Rademacher Averages and Phase Transitions in Glivenko–Cantelli Classes

Shahar Mendelson

***Abstract*—We introduce a new parameter which may replace the fat-shattering dimension. Using this parameter we are able to provide improved complexity estimates for the agnostic learning problem with respect to any L_p norm. Moreover, we show that if $\mathrm{fat}_\varepsilon(F) = O(\varepsilon^{-p})$ then F displays a clear phase transition which occurs at $p = 2$. The phase transition appears in the sample complexity estimates, covering numbers estimates, and in the growth rate of the Rademacher averages associated with the class. As a part of our discussion, we prove the best known estimates on the covering numbers of a class when considered as a subset of L_p spaces. We also estimate the fat-shattering dimension of the convex hull of a given class. Both these estimates are given in terms of the fat-shattering dimension of the original class.**

***Index Terms*—Fat-shattering dimension, Rademacher averages, uniform Glivenko–Cantelli (GC) classes.**

I. Introduction

CLASSES of functions that satisfy the law of large numbers uniformly, i.e., the Glivenko–Cantelli classes, have been thoroughly investigated in the last 30 years.

Formally, the question at hand is as follows: if F is a class of functions on some set Ω, when is it possible to have that for every $\varepsilon > 0$

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

where the supremum is taken with respect to all probability measures μ, X_i are independently sampled according to μ, and $\mathbb{E}_\mu$ is the expectation with respect to μ.

Clearly, the "larger" F is, the less likely it is that it satisfies this uniform law of large numbers. In the sequel, we will always assume that the set consists of functions with a uniformly bounded range.

The problem, besides being intriguing from the theoretical point of view, has important applications in Statistics and in Learning Theory. To demonstrate this, note that (1.1) may be formulated in a "quantified" manner; namely, for every $\varepsilon > 0$ and $0 < \delta < 1$, there exists some integer n_0, such that for every probability measure μ and every $n \geq n_0$

$$\mu\left\{ \sup_{f\in F} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_\mu f \right| \geq \varepsilon \right\} \leq \delta. \tag{1.2}$$

Manuscript received November 3, 2000; revised June 19, 2001.

The author is with the Computer Sciences Laboratory, RSISE, The Australian National University, Canberra 0200, Australia (e-mail: shahar@csl.anu.edu.au).

Communicated by G. Lugosi, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier S 0018-9448(02)00051-2.

For every $\varepsilon > 0$ and $0 \leq \delta < 1$, the smallest possible integer n such that (1.2) is satisfied is called the Glivenko–Cantelli sample complexity estimate associated with the pair ε, δ.

In a *learning problem*, one wishes to find the best approximation of an unknown function by a member of a given set of functions. This approximation is carried out with respect to an $L_p(\mu)$ norm, where μ is an *unknown* probability measure. If one knows that the set is a Glivenko–Cantelli class, then it is possible to reduce the problem to a finite-dimensional approximation problem. Indeed, if one uses a "large enough" sample, and if one is able to find a member of the class which is "close" to the unknown function on the sample points, then with high probability it will also be close to that function in $L_p(\mu)$. Hence, an "almost minimizer" of the L_p empirical distances between the unknown function and the members of the class will be, with high probability, an "almost minimizer" with respect to the $L_p(\mu)$ norm. The terms "close," "high probability," and "large enough" can be made precise using the learning parameters ε and δ and the sample complexity n, respectively.

The method normally used to obtain sample complexity estimates (and proving that a set of functions is indeed a Glivenko–Cantelli class) is to apply covering number estimates. It is possible to show (see Section II or [7] for further details) that the growth rates of the covering numbers of the set in certain L_p spaces characterizes whether or not it is a Glivenko–Cantelli class. Moreover, it is possible to provide sample complexity estimates in terms of the covering numbers.

Though it seems a hard task to estimate the covering numbers of a given set of functions, it is possible to do so using combinatorial parameters, such as the Vapnik–Chervonenkis (VC) dimension for $\{0, 1\}$-valued functions or the fat-shattering dimension in the real-valued case. Those parameters may be used to bound the covering numbers of the class in appropriate L_p spaces, and it is possible to show [23], [2] that they are finite if and only if the set is a Glivenko–Cantelli class.

The goal of this paper is to define another parameter which may replace the combinatorial parameters and, in fact, by using it, may enable one to obtain significantly improved complexity estimates.

This parameter originates from the original proof of the Glivenko–Cantelli theorem, which uses the idea of symmetry. Recall (see, e.g., [10]) that if μ is a probability measure and if X_i are selected independently according to μ, then

$$\mu\left\{ \sup_{f\in F} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_\mu f \right| \geq \varepsilon \right\} \leq 4\tau\left\{ \sup_{f\in F} \left| \sum_{i=1}^{n} r_i f(X_i) \right| \geq \frac{n\varepsilon}{4} \right\}$$

where (r_i) are independent Rademacher random variables on some probability space (Y, Σ', ν) (that is, (r_i) are $\{-1, 1\}$-valued independent symmetric random variables), and τ is the product measure $\mu \times \nu$.

Moreover, the Rademacher averages control the rate of decay of the expected deviation [9]. Indeed, given a measure μ, it is possible to show that if F is a class of functions into $[-M, M]$, then for every integer n

$$\begin{aligned}\mathbb{E}_\mu \sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_\mu f\right| &\le 2\mathbb{E}_\tau \sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^n r_i f(X_i)\right| \\ &\le 2M\mathbb{E}_\mu \sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_\mu f\right| + \mathbb{E}_\nu\left|\frac{1}{n}\sum_{i=1}^n r_i\right|.\end{aligned}$$

In fact, one can show the following.

Theorem 1.1 [9]: Let F be a class of uniformly bounded functions. Then, F is a Glivenko–Cantelli class if and only if

$$\sup_\mu \mathbb{E}_\mu \mathbb{E}_\nu \sup_{f\in F}\left|\sum_{i=1}^n r_i f(X_i)\right| = o(n).$$

If $\{\omega_1, \ldots, \omega_n\}$ is a sample, one can define the Rademacher average associated with that sample by

$$R_n(F/\mu_n) = \frac{1}{\sqrt{n}}\,\mathbb{E}\sup_{f\in F}\left|\sum_{i=1}^n r_i f(\omega_i)\right|$$

where μ_n is the empirical measure supported on the set $\{\omega_1, \ldots, \omega_n\}$. In this paper, we examine the behavior of the supremum of all possible averages of n elements as a function of n. We define a parameter which measures the rate by which those averages increase as a function of n and compare it to the fat-shattering dimension.

The course of action we take is as follows. First, in Section III, we investigate the behavior of the covering numbers of a class when it is considered as a subset of $L_p(\mu_n)$ for an empirical measure μ_n which is supported on a set consisting of at most n elements. We improve the bound on the covering numbers of the class in terms of its fat-shattering dimension. We prove that for every $1 \le p < \infty$ there is a constant c_p such that for any class of functions F into $[-1, 1]$, any empirical measure μ_n, and every $\varepsilon > 0$, the covering numbers of F in $L_p(\mu_n)$ satisfy that

$$\log N(\varepsilon, F, L_p(\mu_n)) \le c_p \mathrm{fat}_{\frac{\varepsilon}{8}}(F)\log^2\left(\frac{2\mathrm{fat}_{\frac{\varepsilon}{8}}(F)}{\varepsilon}\right).$$

Note that the bound we establish is both dimension-free (independent of n) and, up to a logarithmic factor, linear in the fat-shattering dimension. From this we derive several corollaries, the most important of which is an upper estimate on the Rademacher averages associated with the class in terms of the fat-shattering dimension, at least in cases where the fat-shattering dimension is polynomial in (ε^{-1}).

The results we obtain indicate that if $\mathrm{fat}_\varepsilon(F) \le C\varepsilon^{-p}$, then the behavior of the class changes dramatically at $p = 2$. This phase transition appears in the covering numbers estimates, as well as in the growth rate of the Rademacher averages. For example, if $p < 2$, the Rademacher averages are uniformly bounded, whereas if $p > 2$, they may grow at a rate of $n^{\frac{1}{2}-\frac{1}{p}}$, and this bound is tight.

In Section IV, we define a new scale-sensitive parameter which measures the growth rate of the Rademacher averages, called $\mathrm{rav}_\varepsilon(F)$. We present upper and lower bounds on the fat-shattering dimension of F in terms of $\mathrm{rav}_\varepsilon(F)$. This yields a sharper characterization of Glivenko–Cantelli classes than that of Theorem 1.1. Then, we use the fact that the Rademacher averages remain unchanged if one takes the convex hull of the class to establish the best known estimates on the fat-shattering dimension of a convex hull. For example, we show that if $\mathrm{fat}_\varepsilon(F) = O(\varepsilon^{-p})$ then $\mathrm{fat}_\varepsilon(\mathrm{conv}(F)) = O(\varepsilon^{-\max\{2,p\}})$. Another application of our results is a new partial solution to a question in the geometry of Banach spaces which was posed by Elton [8].

