

# On the size of convex hulls of small sets

Shahar Mendelson<sup>1</sup>

## 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.

---

<sup>1</sup>Computer Sciences Laboratory, RSISE, The Australian National University, Canberra 0200, Australia. Email: shahar@csl.anu.edu.au

# 1 Introduction

Convexity plays an important role in machine learning. Its significance can be seen in both sides of the learning game. 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 [8, 10]. 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. The bound we establish has recently appeared in [3]. The main reason we chose to present an alternative proof is because the one in [3] 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. Moreover, our method of proof enables one to provide a bound on the constants appearing in the estimate, which is almost impossible in the proof presented in [3].

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 [6, 9]. The path we take is as follows: we begin by improving the known upper bounds on the fat-shattering dimension of the functionals  $\{x^* \mid \|x^*\| \leq 1\}$  (defined below) when considered as functions on the unit ball of the space. It turns out that the linear fat shattering dimension is determined by the *type* of the Banach space. We use this fact to prove the results regarding 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  $(f(\omega_i))_{i=1}^n$  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 is the same as the *linear* fat-shattering dimension of the image of  $(\omega_i)_{i=1}^n$  in the dual space. Therefore, it is possible to bound the fat-shattering dimension using "linear" methods.

The article is divided to 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 on convex hulls.

We end this introduction with some 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  $\ell_p^n$  be  $\mathbb{R}^n$  equipped with the norm  $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$  and set  $\ell_\infty^n$  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  $\chi_A$ . If  $A$  and  $B$  are sets, let  $A+B = \{a+b | a \in A, b \in B\}$ .

For any probability measure  $\mu$  on a measurable space  $(\Omega, \Sigma)$ , let  $\mathbb{E}_\mu$  denote the expectation with respect to  $\mu$ .  $L_p(\mu)$  is the set of functions which satisfy  $\mathbb{E}_\mu |f|^p < \infty$  and set  $\|f\|_{L_p(\mu)} = (\mathbb{E} |f|^p)^{1/p}$ . 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  $\delta_\omega$  be the point evaluation functional, i.e., for every function  $f$  on  $\Omega$ ,  $\delta_\omega(f) = f(\omega)$ . We denote by  $\mu_n$  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  $\mu_n$  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  $\Omega$ . We say that  $F$  shatters  $\{\omega_1, \dots, \omega_n\} \subset \Omega$ , if for every  $I \subset \{1, \dots, n\}$  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*

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

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

It is possible to use a parametric version of the VC dimension, called the fat-shattering dimension.

**Definition 1.2** For every  $\varepsilon > 0$ , a set  $A = \{\omega_1, \dots, \omega_n\} \subset \Omega$  is said to be  $\varepsilon$ -shattered by  $\mathcal{F}$  if there is some function  $s : A \rightarrow \mathbb{R}$ , such that for every  $I \subset \{1, \dots, n\}$  there is some  $f_I \in \mathcal{F}$  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

$$\text{fat}_\varepsilon(F, \Omega) = \sup \left\{ |A| \mid A \subset \Omega, A \text{ is } \varepsilon\text{-shattered by } F \right\}.$$

The set  $(s_i) = (s(\omega_i))$  is called a witness to the shattering and for every  $I \subset \{1, \dots, n\}$  we call  $f_I$  the shattering function of the set  $I$ .

If  $(X, d)$  is a metric space and if  $F \subset X$ , denote by  $N(\varepsilon, F, d)$  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  $\varepsilon$ -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  $\varepsilon$ -separated if the distance between any two elements of the set is larger than  $\varepsilon$ . Set  $D(\varepsilon, F)$  to be the maximal cardinality of an  $\varepsilon$ -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 [1], 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  $\Omega$  into  $[0, 1]$  and set  $d = \text{fat}_{\varepsilon/4}(F)$ . Then, for every empirical measure  $\mu_n$  on  $\Omega$ ,

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

In the sequel, we require several definitions originating in the theory of Banach spaces. For the basic definitions we refer the reader to [12] or [14].

