omega) = n$. By Theorem 3.11, $n \leq Cd\tau^{-2}(1 + \log n)$ and our claim follows.

■

The best known estimate on the fat-shattering dimension of the convex hull of a VC class is $\mathrm{fat}_\varepsilon\big(\mathrm{conv}(F)\big) = O(d/\varepsilon^2)$. This bound is due to Gurvits (2001) and was obtained using a

similar geometric approach to the one presented here. The one key difference between the two approaches is that Gurvits uses a different notion of type. His result is based on an estimate on the type-2 constant of a certain operator, which is due to Ledoux and Talagrand (1991, Theorem 14.15). It turns out that their bound depends on the fact that the class F has a converging entropy integral, which is the case for VC classes.

This approach is not suitable for arbitrary classes of functions, when one does not have a-priori "global" data (for example, the fat-shattering dimension or covering numbers at every scale), but rather, the fat-shattering dimension at a given scale. It is possible to extend Corollary 3.12 and estimate the fat-shattering dimension of the convex hull of a class using the fat-shattering dimension of the class itself, without resorting to "global" data.

Theorem 3.13 *There is an absolute constant C such that for any class F of functions into $[0,1]$ and every $\varepsilon, \tau \in (0,1)$,*

$$\mathrm{fat}_{\varepsilon+\tau}(K,\Omega) \leq C \frac{\mathrm{fat}_{\frac{\varepsilon}{2}}(F,\Omega)}{\tau^2} \log^2 \frac{2\mathrm{fat}_{\frac{\varepsilon}{2}}(F,\Omega)}{\tau^2 \varepsilon}$$

where K is the symmetric convex hull of F.
In particular, there is an absolute constant C such that for every $\tau > 0$,

$$\mathrm{fat}_{\tau}(K,\Omega) \leq C \frac{\mathrm{fat}_{\frac{\tau}{4}}(F,\Omega)}{\tau^2} \log^2 \frac{2\mathrm{fat}_{\frac{\tau}{4}}(F,\Omega)}{\tau}.$$

Proof: Set $\varepsilon, \delta \in (0,1)$ and let $n = \mathrm{fat}_{2\varepsilon+\tau}(K,\Omega)$. By Theorem 1.3,

$$\log N_\infty(n, 2\varepsilon) \leq C \mathrm{fat}_{\frac{\varepsilon}{2}}(F,\Omega) \log^2 \frac{n}{\varepsilon}.$$

Applying Theorem 3.11 and since $n > 1$ and $\varepsilon \in (0,1)$, there is an absolute constant C such that

$$n \leq C \frac{\mathrm{fat}_{\frac{\varepsilon}{2}}(F,\Omega)}{\tau^2} \log^2 \frac{n}{\varepsilon}.$$

Hence,

$$n \leq C \frac{\mathrm{fat}_{\frac{\varepsilon}{2}}(F,\Omega)}{\tau^2} \log^2 \left(\frac{2\mathrm{fat}_{\frac{\varepsilon}{2}}(F,\Omega)}{\tau^2 \varepsilon} \right),$$

and the claim follows.

■

Remark 3.14 *Using a different approach it is possible to improve Theorem 3.13 when one has "global" data on the fat-shattering dimension of the class. For example, it is possible to recover Gurvits' estimate for the convex hull of VC classes, with a much simpler proof. This approach may be extended to a more general setup. Indeed, if $\mathrm{fat}_\varepsilon(F) = O(\varepsilon^{-p})$ for $p \neq 2$ then $\mathrm{fat}_\varepsilon\big(\mathrm{conv}(F)\big) = O\big(\varepsilon^{-\max\{2,p\}}\big)$. If $p = 2$ then $\mathrm{fat}_\varepsilon\big(\mathrm{conv}(F)\big) = O(\varepsilon^{-2} \log^4 \frac{1}{\varepsilon})$. This result is demonstrated by analyzing the growth rate of the Rademacher averages associated with the class (see Mendelson, 2001c) for further details).*

References

N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale sensitive dimensions, uniform convergence and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

B. Carl. Metric entropy of convex hulls in hilbert spaces. *The Bulletin of the London Mathematical Society*, 29:452–458, 1997.

B. Carl, I.Kyrezi, and A. Pajor. Metric entropy of convex hulls in banach spaces. *Journal of the London Mathematical Society*, 60(2):871–896, 1999.

R.M. Dudley. Universal donsker classes and metric entropy. *Annals of Probability*, 15: 1306–1326, 1987.

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

L. Gurvits. A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces. *Theoretical Computer Science*, 261(1):81–90, 2001.

D. Haussler. Sphere packing numbers for subsets of boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory A*, 69:217–232, 1995.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer-Verlag, Berlin, 1991.

W. S. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980, 1998.

S. Mendelson. Learnability in hilbert spaces with reproducing kernels. *Journal of Complexity*, 2001a. to appear.

S. Mendelson. Learning relatively small classes. In *Proceeding of the 14th Annual Conference on Computational Learning Theory, Lecture Notes in Artificial Intelligence 2111*, pages 273–288, 2001b.

S. Mendelson. Rademacher averages and phase transitions in glivenko-cantelli classes. *IEEE Transactions on Information Theory*, 2001c. to appear.

G. Pisier. Remarques sur un resultat non publiè de B. Maurey. In Centre de Mathematique, editor, *Seminarie d'analyse fonctionelle 1980–1981*, Palaiseau, 1981.

G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge University Press, Cambridge, 1989.

N. Tomczak-Jaegermann. *Banach–Mazur distance and finite–dimensional operator ideals*. Pitman, 1989.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.