Finally, in Section V, we use one of Talagrand's results [21] and prove complexity estimates with respect to any L_q norm for $1 \le q < \infty$. We show that if $\mathrm{fat}_\varepsilon(F) = O(\varepsilon^{-p})$ then the Glivenko–Cantelli sample complexity with respect to any L_q norm is $O(\varepsilon^{-\max\{2,p\}})$, up to a logarithmic factor in $1/\varepsilon$ and $1/\delta$. The complexity estimates we obtain are sharper than the known estimates, and we show that they are optimal if $p > 2$.

II. Preliminaries

We begin with some 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 $||x^*||_{X^*} = \sup_{||x||_X=1}|x^*(x)|$. Let $B(X)$ be the unit ball of X. If $1 \le p < \infty$, let ℓ_p^n be $\mathbb{R}^n$ with respect to the norm

$$||x||_p = \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}$$

and set ℓ_∞^n to be $\mathbb{R}^n$ endowed with the sup norm.

If F is a class of functions, denote by $\ell_\infty(F)$ the set of all bounded functions defined on F. Given $\mathcal{G} \in \ell_\infty(F)$, set

$$||\mathcal{G}||_{\ell_\infty(F)} = \sup_{f\in F}|\mathcal{G}(f)|.$$

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 $||f||_{L_p(\mu)} = (\mathbb{E}_\mu|f|^p)^{1/p}$. $L_\infty(\Omega)$ is the space of bounded functions on Ω, with respect to the norm $||f||_\infty = \sup_{\omega\in\Omega}|f(\omega)|$. 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 shall denote by μ_n an empirical measure supported on a set of n points, hence, $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{\omega_i}$. Given a set A, let $|A|$ be its cardinality, set χ_A to be its characteristic function, and denote by A^c the complement of A. Throughout this paper, all absolute constants are assumed to be positive and are denoted by C or c. Their values may change from line to line or even within the same line.

Given a probability measure μ on Ω, let $\Pr$ be the infinite product measure μ^∞. Uniform Glivenko–Cantelli classes (defined below) are classes of functions on Ω, for which, with high probability, random empirical measures approximate the measure μ uniformly on the elements of the class.

Definition 2.1: Let (Ω, Σ) be a measurable space. A family of measurable functions F on Ω is called a Glivenko–Cantelli class with respect to a family of measures Λ if, for every $\varepsilon > 0$

$$\lim_{n\to\infty} \sup_{\mu\in\Lambda} \Pr\left\{ \sup_{m>n} \sup_{f\in F} |\mathbb{E}_\mu f - \mathbb{E}_{\mu_m} f| \geq \varepsilon \right\} = 0$$

where μ_m is the empirical measure supported on the first m coordinates of the sample. We say that F is a *uniform Glivenko–Cantelli class* if Λ may be selected as the set of all probability measures on (Ω, Σ).

In this paper, we shall refer to Uniform Glivenko–Cantelli classes by the abbreviation "GC classes."

Note that the randomness in Definition 2.1 is in the selection of the empirical measure μ_n, since its atoms are the first n coordinates of a randomly selected sample.

To avoid measurability problems that might be caused by the supremum, one usually uses an outer measure in the definition of GC classes [6]. Actually, only a rather weak assumption (called "image admissibility Suslin") is needed to avoid the measurability problem [7]. We assume henceforth that all the classes we encounter satisfy this condition.

Given two functions f, g and some $1 \leq q < \infty$, let $\mathcal{L}(f, g, q)$ be the q-loss function associated with f and g. Thus, $\mathcal{L}(f, g, q) = |f - g|^q$. Given a class F, a function g, and some $1 \leq q < \infty$, let the q-loss class associated with F and g be

$$\mathcal{L}(F, g, q) = \{|f - g|^q | f \in F\}.$$

For every function g, $\varepsilon > 0$, $0 < \delta \leq 1$, and $1 \leq q < \infty$, set $S_q(\varepsilon, \delta, g)$ to be the GC sample complexity of the loss class $\mathcal{L}(F, g, q)$, that is, the smallest n_0 such that for every $n \geq n_0$

$$\sup_\mu \Pr\left\{ \sup_{f\in F} |\mathbb{E}_\mu (f-g)^q - \mathbb{E}_{\mu_n}(f-g)^q| \geq \varepsilon \right\} \leq \delta.$$

One possibility of characterizing GC classes is through the *covering numbers* of the class in $L_p(\mu_n)$ spaces.

Recall that if (X, d) is a metric space and if $F \subset X$, the ε-covering number of F, denoted by $N(\varepsilon, F, d)$, is the minimal number of open balls with radius $\varepsilon > 0$ (with respect to the metric d) needed to cover 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, where $B(a, \varepsilon)$ is the open ball of radius ε centered at a. In cases where the metric d is clear, we shall denote the covering numbers of F by $N(\varepsilon, F)$.

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. $D(\varepsilon, F)$ are called the packing numbers of F (with respect to the fixed metric d). It is easy to see that $N(\varepsilon, F) \leq D(\varepsilon, F) \leq N(\varepsilon/2, F)$.

There are several results which connect the uniform GC condition of a given class of functions to estimates on the covering numbers of that class. All the results are stated for classes of functions whose absolute value is bounded by 1. The results remain valid for classes of functions with a uniformly bounded range—up to a constant which depends only on that bound.

The next result is due to Dudley, Giné, and Zinn [7].

Theorem 2.2: Let F be a class of functions which map Ω into $[-1, 1]$. Then, F is a GC class if and only if for every $\varepsilon > 0$

$$\sup_{\mu_n} \log N(\varepsilon, F, L_\infty(\mu_n)) = o(n)$$

where the supremum is taken with respect to all empirical measures supported on samples which consist of at most n elements.

Similarly, F is a GC class if and only if for every $\varepsilon > 0$ and $1 \leq p < \infty$

$$\sup_{\mu_n} \log N(\varepsilon, F, L_p(\mu_n)) = o(n).$$

Other important parameters used to analyze GC classes are of a combinatorial nature. Such a parameter was first introduced by Vapnik and Chervonenkis for classes of $\{0, 1\}$-valued functions [23]. Later, this parameter was generalized in various fashions. The parameter which we focus on is the *fat-shattering dimension*.

Definition 2.3: For every $\varepsilon > 0$, a set $A = \{\omega_1, \ldots, \omega_n\} \subset \Omega$ is said to be ε-shattered by F if there is some function $s: A \to \mathbb{R}$, such that for every $I \subset \{1, \ldots, n\}$ 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

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

f_I is called the shattering function of the set I and the set $\{s(\omega_i) | \omega_i \in A\}$ is called a witness to the ε-shattering.

The connection between GC classes and the combinatorial parameters defined above is the following fundamental result [2]:

Theorem 2.4: Let F be a class of functions on Ω. If F is a class of uniformly bounded real-valued functions, then it is a uniform GC class if and only if it has a finite fat-shattering dimension for every $\varepsilon > 0$.

The following result, which is also due to Alon, Ben-David, Cesa-Bianchi, and Haussler [2], enables one to estimate the $L_\infty(\mu_n)$ covering numbers of GC classes in terms of the fat-shattering dimension.

Theorem 2.5: 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

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

In particular, the same estimate holds in $L_2(\mu_n)$.

Note that although $\log D(\varepsilon, F, L_p(\mu_n))$ is almost linear in $\mathrm{fat}_\varepsilon(F)$, this estimate is not dimension-free.

It seems that the fat-shattering dimension governs the growth rate of the covering numbers. Another indication in that direction is the fact that it is possible to provide a lower bound on the covering numbers in empirical L_p spaces [1].

Theorem 2.6: Let F be a class of functions. Then, for any $\varepsilon > 0$,

$$\sup_{\mu_n} N(\varepsilon, F, L_1(\mu_n)) \geq e^{\mathrm{fat}_{16\varepsilon}(F)/8},$$

for $n \geq \mathrm{fat}_{16\varepsilon}(F)$.

In the sequel, we require several definitions originating from the theory of Banach spaces. For the basic definitions we refer the reader to [18] or [22].

Let ℓ_2^n be a real n-dimensional inner product space. We denote the inner product by $\langle\,,\,\rangle$. 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, 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 a convex, symmetric set 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 2.7: If F is a bounded subset of ℓ_2^n, let

$$F^\circ = \left\{ x \in \ell_2^n \,\middle|\, \sup_{f\in F} |\langle f, x\rangle| \le 1 \right\}.$$

F° is called the polar of F.