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  $\| \cdot \|_K$ , and given by

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

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

**Definition 1.4** If  $F$  is a bounded subset of  $\ell_2^n$ , let

$$F^\circ = \{x \in \ell_2^n \mid \sup_{f \in F} |\langle f, x \rangle| \leq 1\},$$

where  $\langle -, - \rangle$  is the inner product in  $\ell_2^n$ . The set  $F^\circ$  is called the polar of  $F$ .

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

$$\text{absconv}(F) = \left\{ \sum_{i=1}^n a_i f_i \mid n \in \mathbb{N}, f_i \in F, \sum_{i=1}^n |a_i| = 1 \right\}.$$

It is easy to see that  $F^\circ = (\text{absconv}(F))^\circ$  and that if  $G \subset F$  then  $F^\circ \subset G^\circ$ . Note that  $F^\circ$  is the unit ball of the norm  $\|\cdot\|_{K^*}$ , where  $K = \text{absconv}(F)$ .

Given a class  $F$  and an empirical measure  $\mu_n$ , we endow  $\mathbb{R}^n$  with the Euclidean structure of  $L_2(\mu_n)$ , which is isometric to  $\ell_2^n$ . Let  $F/\mu_n$  be the image of  $F$  in  $L_2(\mu_n)$  under the inclusion operator, i.e.,

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

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

$$F/\mu_n = \left\{ n^{-1/2} \sum_{i=1}^n f(\omega_i) e_i \mid f \in F \right\}.$$

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

## 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  $\varepsilon^{-1}$ .

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  $\varepsilon^{-1}$ . 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  $N(\varepsilon, F, L_2(\mu)) = O(\varepsilon^{-p})$  for some  $p > 0$ , and the second, (which is more difficult), when  $\log N(\varepsilon, F, L_2(\mu)) = O(\varepsilon^{-p})$  for  $0 < p < 2$ . The first case was investigated by Dudley [4]. He showed that for every  $\delta > 0$ , the log-covering numbers of the convex hull of  $F$  are polynomial in  $\varepsilon^{-1}$  with exponent  $\delta + 2p(p+2)^{-1}$ . This result was improved independently by Van der Vaart and Wellner [15] and by Carl [2] who removed the superfluous  $\delta$ . Those results indicate that if  $N(\varepsilon, F, L_2(\mu)) \leq \gamma\varepsilon^{-p}$  for some  $\gamma > 0$  and  $p > 0$ , and if  $K$  is the symmetric convex hull of  $F$ , then the entropy integral  $\int_0^1 \log^{1/2} N(\varepsilon, K, L_2(\mu)) d\varepsilon$  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 [5] for more details).

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

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

Thus, it is only natural to try and find the value of that constant. 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  $\log N(\varepsilon, F, L_2(\mu)) = O(\varepsilon^{-p})$  for some  $0 < p < 2$ . We present partial answers to both questions.

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

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

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

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

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

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

The claims of theorem 2.1 were recently published in [3], 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 the an idea which was used in [15], and if fact, the first claim follows from a careful analysis of the proof in [15].

We shall require three preliminary results. The first result is due to B. Maurey [13], 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  $\text{diam}(F)$ . Then, for every  $\varepsilon > 0$ ,*

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

Let  $(\delta_i)$  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  $\delta_i$ -separated set in  $F$  which is maximal with respect to inclusion (i.e., each  $F_i$  is  $\delta_i$ -separated and if  $F_i \subset A \subset F$  then  $A$  is not  $\delta_i$ -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 reveals our strategy: 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.

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

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

and

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

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

Set  $F_i = \{x_1, \dots, x_{|F_i|}\}$  and  $F_j = \{x_1, \dots, x_{|F_j|}\} \subset F_i$ . If  $z \in \text{conv}(F_i)$ , then there are  $\lambda_i \geq 0$  such that  $\sum_{i=k}^{|F_i|} \lambda_k = 1$  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  $z_1 \in \text{conv}(F_j)$  and  $z_2 \in \text{conv}(G_{ij} \cup \{0\})$ . Hence,

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

and the first assertion is verified.