It is easy to see that F° is the unit ball of the norm $\|\ \|_{K^*}$, where K is the symmetric convex hull of F, denoted by $\mathrm{absconv}(F)$. Formally

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

Given a class F and an empirical measure μ_n, we endow $\mathbb{R}^n$ with the Euclidean structure of $L_2(\mu_n)$, which is isometric to ℓ_2^n. Let F/μ_n be the image of F in $L_2(\mu_n)$ under the inclusion operator. Thus,

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

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

$$F/\mu_n = \left\{ n^{-\frac{1}{2}} \sum_{i=1}^n f(\omega_i)e_i \,\middle|\, f \in F \right\} \subset \ell_2^n$$

where (e_i) is an orthonormal basis in ℓ_2^n.

Note that if $f_1, f_2 \in F$ and if μ_n is the empirical measure supported on the sample $\{\omega_1, \ldots, \omega_n\}$, then

$$\begin{aligned}\|f_1/\mu_n - f_2/\mu_n\|_{\ell_2^n}^2 &= \frac{1}{n}\sum_{i=1}^n (f_1(\omega_i) - f_2(\omega_i))^2 \\ &= \|f_1 - f_2\|_{L_2(\mu_n)}^2.\end{aligned}$$

Throughout this paper, 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 main tools we use are probabilistic averaging techniques. To that end, we define Gaussian and Rademacher averages of a subset of ℓ_2^n.

Definition 2.8: For $F \subset \ell_2^n$, let

$$\ell(F) = \mathbb{E}\left\| \sum_{i=1}^n g_i e_i \right\|_{F^\circ} \tag{2.1}$$

and

$$R(F) = \mathbb{E}\left\| \sum_{i=1}^n r_i e_i \right\|_{F^\circ} \tag{2.2}$$

where $(e_i)_{i=1}^n$ is an orthonormal basis of ℓ_2^n, $(g_i)_{i=1}^n$ are independent standard Gaussian random variables, and $(r_i)_{i=1}^n$ are independent Rademacher random variables.

In the sequel, we will be interested in sets of the form F/μ_n. Note that if $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{\omega_i}$ then

$$\begin{aligned}\ell(F/\mu_n) &= \mathbb{E}\left\| \sum_{i=1}^n g_i e_i \right\|_{F^\circ} \\ &= \mathbb{E}\sup_{f\in F/\mu_n} \left| \left\langle f, \sum_{i=1}^n g_i e_i \right\rangle \right| \\ &= \mathbb{E}\sup_{f\in F/\mu_n} \left| \left\langle \frac{1}{\sqrt{n}}\sum_{i=1}^n f(\omega_i)e_i, \sum_{i=1}^n g_i e_i \right\rangle \right| \\ &= \frac{1}{\sqrt{n}}\,\mathbb{E}\sup_{f\in F}\left| \sum_{i=1}^n g_i f(\omega_i) \right|.\end{aligned}$$

In a similar fashion

$$R(F/\mu_n) = \frac{1}{\sqrt{n}}\,\mathbb{E}\sup_{f\in F}\left| \sum_{i=1}^n r_i f(\omega_i) \right|.$$

Remark 2.9: It is important to note that the Rademacher and Gaussian averages do not change if one takes the convex hull of F. Therefore,

$$R(F) = R(\mathrm{absconv}(F)) \quad \text{and} \quad \ell(F) = ((\mathrm{absconv}(F)).$$

It is known that Gaussian and Rademacher averages are closely related, even in a much more general context than the one used here (for further details, see [22] or [13]). All we shall use is the following connection.

Theorem 2.10: There is an absolute constant C such that for every integer n and every $F \subset \ell_2^n$, $CR(F) \le \ell(F)$.

The following deep result provides a connection between the ℓ-norm of a set and its covering numbers in ℓ_2^n. The upper bound was established by Dudley in [5] while the lower bound is due to Sudakov [19]. A proof of both bounds may be found in [18].

Theorem 2.11: There are absolute positive constants c and C, such that for any $F \subset \ell_2^n$

$$\begin{aligned}c\sup_{\varepsilon>0}\varepsilon\log^{\frac{1}{2}}(N(\varepsilon, F, \ell_2^n)) \\ \le \ell(F) \le C\int_0^\infty \log^{\frac{1}{2}}(N(\varepsilon, F, \ell_2^n))\,d\varepsilon.\end{aligned}$$

Hence, there are absolute constants C and c such that for any class of uniformly bounded functions F and any empirical measure μ_n

$$\begin{aligned}c\sup_{\varepsilon>0}\varepsilon\log^{\frac{1}{2}}(N(\varepsilon, F, L_2(\mu_n))) \\ \le \ell(F/\mu_n) \le C\int_0^\infty \log^{\frac{1}{2}}(N(\varepsilon, F, L_2(\mu_n)))\,d\varepsilon.\end{aligned}$$

III. The Covering Theorem and Its Applications

The main result presented in this section is an estimate on the covering numbers of a GC class when considered as a subset of $L_p(\mu)$, for an arbitrary probability measure μ. The estimate is based on the fat-shattering dimension of the class, and the goal is to produce a dimension-free estimate which is "almost" linear in $\mathrm{fat}_\varepsilon(F)$. Thus far, the only way to obtain such a result in every L_p space was through the L_∞ estimates (Theorem 2.5).

Unfortunately, those estimates may be applied only in the case where μ is an empirical measure supported on a set I, and carry a factor of $\log^2 |I|$. Hence, the estimate one obtains is not dimension-free. There are dimension-free results similar to those obtained here, but only with respect to the L_1 norm [3].

The proof we present here is based on a result which is due to Pajor [17]. First, we demonstrate that if μ_n is supported on $\{\omega_1, \ldots, \omega_n\}$ and if a set $F \subset B(L_\infty(\Omega))$ (that is, a subset of the $L_\infty(\Omega)$ unit ball) is well separated in $L_2(\mu_n)$, then there is a "small" subset $I \subset \{\omega_1, \ldots, \omega_n\}$ such that F is "well separated" in $L_\infty(I)$. The next step in the proof is to apply the bound on the packing numbers of F in $L_\infty(I)$ in terms of the fat-shattering dimension of F. Our result is stronger than Pajor's because we use a sharper upper bound on the packing numbers.

Lemma 3.1: Let $F \subset B(L_\infty(\Omega))$ and suppose that μ_n is the empirical measure supported on $A = \{\omega_1, \ldots, \omega_n\}$. Fix $\varepsilon > 0$ and $p \geq 1$, set $d_p = D(\varepsilon, F, L_p(\mu_n))$, and assume that $d_p > 1$. Then, there is a constant c_p that depends only on p, and a subset $I \subset A$, such that

$$|I| \leq c_p \frac{\log d_p}{\varepsilon^p}$$

and

$$\log D(\varepsilon, F, L_p(\mu_n)) \leq \log D(\varepsilon/2, F, L_\infty(I)).$$

Proof: Fix any integer n and $p \geq 1$ and let $\{f_1, \ldots, f_{d_p}\} \subset F$ be ε-separated in $L_p(\mu_n)$. Hence, for every $i \neq j$

$$\varepsilon^p < \frac{1}{n} \sum_{k=1}^{n} |f_i(\omega_k) - f_j(\omega_k)|^p.$$

Let $L(i, j)$ be the set of indexes on which $|f_i(\omega_k) - f_j(\omega_k)| \leq \varepsilon/2$. Note that for every $i \neq j$

$$\begin{aligned} n\varepsilon^p &\leq \sum_{k=1}^{n} |f_i(\omega_k) - f_j(\omega_k)|^p \\ &= \sum_{k \in L(i,j)} |f_i(\omega_k) - f_j(\omega_k)|^p \\ &\quad + \sum_{k \in L(i,j)^c} |f_i(\omega_k) - f_j(\omega_k)|^p \\ &\leq |L(i, j)| \left(\frac{\varepsilon}{2}\right)^p + 2^p(n - |L(i, j)|). \end{aligned}$$

A straightforward computation shows that

$$|L(i, j)| \leq \left(1 - \left(\frac{2^p - 1}{4^p}\right)\varepsilon^p\right) n.$$

Let $(X_k)_{1 \leq k \leq t}$ be t independent random variables, uniformly distributed on $\{1, \ldots, n\}$. Clearly, for every pair $i < j$, the probability that for every $1 \leq k \leq t$, $X_k \in L(i, j)$ is smaller than $(1 - (\frac{2^p - 1}{4^p})\varepsilon^p)^t$. Therefore, the probability that there is a pair $i < j$ such that for every $1 \leq k \leq t$, $X_k \in L(i, j)$, is smaller than

$$\frac{d_p(d_p - 1)}{2}\left(1 - \left(\frac{2^p - 1}{4^p}\right)\varepsilon^p\right)^t =: \Theta.$$

If $\Theta < 1$, there is a set $I \subset \{\omega_1, \ldots, \omega_n\}$ such that $|I| \leq t$ and for every $i \neq j$, $\|f_i - f_j\|_{L_\infty(I)} \geq \varepsilon/2$, as claimed. Thus, all it requires is that

$$t \geq c_p \frac{\log d_p}{\varepsilon^p}$$

where c_p is a constant which depends only on p, and our claim follows. □

Theorem 3.2: If $F \subset B(L_\infty(\Omega))$ then for every $p \geq 1$ there is some constant c_p, which depends only on p, such that for every empirical measure μ_n and every $\varepsilon > 0$

$$\log D(\varepsilon, F, L_p(\mu_n)) \leq c_p \mathrm{fat}_{\frac{\varepsilon}{8}}(F) \log^2\left(\frac{2\mathrm{fat}_{\frac{\varepsilon}{8}}(F)}{\varepsilon}\right).$$

Proof: Fix $\varepsilon > 0$. By Lemma 3.1 and Theorem 2.5, there is a subset $I \subset \{\omega_1, \ldots, \omega_n\}$ such that

$$|I| \leq c_p \frac{\log D(\varepsilon, F, L_p(\mu_n))}{\varepsilon^p}$$

and

$$\begin{aligned} &\log D(\varepsilon, F, L_p(\mu_n)) \\ &\quad \leq \log D\left(\frac{\varepsilon}{2}, F, L_\infty(I)\right) \\ &\quad \leq c_p \mathrm{fat}_{\frac{\varepsilon}{8}}(F) \log^2\left(\frac{2\log D(\varepsilon, F, L_p(\mu_n))}{\varepsilon}\right). \end{aligned}$$

Therefore,

$$\log D(\varepsilon, F, L_p(\mu_n)) \leq c_p \mathrm{fat}_{\frac{\varepsilon}{8}}(F) \log^2\left(\frac{2\mathrm{fat}_{\frac{\varepsilon}{4}}(F)}{\varepsilon}\right)$$

as claimed. □

A. First Phase Transition: Universal Central Limit Theorem

The first application of Theorem 3.2 is that if $\mathrm{fat}_\varepsilon(F) = O(\varepsilon^{-p})$ for some $0 < p < 2$ then F is a *universal Donsker class*, that is, it satisfies the uniform central limit theorem for every probability measure. We shall not present all the necessary definitions, but rather, refer the reader to [6] or [9] for the required information.

Definition 3.3: Let $F \subset B(L_\infty(\Omega))$, set P to be a probability measure on Ω, and assume G_P to be a Gaussian process indexed by F which has mean 0 and covariance

$$\mathbb{E}G_P(f)G_P(g) = \int fgdP - \int fdP \int gdP.$$

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

It is possible to show that if F satisfies certain measurability conditions (which we omit) and if F is a universal Donsker class then

$$\sup_{f \in F} \left| n^{\frac{1}{2}}(\mathbb{E}_{\mu_n} f - \mathbb{E}_\mu f) \right| \to \sup_{f \in F} G_P(f)$$

as $n \to \infty$, where the convergence is in distribution. Moreover, the universal Donsker property is connected to covering numbers estimates.

Theorem 3.4 [6]: Let $F \subset B(L_\infty(\Omega))$. If

$$\int_0^\infty \sup_n \sup_{\mu_n} \log^{\frac{1}{2}} N(\varepsilon, F, L_2(\mu_n))\, d\varepsilon < \infty$$

then F is a universal Donsker class. On the other hand, if F is a Donsker class then there is some constant C such that for every $\varepsilon > 0$

$$\sup_n \sup_{\mu_n} \log N(\varepsilon, F, L_2(\mu_n)) \le \frac{C}{\varepsilon^2}.$$

The sufficient condition in the theorem above is called *Pollard's entropy condition*.

Lemma 3.5: Let $F \subset B(L_\infty(\Omega))$ such that $\mathrm{fat}_\varepsilon(F) = O(\varepsilon^{-p})$ for some $0 < p < 2$ and $0 < \varepsilon \le 1$. Then

$$\int_0^\infty \sup_n \sup_{\mu_n} \log^{1/2} N(\varepsilon, F, L_2(\mu_n))\, d\varepsilon < \infty. \tag{3.1}$$

This Lemma follows immediately from Theorem 3.2.

Corollary 3.6: Let $F \subset B(L_\infty(\Omega))$. If there is some constant C such that $\mathrm{fat}_\varepsilon(F) \le C\varepsilon^{-p}$ for $0 < p < 2$, then F is a universal Donsker class. On the other hand, if $\mathrm{fat}_\varepsilon(F) \ge C\varepsilon^{-p}$ for $p > 2$ then F is not a universal Donsker class.

Proof: The first part of our claim follows by Lemma 3.5, since F satisfies Pollard's entropy condition. For the second part, recall that by Theorem 2.6

$$\sup_{\mu_n} \log N(\varepsilon, F, L_2(\mu_n)) \ge \frac{\mathrm{fat}_{16\varepsilon}(F)}{8}$$

provided that $n \ge \mathrm{fat}_{16\varepsilon}(F)$. Therefore, for any $\varepsilon > 0$

$$\sup_n \sup_{\mu_n} \log N(\varepsilon, F, L_2(\mu_n)) \ge \frac{c_p}{\varepsilon^p}$$

for $p > 2$. But, if F is a Donsker class then

$$sup_n \sup_{\mu_n} \log N(\varepsilon, F, L_2(\mu_n)) = O(\varepsilon^{-2})$$

arriving to a contradiction. □

B. ℓ-Norm Estimates

We now establish bounds on the empirical ℓ-norms of function classes, based on their fat-shattering dimension. The estimates are established via an indirect route using the estimate on the $L_2(\mu_n)$ covering numbers proved in Theorem 3.2.

We begin with the following lemma, which is based on the proof of the upper bound in Theorem 2.11 (see [18]). Exactly the same argument was used in [14], so its details are omitted.

Lemma 3.7: Let μ_n be an empirical measure on Ω, put $F \subset B(L_\infty(\Omega))$ and set $(\varepsilon_k)_{k=0}^\infty$ to be a monotone sequence decreasing to 0 such that $\varepsilon_0 = 1$. Then, there is an absolute constant C such that for every integer N

$$\ell(F/\mu_n) \le C \sum_{k=1}^N \varepsilon_{k-1} \log^{\frac{1}{2}} N(\varepsilon_k, F, L_2(\mu_n)) + 2\varepsilon_N n^{\frac{1}{2}}.$$

In particular,

$$\ell(F/\mu_n) \le C \sum_{k=1}^N \varepsilon_{k-1} \mathrm{fat}^{\frac{1}{2}}_{\frac{\varepsilon_k}{8}}(F) \log\left(\frac{2\mathrm{fat}_{\frac{\varepsilon_k}{8}}(F)}{\varepsilon}\right) + 2\varepsilon_N n^{\frac{1}{2}}. \tag{3.2}$$

The latter part of Lemma 3.7 follows from its first part and Theorem 3.2.

Theorem 3.8: Let $F \subset B(L_\infty(\Omega))$ and assume that there is some $\gamma > 1$ such that for any $\varepsilon > 0$, $\mathrm{fat}_\varepsilon(F) \le \gamma\varepsilon^{-p}$. Then, there are absolute constants C_p, which depend only on p, such that for any empirical measure μ_n

$$\ell(F/\mu_n) \le \begin{cases} C_p \gamma^{\frac{1}{2}} \log \gamma, & \text{if } 0 < p < 2 \\ C_2(\gamma^{\frac{1}{2}} \log \gamma) \log^2 n, & \text{if } p = 2 \\ C_p(\gamma^{\frac{1}{2}} \log \gamma) n^{\frac{1}{2}-\frac{1}{p}}, & \text{if } p > 2. \end{cases}$$

Proof: Let μ_n be an empirical measure on Ω. If $p < 2$, then by Theorem 3.2

$$\int_0^\infty \log^{\frac{1}{2}} N(\varepsilon, F, L_2(\mu_n))\, d\varepsilon \le C_p \gamma^{\frac{1}{2}} \log \gamma$$

and the bound on the ℓ-norm follows from the upper bound in Theorem 2.11.