It is easy 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

$$N(\varepsilon, \text{conv}(F_i)) \leq N\left(\frac{\varepsilon}{2}, \text{conv}(F_j)\right) \cdot N\left(\frac{\varepsilon}{2}, \text{conv}(G_{ij} \cup \{0\})\right). \quad (2.1)$$

To estimate the first term, note that  $\text{span}(F_j)$  can be isometrically embedded in  $\ell_2^{|F_j|}$ . Hence, 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 straightforward volume estimate (see [12]), one can show that for every  $\varepsilon, \varepsilon'$ ,

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

Hence,

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

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

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

■

As the next two technical results indicate, one has to select an appropriate rate of decay of the “mesh” sequence  $(\delta_n)$  which define the increasing family  $(F_n)$ .

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

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

*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  $k$ ,*

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

*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 is some  $\gamma > 0$  such that for every  $\varepsilon > 0$*

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

*for some  $0 < p < 2$ . 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  $(B_k)$  such that for every integers  $n \geq 2$  and  $k \geq 1$*

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

*where  $\alpha = 4p/(2-p)$ ,  $\sup A_k \leq A'_p$  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 to Lemma 2.5. The proof to Lemma 2.4 follows by a similar argument. The idea behind the proof of Lemma 2.4 is due to Van-der-Vaart and Wellner [15]. 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  $\delta_n$  separated in  $F$ . Thus

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

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

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

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

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

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

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

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

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

as claimed.

Now, we fix some  $n \geq 2$  and use induction with respect to  $k$ . Let  $i = \lceil n^{k^\alpha} \rceil$  and  $j = \lceil n^{(k-1)^\alpha} \rceil$ . Note that  $|G_{ij} \cup \{0\}| \leq \lceil n^{k^\alpha} \rceil$  and that  $\text{diam}(G_{ij} \cup \{0\}) \leq 2\delta_j = 4\gamma^{1/p} \log^{-1/p} \lceil n^{(k-1)^\alpha} \rceil$ . By Lemma 2.2,

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

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

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

Also,

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

Using the above estimates and the definition of  $\alpha$  it follows that

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

On the other hand, by the induction hypothesis,

$$\log\left(A_{k-1} \varepsilon_n, \text{conv } F_{\lceil n^{(k-1)^\alpha} \rceil}\right) \leq B_{k-1} \frac{n^2}{\log^{\frac{4}{p}-1} n}.$$

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  $\text{conv } F_{\lceil n^{k^\alpha} \rceil}$ . Hence,

$$\log\left(A_k \varepsilon_n, \text{conv } F_{\lceil n^{k^\alpha} \rceil}\right) \leq \left(B_{k-1} + C(p, \gamma) \frac{\log k}{k^2}\right) \frac{n^2}{\log^{\frac{4}{p}-1} n}.$$

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  $n \geq 2$  and let  $\varepsilon_n = \gamma^{1/p} n^{-1/2-1/p}$ . By Lemma 2.4 it follows that for every integer  $k$ ,

$$\log N(A_k \varepsilon_n, \text{conv } F_{nk^{3p}}) \leq B_k n \leq B p n = B p \left( \frac{1}{\varepsilon_n} \right)^{\frac{2p}{2+p}}.$$

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

$$\log N(A \varepsilon_n, \text{conv } F_{nk^{3p}}) \leq \log N(A_k \varepsilon_n, \text{conv } F_{nk^{3p}}) \leq B p \left( \frac{1}{\varepsilon_n} \right)^{\frac{2p}{2+p}}.$$

Taking  $k$  to infinity,

$$\log N(A \varepsilon_n, \text{conv } F) \leq B p \left( \frac{1}{\varepsilon_n} \right)^{\frac{2p}{2+p}}.$$

The claim for a general  $\varepsilon$  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  $n \geq 2$ ,

$$\log N(A' \varepsilon_n, \text{conv } F) \leq B' \frac{n^2}{\log^{\frac{4}{p}-1} n}.$$

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

$$\log N(A' \varepsilon_n, \text{conv } F) \leq B' \frac{1}{\varepsilon_n^2} \log^{1-\frac{2}{p}} \frac{1}{\varepsilon_n}$$

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  $1/\varepsilon$  [7, 15]:

**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  $\mu$  on  $\Omega$ ,  $N(\varepsilon, F, L_2(\mu)) \leq C d (4e)^d \varepsilon^{-2d}$ .*

**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*

$$\log N(\varepsilon, \text{conv}(F), L_2(\mu)) \leq Cd \left( \frac{1}{\varepsilon} \right)^{\frac{2d}{1+d}},$$

where  $C$  is an absolute constant.