Assume that $p \ge 2$ and let ε_k and N be as in Lemma 3.7. Select $\varepsilon_k = 2^{-k}$ and $N = p^{-1} \log n$. By (3.2)

$$\begin{aligned} \ell(F/\mu_n) &\le C_p \gamma^{\frac{1}{2}} \log \gamma \sum_{i=1}^N \varepsilon_k^{1-\frac{p}{2}} \log\left(\frac{2}{\varepsilon_k}\right) + 2\varepsilon_N n^{\frac{1}{2}} \\ &\le C_p \gamma^{\frac{1}{2}} \log \gamma \sum_{i=1}^N k 2^{k(\frac{p}{2}-1)} + 2n^{\frac{1}{2}-\frac{1}{p}}. \end{aligned}$$

If $p = 2$, the geometric sum is bounded by

$$C_p(\gamma^{\frac{1}{2}} \log \gamma) N^2 \le C_p(\gamma^{\frac{1}{2}} \log \gamma) \log^2 n$$

whereas if $p > 2$, it is bounded by

$$C_p(\gamma^{\frac{1}{2}} \log \gamma) n^{\frac{1}{2}-\frac{1}{p}}$$

and our claim follows. □

Remark 3.9: In Section IV, we shall show that this bound is tight for $p > 2$, in the sense that there is a constant $c(p, \gamma) > 0$ such that if $\mathrm{fat}_\varepsilon(F) \ge \gamma\varepsilon^{-p}$, then for every integer n there is some empirical measure μ_n such

$$\ell(F/\mu_n) \ge c(p, \gamma) n^{\frac{1}{2}-\frac{1}{p}}.$$

This result indicates a second phase transition. If

$$c\varepsilon^{-p} \le \mathrm{fat}_\varepsilon(F) \le C\varepsilon^{-p}$$

the growth rate of $\ell(F/\mu_n)$ changes at $p = 2$; if $p < 2$ then $\ell_n = \sup_{\mu_n} \ell(F/\mu_n)$ are uniformly bounded, and if $p > 2$ they increase polynomially.

In the sequel, we will be interested in sample complexity estimates for q-loss classes. Hence, we will be interested to derive a result similar to Theorem 3.8 for classes of the form

$$|F - g|^q = \{|f - g|^q | f \in F\}, \qquad 1 \le q < \infty$$

for any $g \in B(L_\infty(\Omega))$. Note that the proof of Theorem 3.8 was based only on covering number estimates; thus, our first order of business is to establish such bounds on the class $|F - g|^q$.

Lemma 3.10: If $F \subset B(L_\infty(\Omega))$, then for every $1 \le q < \infty$ there is a constant C_q, which depends only on q, such that for every $\varepsilon > 0$, every $g \in B(L_\infty(\Omega))$ and any probability measure μ

$$\log N(\varepsilon, |F - g|^q, L_2(\mu)) \le \log N(C_q\varepsilon, F, L_2(\mu)).$$

In particular, if there is some $\gamma > 1$ and p such that $\mathrm{fat}_\varepsilon(F) \leq \gamma\varepsilon^{-p}$, then

$$\log N(\varepsilon, |F-g|^q, L_2(\mu)) \leq C(p, q, \gamma)\left(\frac{1}{\varepsilon^p}\log^2\frac{2}{\varepsilon}\right).$$

The proof of the first part of the lemma is standard and is omitted. The second one follows from Theorem 3.2.

Corollary 3.11: Assume that F and g are as in Lemma 3.10 and $G = |F-g|^q$. Then, there are constants $C(p, q, \gamma)$ such that for every empirical measure μ_n

$$\ell(G/\mu_n) \leq \begin{cases} C(p, q, \gamma), & \text{if } 0 < p < 2 \\ C(2, q, \gamma)\log^2 n, & \text{if } p = 2 \\ C(p, q, \gamma)n^{\frac{1}{2}-\frac{1}{p}}, & \text{if } p > 2. \end{cases}$$

C. General Covering Estimates

The final direct corollary we derive from Theorem 3.2 is a general estimate on the $L_p(\mu)$ covering numbers of the class F with respect to *any* probability measure μ.

Corollary 3.12: Let F be a GC 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 μ

$$\log D(\varepsilon, F, L_p(\mu)) \leq C_p \mathrm{fat}_{\frac{\varepsilon}{32}}(F)\log^2\left(\frac{2\mathrm{fat}_{\frac{\varepsilon}{32}}(F)}{\varepsilon}\right).$$

Proof: By a standard argument, if F is a GC class then for every $1 \leq p < \infty$, $|F-F|^p$ is also a GC class. Thus, for every $\varepsilon > 0$ there exists some integer n and an empirical measure μ_n such that

$$\sup_{f,g\in F}\left|\|f-g\|^p_{L_p(\mu_n)} - \|f-g\|^p_{L_p(\mu)}\right| \leq 2^p\varepsilon^p.$$

Let $m = N(\varepsilon, F, L_p(\mu_n))$. Therefore, there is a set $\{f_1, \ldots, f_m\} \subset F$ which is a 2ε cover of F in $L_p(\mu_n)$. By the selection of n it follows that this set is a $(2^{p+1}\varepsilon^p)^{1/p}$ cover of F in $L_p(\mu)$. Hence

$$N(4\varepsilon, F, L_p(\mu)) \leq N(\varepsilon, F, L_p(\mu_n)).$$

Our claim follows by Theorem 3.2. □

IV. Averaging Techniques

As stated in the Introduction, our aim is to connect the fat-shattering dimension and the growth rate of the Rademacher averages associated with the class.

The Rademacher averages appear naturally in the analysis of GC classes. Usually, the first step in estimating the deviation of the empirical means from the actual mean is to apply a symmetrization method [7], [23]

$$\Pr\left\{\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f\right| \geq \varepsilon\right\} \leq 4\Pr\left\{\sup_{f\in F}\left|\sum_{i=1}^n r_i f(X_i)\right| \geq \frac{n\varepsilon}{4}\right\} = (*).$$

The path usually taken at this point is to estimate $(*)$ using the covering numbers of F combined with Hoeffding's inequality. Instead, we shall provide direct estimates on the growth rate of the Rademacher averages and combine it with a different concentration inequality.

We start with the definition of the new learning parameter based on the growth rate of the Rademacher averages. Since we want to compare the known results and those obtained here, we establish a lower bound on the fat-shattering dimension in terms of Gaussian averages. This enables us to estimate the fat-shattering dimension in terms of the growth rate of the Rademacher averages. We present several additional applications of this bound. First, we improve the best known estimate on the fat-shattering dimension of the convex hull of a class, at least when $\mathrm{fat}_\varepsilon(F) = O(\varepsilon^{-p})$ for some $0 < p < \infty$. Second, we prove a sharper characterization of GC classes in terms of the empirical ℓ-norms. Finally, we present a partial solution to a problem from the geometry of Banach spaces.

A. Averaging and Fat Shattering

Definition 4.1: Let μ be a probability measure on Ω. Let $R_n = \sup_{\mu_n} R(F/\mu_n)$ and $\overline{R}_{n,\mu} = \mathbb{E}_\mu R(F/\mu_n)$. Thus,

$$\overline{R}_{n,\mu} = \mathbb{E}_\mu\mathbb{E}_\nu \sup_{f\in F} n^{-\frac{1}{2}}\left|\sum_{i=1}^n r_i f(X_i)\right|$$

where r_i are independent Rademacher random variables on (Y, ν) and (X_i) are independent, distributed according to μ. Similarly, it is possible to define ℓ_n and $\overline{\ell}_{n,\mu}$ using Gaussian averages instead of the Rademacher averages.

The connections between R_n and $\overline{R}_{n,\mu}$ are analogous to those between the VC dimension and the VC entropy; R_n is a "worst case" parameter whereas $\overline{R}_{n,\mu}$ is an averaged version, which takes into account the particular measure according to which one is sampling.

The following is a definition of a parameter which may replace the fat-shattering dimension.

Definition 4.2: Let $F \subset B(L_\infty(\Omega))$. For every $\varepsilon > 0$, let

$$\mathrm{rav}_\varepsilon(F) = \sup\{n \in \mathbb{N} | R_n(F) \geq \varepsilon n^{\frac{1}{2}}\}.$$

To see the connection between $\mathrm{fat}_\varepsilon(F)$ and $\mathrm{rav}_\varepsilon(F)$, assume that $\{\omega_1, \ldots, \omega_n\}$ is ε-shattered. Let

$$(\sigma_1, \ldots, \sigma_n) \in \{-1, 1\}^n$$

and set $I = \{\omega_i | \sigma_i = 1\}$. For every $J \subset \{\omega_1, \ldots, \omega_n\}$, let f_J be the function shattering J. Then, by the triangle inequality, and setting $f = f_I$, $f' = f_{I^c}$, it follows that

$$\begin{aligned} &\frac{1}{\sqrt{n}}\sup_{f\in F}\left|\sum_{i=1}^n \sigma_i f(\omega_i)\right| \\ &\quad\geq \frac{1}{2\sqrt{n}}\sup_{f,f'\in F}\left|\sum_{i=1}^n \sigma_i(f(\omega_i) - f'(\omega_i))\right| \\ &\quad\geq \frac{1}{2\sqrt{n}}\left|\sum_{i=1}^n \sigma_i(f_I(\omega_i) - f_{I^c}(\omega_i))\right| \geq \sqrt{n}\varepsilon. \end{aligned}$$

Thus, if $\{\omega_1, \ldots, \omega_n\}$ is ε-shattered, then for every realization of the Rademacher random variables

$$n^{-\frac{1}{2}} \sup_{f\in F}\left|\sum_{i=1}^{n} r_i(t) f(\omega_i)\right| \geq \sqrt{n}\varepsilon$$

while $\mathrm{rav}_\varepsilon(F)$ is determined by averaging such realizations. Hence

$$\mathrm{rav}_\varepsilon(F) \geq \mathrm{fat}_\varepsilon(F).$$

It is considerably more difficult to find an upper bound on $\mathrm{rav}_\varepsilon(F)$ in terms of the fat-shattering dimension. The first step in that direction is to estimate the fat-shattering dimension of the class in terms of the empirical ℓ-norms.

Theorem 4.3: Let $F \subset B(L_\infty(\Omega))$ and let μ_n be an empirical measure. Denote by ℓ the ℓ-norm of F/μ_n. There are absolute constants C, c such that $\rho = c\ell/n^{1/2}$ and

$$\mathrm{fat}_\rho(F) \geq C\left(\frac{\ell}{\log n}\right)^2. \tag{4.1}$$

The idea behind the proof of this result is due to Pajor [17]. Our contribution is the application of the improved bound on the L_2 covering numbers of F, which yields a better bound.

Lemma 4.4: Let $F \subset L_2(\mu_n)$. Then, for every $\varepsilon > 0$

$$D(\varepsilon, F, L_2(\mu_n)) \leq \left(1 + \frac{(2\pi)^{1/2}\ell(F)}{\varepsilon n^{\frac{1}{2}}}\right)^n.$$

Proof: Let $\varepsilon > 0$ and set $K = (F + \frac{\varepsilon}{2}B_2^n)$. Note that if A is an ε-separated subset of F in $L_2(\mu_n)$ and if B_2^n is the unit ball in $L_2(\mu_n)$ then $A + \frac{\varepsilon}{2}B_2^n \subset K$. By comparing volumes

$$\left(\frac{\varepsilon}{2}\right)^n D(\varepsilon, F, L_2(\mu_n)) \leq \frac{\mathrm{vol}(K)}{\mathrm{vol}(B_2^n)}.$$

Set $c_n = \mathbb{E}(\sum_{i=1}^n |g_i|^2)^{1/2}$ and let $d\sigma(t)$ be the Haar measure on S^{n-1}, which is the unit sphere in $\mathbb{R}^n$. Using Uryson's inequality and the standard connections between the Haar measure on the sphere and the Gaussian measure on $\mathbb{R}^n$ (see, e.g., [18])

$$\begin{aligned}\left(\frac{\mathrm{vol}(K)}{\mathrm{vol}(B_2^n)}\right)^{\frac{1}{n}} &\leq \int_{S^{n-1}} \|t\|_{K^\circ}\, d\sigma(t)\\ &= \int_{S^{n-1}} \sup_{f\in F,\, s\in B_2^n} \left\langle f + \frac{\varepsilon}{2}s, t\right\rangle d\sigma(t)\\ &\leq \frac{\varepsilon}{2} + \frac{\ell(F)}{c_n}.\end{aligned}$$

Our claim follows since for every n, $c_n \geq (2/\pi)^{1/2} n^{1/2}$. □

Proof of Theorem 4.3: By the upper bound in Theorem 2.11

$$\ell \leq C_0 \int_0^\infty \log^{\frac{1}{2}} N(\varepsilon, F, L_2(\mu_n))\, d\varepsilon.$$

Applying Lemma 4.4, it follows that for every $0 < x < 1$

$$\begin{aligned}&\int_0^{x\ell/n^{1/2}} \log^{\frac{1}{2}} N(\varepsilon, F, L_2(\mu_n))\, d\varepsilon\\ &\qquad \leq \int_0^{x\ell/n^{1/2}} n^{\frac{1}{2}} \log^{\frac{1}{2}}\left(1 + \frac{\sqrt{2\pi}\ell(F)}{\varepsilon n^{\frac{1}{2}}}\right) d\varepsilon = (*).\end{aligned}$$

Changing the integration variable to $u = \varepsilon n^{1/2}\ell^{-1}$ and by a straightforward estimate of the integral, it follows that

$$(*) = \ell \int_0^x \log^{\frac{1}{2}}\left(1 + \frac{(2\pi)^{1/2}}{u}\right) du \leq 2\ell\left((2\pi)^{\frac{1}{2}}x\right)^{\frac{1}{2}}. \tag{4.2}$$

Let $\alpha = (32\pi)^{-1/2}C_0^{-2}$ and $\beta = \alpha\ell/n^{1/2}$. It is easy to see that $\ell \leq \sqrt{2/\pi}n^{1/2}$, implying that $\beta < 1$. Since $F \subset B(L_\infty(\Omega))$ and by Theorem 3.2 and (4.2)

$$\begin{aligned}&C_0 \int_0^\infty \log^{\frac{1}{2}} N(\varepsilon, F, L_2(\mu_n))\, d\varepsilon\\ &\quad = C_0 \int_0^\beta \log^{\frac{1}{2}} N(\varepsilon, F, L_2(\mu_n))\, d\varepsilon\\ &\qquad + C_0 \int_\beta^1 \log^{\frac{1}{2}} N(\varepsilon, F, L_2(\mu_n))\, d\varepsilon\\ &\quad \leq \frac{\ell}{2} + C\mathrm{fat}^{\frac{1}{2}}_{\frac{\beta}{8}}(F) \log\left(2\frac{\mathrm{fat}_{\frac{\beta}{8}}(F)}{\beta}\right).\end{aligned}$$

Thus,

$$\ell \leq C\mathrm{fat}^{\frac{1}{2}}_{\frac{\beta}{8}}(F) \log\left(\frac{2\mathrm{fat}_{\frac{\beta}{8}}(F)}{\beta}\right)$$

implying that

$$\mathrm{fat}_{\frac{\beta}{8}}(F) \geq C\left(\frac{\ell}{\log n}\right)^2.$$ □

Corollary 4.5: Let $F \subset B(L_\infty(\Omega))$. Then, there are absolute constants c and C such that for every $\varepsilon > 0$

$$\mathrm{rav}_\varepsilon(F) \leq C\frac{\mathrm{fat}_{c\varepsilon}(F)}{\varepsilon^2}\log^2\frac{\mathrm{fat}_{c\varepsilon}(F)}{\varepsilon}.$$

Proof: Assume that μ_n is an empirical measure such that $R(F/\mu_n) \geq \varepsilon n^{1/2}$. Using the connections between Gaussian and Rademacher averages (Theorem 2.10), it follows that there is an absolute constant C such that $\ell(F/\mu_n) \geq C\varepsilon n^{1/2}$. Therefore, by Theorem 4.3

$$\begin{aligned}\mathrm{fat}_{C\varepsilon}(F) &\geq \mathrm{fat}_{C\ell n^{-1/2}}(F) \geq C\left(\frac{\ell}{\log n}\right)^2\\ &\geq C\frac{\varepsilon^2 n}{\log^2 n}.\end{aligned}$$

Thus,

$$n \leq C\frac{\mathrm{fat}_{c\varepsilon}(F)}{\varepsilon^2}\log^2\frac{\mathrm{fat}_{c\varepsilon}(F)}{\varepsilon}$$

implying that $\mathrm{rav}_\varepsilon(F)$ satisfies the same inequality. □

It is interesting to note that if $\mathrm{fat}_\varepsilon(F)$ is polynomial in $1/\varepsilon$ then $\mathrm{rav}_\varepsilon(F)$ and $\mathrm{fat}_\varepsilon(F)$ are equivalent for "large" exponents ($p > 2$), but behave differently for $p \leq 2$. The latter follows since by its definition, $\mathrm{rav}_\varepsilon(F) = \Omega(\varepsilon^{-2})$.