Next, we establish a similar estimate in the case where the fat shattering dimension satisfies  $\text{fat}_\varepsilon(F) = O(\varepsilon^{-p})$  for some  $0 < p < 2$ . It is possible to connect the fat shattering dimension and the  $L_2(\mu)$  covering numbers [11].

**Theorem 2.8** *Let  $F$  be a class of functions into  $[0, 1]$ . Then, for every  $1 \leq p < \infty$  there is some constant  $C_p$  such that for every probability measure  $\mu$ ,*

$$\log N(\varepsilon, F, L_2(\mu)) \leq C_p \text{fat}_{\frac{\varepsilon}{32}}(F) \log^2 \left( \frac{2 \text{fat}_{\frac{\varepsilon}{32}}(F)}{\varepsilon} \right).$$

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

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

### 3 Fat-shattering of convex hulls

In this section we investigate the fat-shattering dimension of convex hulls. First, we show that in some sense, fat-shattering is equivalent to linear fat-shattering dimension in an appropriate Banach space (Lemma 3.1). Then, we present an upper bound on the linear fat-shattering dimension in Banach spaces using their type (see definition below). Finally, we show that  $\text{absconv}(F/\mu_n)$  may be approximated by a set  $G$  which has almost the same fat-shattering dimension as  $\text{absconv}(F/\mu_n)$  but  $G^\circ$  has a relatively small type constant.

Similar to what was done in previous sections, we endow  $\mathbb{R}^n$  with the Euclidean structure of  $L_2(\mu_n)$  and denote by  $(e_i)$  the orthonormal basis in  $L_2(\mu_n)$ .

The next lemma is straightforward and its proof is omitted.

**Lemma 3.1** *Let  $S = \{\omega_1, \dots, \omega_n\}$  be a sample and let  $\mu_n$  be the empirical measure supported on  $S$ .*

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

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

**Definition 3.2** A Banach space  $X$  has Rademacher type  $p$  if there is some  $C$  such that for every integer  $n$  and every  $x_1, \dots, x_n \in X$ ,

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

where  $(\varepsilon_i)$  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  $T_p(X)$ .

### 3.1 The fat-shattering dimension in of linear functionals

**Theorem 3.3** Let  $X$  be a Banach space which has type  $p$  for some  $1 < p \leq 2$  with a type constant  $T_p(X)$ . Then

$$\text{fat}_\varepsilon(B(X^*), B(X)) \leq \left( \frac{T_p(X)}{\varepsilon} \right)^{\frac{p}{p-1}}.$$

**Proof:** Assume that the set  $\{x_1, \dots, x_n\} \subset B(X)$  is  $\varepsilon$ -shattered by  $B(X^*)$  and set  $\{s_1, \dots, s_n\}$  to be a witness to the shattering. Let  $I \subset \{1, \dots, n\}$  and set  $x_I^*$  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,

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

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

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

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

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

Thus, there is a realization of the random variables  $(\varepsilon_i)$  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 3.4** *It is possible to show that this bound is tight [9]. Indeed, one can show that if  $p^* = \sup\{p | X \text{ has type } p\}$  then for every  $\varepsilon > 0$ ,*

$$\left( \frac{1}{\varepsilon} \right)^{\frac{p^*}{p^*-1}} - 1 \leq \text{fat}_\varepsilon(B(X^*), B(X)).$$

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

$$\left( \frac{1}{\varepsilon^2} \right) - 1 \leq \text{fat}_\varepsilon(B(X^*), B(X)) \leq \left( \frac{1}{\varepsilon^2} \right).$$

## 3.2 Covering numbers and type constants

In order to bound the type-2 constant of  $(F/\mu_n)^o$ , we shall 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 [14] that there are absolute constants  $c$  and  $C$  such that for every integer  $n$ ,

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

The second fact we shall use is that if the cardinality of  $F$  is small then  $F^o$  can be isometrically embedded into  $\ell_\infty^n$  for a relatively small  $n$ .