Theorem 4.6: Let $F \subset B(L_\infty(\Omega))$ and assume that $\mathrm{fat}_\varepsilon(F) \leq \gamma\varepsilon^{-p}$ for some $\gamma > 1$ and every $\varepsilon > 0$. Then

$$\mathrm{rav}_\varepsilon(F) \leq C_p\begin{cases}(\gamma\log^2\gamma)\varepsilon^{-2}, & \text{if } 0 < p < 2\\ \gamma\varepsilon^{-2}\log^4\frac{\gamma}{\varepsilon}, & \text{if } p = 2\\ (\gamma^{\frac{p}{2}}\log^p\gamma)\varepsilon^{-p} & \text{if } p > 2.\end{cases}$$

Proof: The proof follows from the ℓ-norm estimates proved in Section III-B. We shall present a complete proof only in the case $p < 2$. By Theorem 3.8, it follows that for every empirical measure μ_n

$$R(F/\mu_n) \leq C\ell(F/\mu_n) \leq C_p\gamma^{1/2}\log\gamma.$$

Hence, if $\mathrm{rav}_\varepsilon(F) \geq n$, there is some empirical measure μ_n such that

$$\varepsilon n^{\frac{1}{2}} \leq R(F/\mu_n) \leq C_p\gamma^{1/2}\log\gamma$$

implying that $n \leq C_p(\gamma\log^2\gamma)\varepsilon^{-2}$ as claimed.

The other proofs follow using a similar argument. □

Using the bounds on $\mathrm{rav}_\varepsilon(F)$ it is possible to bound $\mathrm{fat}_\varepsilon(\mathrm{absconv}(F))$.

Corollary 4.7: Let $F \subset B(L_\infty(\Omega))$. If $\mathrm{fat}_\varepsilon(F) = O(\varepsilon^{-p})$ for some $p \neq 2$ then

$$\mathrm{fat}_\varepsilon(\mathrm{absconv}(F)) = O(\varepsilon^{-\max\{2,p\}}).$$

For $p = 2$, one has an additional logarithmic factor. In general, there are absolute constants c and C such that for every $\varepsilon > 0$

$$\mathrm{fat}_\varepsilon(\mathrm{absconv}(F)) \leq C\,\frac{\mathrm{fat}_{c\varepsilon}(F)}{\varepsilon^2}\log^2\frac{\mathrm{fat}_{c\varepsilon}(F)}{\varepsilon}.$$

Proof: Since the Rademacher averages do not change when one takes the symmetric convex hull, then

$$\mathrm{rav}_\varepsilon(F) = \mathrm{rav}_\varepsilon(\mathrm{absconv}(F)).$$

Hence

$$\mathrm{rav}_\varepsilon(F) = \mathrm{rav}_\varepsilon(\mathrm{absconv}(F)) \geq \mathrm{fat}_\varepsilon(\mathrm{absconv}(F)). \quad (4.3)$$

Now, if $\mathrm{fat}_\varepsilon(F) = O(\varepsilon^{-p})$ for some $p \neq 2$, then by Theorem 4.6

$$\mathrm{rav}_\varepsilon(F) = O\left(\varepsilon^{-\max\{2,p\}}\right).$$

The case $p = 2$ follows from a similar argument, while the general inequality may be derived from Corollary 4.5 and (4.3). □

Remark 4.8: Theorem 4.6 and Corollary 4.7 indicate the same phase transition which occurs when $p = 2$.

Using theorem 4.3 we can prove the following result.

Theorem 4.9: Let $F \subset B(L_\infty(\Omega))$. If F is a GC class then

$$\lim_{n\to\infty} \ell_n/n^{1/2} = 0.$$

Note that in the converse direction, a weaker condition is needed to imply GC. Indeed, it is possible to show that $\sup_\mu \overline{R}_{n,\mu} = o(n^{1/2})$ if and only if F is a GC class [7]. Hence, Theorem 4.3 is a characterization of GC classes.

Proof: If $\ell_n/n^{1/2}$ does not converge to 0, there is a sequence $n_k \to \infty$ and some $\varepsilon > 0$ such that for every n_k, $\ell_{n_k}/n_k^{1/2} \geq \varepsilon$. By Theorem 4.3, there is some constant C such that for every n_k

$$\mathrm{fat}_{C\varepsilon}(F) \geq \mathrm{fat}_{C\ell_{n_k}n_k^{-1/2}}(F) \geq C\left(\frac{\ell_{n_k}}{\log n_k}\right)^2 \geq C\left(\frac{\varepsilon n_k^{1/2}}{\log n_k}\right)^2.$$

Thus, $\mathrm{fat}_{C\varepsilon}(F) = \infty$, and F is not a GC class. □

B. A Geometric Interpretation of the Fat-Shattering Dimension

We begin by exploring the connections between the fat-shattering dimension of F and the fact that F^o contains a copy of ℓ_1^n.

Definition 4.10: Let X be a Banach space and let $(x_i)_{i=1}^n \subset B(X)$. We say that the set $(x_i)_{i=1}^n$ is ε-equivalent to an ℓ_1^n unit-vector basis, if, for every set $(a_i)_{i=1}^n$ of scalars

$$\varepsilon\sum_{i=1}^n |a_i| \leq \left\|\sum_{i=1}^n a_i x_i\right\| \leq \sum_{i=1}^n |a_i|.$$

Clearly, since the vectors (x_i) belong to $B(X)$, the upper bound is always true. Also, note that the set $(x_i)_{i=1}^n \subset B(X)$ is ε-equivalent to an ℓ_1^n unit-vector basis if and only if the operator $T\colon \ell_1^n \to \mathrm{Im}(T) \subset X$ which maps each unit vector v_i to x_i satisfies that $\|T^{-1}\| \leq \varepsilon^{-1}$.

Theorem 4.11: Let

$$F \subset B(L_\infty(\Omega)).$$

If the set $\{\omega_1, \ldots, \omega_n\}$ is ε-shattered by F, then the set $(n^{1/2}e_i)_{i=1}^n \subset (F/\mu_n)^o$ is ε-equivalent to ℓ_1^n unit-vector basis.

Proof: Let $(a_i)_{i=1}^n \subset \mathbb{R}$ and set $A = \{i | a_i \geq 0\}$. Denote by f_A the shattering function of the set A and f_{A^c} is the shattering function of its complement. By the triangle inequality

$$\left\|\sum_{i=1}^n a_i n^{\frac{1}{2}} e_i\right\|_{(F/\mu_n)^o} = \sup_{f\in F}\left|\sum_{i=1}^n a_i f(\omega_i)\right| \geq \tfrac{1}{2}\sup_{f,\,f'\in F}\left|\sum_{i=1}^n a_i(f(\omega_i) - f'(\omega_i))\right|.$$

Selecting $f = f_A$ and $f' = f_{A^c}$ it follows that

$$\begin{aligned}\sup_{f,\,f'\in F}\frac{1}{2}&\left|\sum_{i=1}^n a_i(f(\omega_i) - f'(\omega_i))\right| \\ &\geq \frac{1}{2}\left(\sum_{i\in A} a_i(f_A(\omega_i) - f_{A^c}(\omega_i)) + \sum_{i\in A^c} a_i(f_A(\omega_i) - f_{A^c}(\omega_i))\right) \\ &\geq \varepsilon\sum_{i=1}^n |a_i|.\end{aligned}$$

□

This result has a partial converse, namely, that if $(n^{1/2}e_i)_{i=1}^n$ is ε-equivalent to an ℓ_1^n unit-vector basis, then $\{\omega_1, \ldots, \omega_n\}$ is ε-shattered by the symmetric convex hull of F.

Theorem 4.12: Assume that $F \subset B(L_\infty(\Omega))$ and μ_n is an empirical measure. If $(n^{1/2}e_i)_{i=1}^n \subset (F/\mu_n)^o$ is ε-equivalent to an ℓ_1^n unit-vector basis, then $\{\omega_1, \ldots, \omega_n\}$ is ε-shattered by $\mathrm{absconv}(F)$.

Proof: Let $(v_i)_{i=1}^n$ be the unit vectors in ℓ_1^n. By our assumption, the operator $T\colon \ell_1^n \to (F/\mu_n)^o$ defined by $T(v_i) = n^{1/2}e_i$ is such that $\|T^{-1}\| \leq \varepsilon^{-1}$. Let $I \subset \{1, \ldots, n\}$ and select $v^* \in B(\ell_\infty^n)$ (which is the unit ball of the dual space of ℓ_1^n) such that $v^*(v_i) = 1$ if $i \in I$ and $v^*(v_i) = -1$ otherwise. If (T^{-1^*}) is the dual operator to T^{-1}, and if $u = (T^{-1})^*(v^*)$ then $\varepsilon u \in (F/\mu_n)^{oo} = \mathrm{absconv}(F)$. Also, for every $i \in I$

$$\varepsilon u(\omega_i) = \varepsilon\langle u, n^{\frac{1}{2}}e_i\rangle = \varepsilon\left\langle (T^{-1})^*v, n^{\frac{1}{2}}e_i\right\rangle = \varepsilon v^*(v_i) = \varepsilon$$

and, similarly, if $i \in I^c$ then $\varepsilon u(\omega_i) = -\varepsilon$. Since that set I is an arbitrary subset of $\{\omega_1, \ldots, \omega_n\}$, the set $\{\omega_1, \ldots, \omega_n\}$ is ε-shattered by $\mathrm{absconv}(F)$. □

Using Theorem 4.11 we can show that the bounds obtained for $p > 2$ in Theorem 3.8 are tight.