**Lemma 3.5** *Let  $F \subset \ell_2^n$  be a finite set. Then  $F^o$  can be isometrically embedded into  $\ell_\infty^{|F|}$ .*

**Proof:** Let  $F = \{f_1, \dots, f_m\}$  and define  $T : (\mathbb{R}^n, \|\cdot\|_{F^o}) \rightarrow \ell_\infty^{|F|}$  by  $Tx^* = (x^*(f))_{i=1}^m$ . 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  $F/\mu_n$  are relatively small sets. In the real valued case, the  $L_\infty(\mu_n)$  covering numbers of  $F$  may be bounded using  $\text{fat}_\varepsilon(F)$  (theorem 1.3), whereas for VC classes, one may use the following version of Sauer's Lemma [15]:

**Lemma 3.6** *There is an absolute constant  $C$  such that if  $F$  is a class of  $\{0, 1\}$  valued functions on  $\Omega$  with  $VC(F) = d$ , then for every empirical measure  $\mu_n$ ,  $|F/\mu_n| \leq Cn^d$ .*

To simplify notation, given a class  $F$  and a fixed empirical measure  $\mu_n$ , let  $K$  be the symmetric convex hull of the set  $F/\mu_n$  and  $F^o = (F/\mu_n)^o$ . It is interesting to note that the type-2 constants of  $(F/\mu_n)^o$  characterize VC classes.

**Theorem 3.7** *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  $\mu_n$ ,  $T_2((F/\mu_n)^o) \leq C(1 + \log n)^{1/2}$ .*

**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,  $F^o$  can be isometrically embedded in  $\ell_\infty^{Cn^d}$ . It follows that  $T_2(F^o) \leq C(1 + d \log n)^{1/2}$  as claimed.

Conversely, assume that there is a constant  $C$  such that for every integer  $n$  and every empirical measure  $\mu_n$ ,  $T_2(F^o) \leq C(1 + \log n)^{1/2}$ . Let  $S = \{\omega_1, \dots, \omega_n\}$  be

shattered by  $F$  and set  $\mu_n$  to be an empirical measure supported on  $S$ . By the first part of lemma 3.1 and theorem 3.3,

$$n \leq \text{fat}_1(F^o, F) \leq C(T_2^2(F^o) + 1) \leq C(1 + \log n),$$

implying that  $VC(F, \Omega) < \infty$ . ■

**Remark 3.8** *The proof of theorem 3.7 implies that if  $F$  is a  $\{0, 1\}$  class such that  $\sup_{\mu_n} T_2(F^o) = o(n)$  then  $F$  is a VC class.*

Note that the upper bound on  $T_2(F^o)$  can not be significantly 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  $\mu_n$  of  $[-1, 1]$ ,

$$F/\mu_n = \left\{ (1, 0, \dots, 0), (1, 1, 0, \dots, 0), \dots, (1, 1, \dots, 1) \right\}.$$

Therefore,  $(F/\mu_n)^o$  is isometric to  $\ell_\infty^n$  and  $T_2((F/\mu_n)^o) = C(1 + \log n)^{1/2}$ .

Also, Theorem 3.7 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  $\ell_p$ . Given an integer  $n$ , let  $\mu_n$  be the empirical measure supported on  $\{e_1, \dots, e_n\} \subset B(\ell_p)$ . It is easy to see that  $F/\mu_n = n^{-1/2}B(\ell_q^n)$ . Therefore,  $F^o$  is isometric to  $\ell_p^n$  and  $T_2(F^o) \geq n^{\frac{1}{p}-\frac{1}{2}}$  (see [14]).

Finally, we shall 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  $n$ ,  $1 \leq p \leq \infty$  and  $\varepsilon > 0$ , let

$$N_p(n, \varepsilon) = \sup_{m \leq n} \sup_{\mu_m} N(\varepsilon, F, L_p(\mu_m)).$$

Thus,  $N_p(n, \varepsilon)$  is the supremum of the  $\varepsilon$ -covering numbers of  $F$  in empirical  $L_p$  spaces, where the empirical measure is supported on at most  $n$  elements of  $\Omega$ .

**Theorem 3.9** *Let  $F$  be a class of functions whose range is a subset of  $[0, 1]$  and set  $K$  to be its convex hull. Then, for every  $\varepsilon, \tau > 0$ ,*

$$\text{fat}_{2\varepsilon+\tau}(K, \Omega) \leq C(1 + \log N_\infty(n, 2\varepsilon)) \frac{1}{\tau^2}.$$

The following lemma indicates that an  $\varepsilon$ -cover of a set in  $L_\infty$  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  $\delta$ -shatters the set  $A = \{\omega_1, \dots, \omega_n\}$ , and let  $\mu_n$  be the empirical measure on supported  $A$ . Assume that  $\delta > 2\varepsilon$  and that  $H$  is an  $\varepsilon$ -cover of  $F/\mu_n$  in  $L_\infty(\mu_n)$ . Then,  $A' = \{\sqrt{n}e_1, \dots, \sqrt{n}e_n\} \subset F^\circ$  is  $(\delta - 2\varepsilon)$ -shattered by  $H$ , where the elements of  $H$  are viewed as linear functionals on  $F^\circ$ .*

**Proof:** By Lemma 3.1,  $A'$  is  $\delta$ -shattered by  $F/\mu_n$  and set  $(s_i)$  to be a witness to the shattering. Fix  $I \subset \{1, \dots, n\}$  and let  $f_I \in F/\mu_n$  to be the  $\delta$ -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  $i \in I$

$$\begin{aligned} \langle h_I, \sqrt{n}e_i \rangle &= \langle f_I, \sqrt{n}e_i \rangle + \langle h_I - f_I, \sqrt{n}e_i \rangle \\ &\geq s_i + \frac{\delta}{2} - \|h_I - f_I\|_{L_\infty(\mu_n)} \|\sqrt{n}e_i\|_{L_1(\mu_n)} \\ &\geq s_i + \frac{\delta}{2} - \varepsilon, \end{aligned}$$

and in a similar fashion, if  $j \in I^c$  then  $\langle h_I, \sqrt{n}e_j \rangle \leq s_j - \frac{\delta}{2} + \varepsilon$ , and our claim follows.  $\blacksquare$

**Proof of Theorem 3.9:** Fix  $\varepsilon > 0$  and  $\tau > 0$ , assume that  $A = \{\omega_1, \dots, \omega_n\}$  is  $(2\varepsilon + \tau)$ -shattered by  $K = \text{conv}(F)$  and let  $\mu_n$  be an empirical measure supported on  $A$ . Let  $H \subset F/\mu_n$  be an  $\varepsilon$ -cover to  $F/\mu_n$  in  $L_\infty(\mu_n)$  and set  $G$  to be the symmetric convex hull of  $H$ . Clearly,  $|H| \leq N(2\varepsilon, F/\mu_n, L_\infty(\mu_n))$ .

Since  $H$  is an  $\varepsilon$ -cover of  $F/\mu_n$  in  $L_\infty(\mu_n)$  then  $G$  is an  $\varepsilon$ -cover to  $K$  in  $L_\infty(\mu_n)$ . By Lemma 3.10, if the set  $A$  is  $2\varepsilon + \tau$  shattered by  $K$  then the set  $A' = \{\sqrt{n}e_1, \dots, \sqrt{n}e_n\}$  is  $\tau$ -shattered by  $G$ , when members of  $G$  are viewed as linear functionals on  $F^\circ$ . Also, note that  $A' \subset F^\circ$ . Thus,

$$n \leq \text{fat}_\tau(G, (\sqrt{n}e_i)_{i=1}^n) \leq \text{fat}_\tau(G, F^\circ).$$

Since  $H \subset F/\mu_n$ , then by taking convex hulls  $G \subset K$ . Using the properties of the polar,  $F^\circ = K^\circ \subset G^\circ$ , thus  $\text{fat}_\tau(G, F^\circ) \leq \text{fat}_\tau(G, G^\circ)$ . By Lemma 3.5 there is an absolute constant  $C$  such that

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

Hence, by Theorem 3.3,

$$\text{fat}_\tau(G, G^\circ) \leq C \left( 1 + \log N(\varepsilon, F/\mu_n, L_\infty(\mu_n)) \right) \frac{1}{\tau^2}.$$

Therefore,

$$\text{fat}_{2\varepsilon+\tau}(K, \Omega) \leq C \left(1 + \log N(\varepsilon, F/\mu_n, L_\infty(\mu_n))\right) \frac{1}{\tau^2}.$$

■

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

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

where  $K$  is the convex hull of  $F$  and  $VC(F) = d$ .

**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  $\text{fat}_{2\varepsilon+\tau}(K, \Omega) = n$ . By theorem 3.9,  $n \leq Cd\tau^{-2}(1 + \log n)$  and our claim follows.

■

It is possible to extend Corollary 3.11 and estimate the fat-shattering dimension of the convex hull of a class using the fat-shattering dimension of the class itself.

**Theorem 3.12** *There is an absolute constant  $C$  such that for any class  $F$  of functions into  $[0, 1]$  and every  $\varepsilon, \tau \in (0, 1)$ ,*

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

where  $K$  is the convex hull of  $F$ .

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

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

By theorem 3.9 and since  $n > 1$  and  $\varepsilon \in (0, 1)$ , there is an absolute constant  $C$  such that

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

Hence,

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

and the claim follows.

■

**Remark 3.13** *Using a different approach it is possible to improve Theorem 3.12 when one has “global data” on the fat-shattering dimension of the class. Indeed, if  $\text{fat}_\varepsilon(F) = O(\varepsilon^{-p})$  for  $p \neq 2$  then  $\text{fat}_\varepsilon(\text{conv}(F)) = O(\varepsilon^{-\max\{2,p\}})$ . If  $p = 2$  then  $\text{fat}_\varepsilon(\text{conv}(F)) = O(\varepsilon^{-2} \log^4 \frac{1}{\varepsilon})$ . This result is demonstrated by analyzing the growth rate of the Rademacher averages associated with the class (see [11] for further details).*

## References

- [1] 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.
- [2] B. Carl: Metric entropy of convex hulls in Hilbert spaces, Bull. London Math. Soc., 29, 452–458, 1997
- [3] B. Carl, I.Kyrezi, A. Pajor: Metric entropy of convex hulls in Banach spaces, J. London. Math. Soc. 60, 2, 871-896, 1999.
- [4] R.M. Dudley: Universal Donsker classes and metric entropy, Annals of Probability 15, 1306-1326, 1987.
- [5] R.M. Dudley: *Uniform Central Limit Theorems* Cambridge Studies in Advanced Mathematics 63, Cambridge University Press 1999
- [6] L. Gurvits: A note on the scale-sensitive dimension of linear bounded functionals in Banach spaces, NEC Res. Inst. Technical Report, 1997
- [7] D. Haussler: Sphere packing numbers for subsets of Boolean  $n$ -cube with bounded Vapnik-Chervonenkis dimension, Journal of Combinatorial Theory A 69, 217-232.
- [8] W.S.Lee, P.L. Bartlett, R.C. Williamson: The Importance of Convexity in Learning with Squared Loss, IEEE Transactions on Information Theory 44 5, 1974-1980, 1998.
- [9] S. Mendelson: Learnability in Hilbert Spaces with Reproducing Kernels, to appear in Journal of Complexity.
- [10] S. Mendelson: Learning relatively small classes, preprint.
- [11] S. Mendelson: Rademacher averages and phase transitions in Glivenko-Cantelli classes, preprint.

- [12] G. Pisier: *The volume of convex bodies and Banach space geometry*, Cambridge University Press, 1989.
- [13] G. Pisier: Remarques sur un résultat non publié de B. Maurey, Séminaire d'analyse fonctionnelle, 1980-1981, exposé No. 5, École Polytechnique, Palaiseau
- [14] N. Tomczak-Jaegermann: *Banach-Mazur distance and finite-dimensional operator Ideals*, Pitman monographs and surveys in pure and applied Mathematics 38, 1989
- [15] A.W. Van-der-Vaart, J.A. Wellner: *Weak convergence and Empirical Processes*, Springer-Verlag, 1996.