Corollary 4.13: Let $F \subset B(L_\infty(\Omega))$ and suppose that there is some $\gamma > 1$ such that for every $\varepsilon > 0$, $\mathrm{fat}_\varepsilon(F) \geq \gamma\varepsilon^{-p}$. Then, there is an absolute constant C such that for every integer n and all empirical measures μ_n

$$\ell(F/\mu_n) \geq \left(\frac{2}{\pi}\right)^{\frac{1}{2}} \gamma^{\frac{1}{p}} n^{\frac{1}{2}-\frac{1}{p}}.$$

Proof: Since $\mathrm{fat}_\varepsilon(F) \geq \gamma\varepsilon^{-p}$, then for every integer n there is a set $I \subset \Omega$ such that $|I| \geq n$ and I is $(\gamma/n)^{1/p}$ shattered by F. Let μ_n be the empirical measure supported on the first n elements of I. By Theorem 4.11, the set

$$(n^{1/2}e_i)_{i=1}^n \subset (F/\mu_n)^o$$

is $(\gamma/n)^{1/p}$-equivalent to an ℓ_1^n unit-vector basis. Therefore,

$$\begin{aligned}\ell(F/\mu_n) &= n^{-\frac{1}{2}}\mathbb{E}\left\|\sum_{i=1}^n g_i n^{\frac{1}{2}} e_i\right\|_{(F/\mu_n)^o} \\ &\geq \gamma^{\frac{1}{p}} n^{-(\frac{1}{2}+\frac{1}{p})} \sum_{i=1}^n \mathbb{E}|g_i| \\ &= (\mathbb{E}|g_1|)\gamma^{\frac{1}{p}} n^{\frac{1}{2}-\frac{1}{p}}\end{aligned}$$

as claimed. □

C. The Elton–Pajor Theorem

Theorem 4.3 has an application in the theory of Banach spaces. The question at hand is as follows: consider a set $A = \{x_1, \ldots, x_n\}$ of vectors in some Banach space X. Let

$$R = \mathbb{E}\left\|\sum_{i=1}^n r_i x_i\right\|_X \quad \text{and} \quad E = \mathbb{E}\left\|\sum_{i=1}^n g_i x_i\right\|_X.$$

If R or E are large, is there a large subset of A which is "almost" equivalent to an ℓ_1^m unit-vector basis (see Definition 4.10)?

This question was first tackled by Elton [8] who showed that if $R \geq \varepsilon n$, there is a set $I \subset A$, such that $|I| \geq c(\varepsilon)n$ which is $K(\varepsilon)$ equivalent to an $\ell_1^{|I|}$ unit-vector basis, where $K \to 1$ and $c \to 1/2$ as $\varepsilon \to 1$. This result was improved by Pajor [17] who showed that it is possible to select $c(\varepsilon) = C\varepsilon^2$ and $K(\varepsilon) = C\varepsilon^3$ for some absolute constant C. Talagrand [20] was able to show the following result.

Theorem 4.14: There is some absolute constant K such that for every set $A = \{x_1, \ldots, x_n\} \subset B(X)$, there is a subset I, such that $|I| \geq E^2/Kn$ which is

$$\left(\frac{Kn}{E^2}|I|\right)^{-\frac{1}{2}} \left(\log\left(\frac{Kn}{E^2}|I|\right)\right)^{-K}$$

equivalent to an $\ell_1^{|I|}$ unit-vector basis.

We can derive a similar result using Theorem 4.3:

Theorem 4.15: Let $A = \{x_1, \ldots, x_n\} \subset B(X)$, and set

$$E = \mathbb{E}\left\|\sum_{i=1}^n g_i x_i\right\|_X.$$

Then, there is a subset $I \subset A$, such that

$$|I| \geq C\frac{E^2}{n\log^2 n}$$

which is CE/n equivalent to an ℓ_1^m unit-vector basis, where C is an absolute constant.

Proof: Let $\Omega = \{x_1, \ldots, x_n\}$ and set $F = B(X^*)$, implying that $F \subset B(L_\infty(\Omega))$. Moreover, if $I \subset \Omega$, $|I| = m$, and μ_m is the empirical measure supported on I, then for every set of scalars $(a_i)_{i=1}^m$

$$\begin{aligned}\left\|\sum_{i=1}^m a_i x_i\right\|_X &= \sup_{x^* \in B(X^*)} \sum_{i=1}^m a_i x^*(x_i) \\ &= \sup_{x^* \in B(X^*)} \left\langle m^{-\frac{1}{2}} \sum_{i=1}^m x^*(x_i)e_i, \sum_{i=1}^m a_i m^{\frac{1}{2}} e_i \right\rangle \\ &= \left\|\sum_{i=1}^m a_i m^{\frac{1}{2}} e_i\right\|_{(F/\mu_m)^o}.\end{aligned}$$

Thus, the set $(m^{1/2}e_i)_{i=1}^m \subset (F/\mu_m)^o$ is ε-equivalent to an ℓ_1^m unit-vector basis, if and only if $I \subset X$ is also ε-equivalent.

If μ_n is the empirical measure supported on the set $\{x_1, \ldots, x_n\}$, then

$$\ell(F/\mu_n) = n^{-\frac{1}{2}}E.$$

By Theorem 4.3

$$\mathrm{fat}_{CE/n}(F) \geq C\frac{E^2}{n\log^2 n}$$

hence, by Theorem 4.11, there is a subset $B \subset \{x_1, \ldots, x_n\}$ such that

$$m = |B| \geq C\frac{E^2}{n\log^2 n}$$

and B is CE/n-shattered by F. Therefore, if μ_m is the empirical measure supported on B, then $\{m^{1/2}e_i\}_{i=1}^m \subset (F/\mu_m)^o$ is CE/n equivalent to an ℓ_1^m unit-vector basis, and our claim follows. □

Now, we can derive a similar result to that of Pajor:

Corollary 4.16: Let $A = \{x_1, \ldots, x_n\} \subset B(X)$. If

$$\mathbb{E}\left\|\sum_{i=1}^n r_i x_i\right\|_X \geq \varepsilon n$$

then there is a subset $I \subset A$, such that

$$|I| \geq C\frac{n\varepsilon^2}{\log^2 n}$$

which is $C\varepsilon$-equivalent to an $\ell_1^{|I|}$ unit-vector basis for some absolute constant C.

V. Complexity Estimates

In this section, we prove sample complexity estimates for an agnostic learning problem with respect to any q-loss function. We use a concentration result which yields an estimate on the deviation of the empirical means from the actual mean in terms of the Rademacher averages. We then apply the estimates on those averages in terms of the fat-shattering dimension obtained is Section III-B and improve the known complexity estimates. It turns out that $\mathrm{rav}_\varepsilon(F)$ measures precisely the sample complexity.

We begin with the following result which is due to Talagrand [21].

Theorem 5.1: There are two absolute constants K and $a \leq 1$ with the following property: consider a class of functions F whose range is a subset of $[0, 1]$, such that

$$\sup_{f \in F} \mathbb{E}(f - \mathbb{E}f)^2 \leq a.$$

If μ is any probability measure on Ω and

$$n^{\frac{1}{2}} \geq K\overline{R}_{n,\mu} \qquad M \geq K\overline{R}_{n,\mu}$$

then

$$\Pr\left\{\sup_{f \in F} |\mathbb{E}_{\mu_n} f - \mathbb{E}_\mu f| \geq M n^{-\frac{1}{2}}\right\} \leq K \exp(-11M^2).$$

Let G be a class of functions whose range is a subset of $[0, 1]$, let g be some function whose range is a subset of $[0, 1]$ and fix some q, such that $1 \leq q < \infty$. If $F = |G - g|^q$ then F is also a class of function whose range is a subset of $[0, 1]$. Let a be as in Theorem 5.1 and denote by $F^a = \{\sqrt{a}f | f \in F\}$. Therefore, $\sup_{f \in F^a} \mathbb{E}(f - \mathbb{E}f)^2 \leq a$.

Lemma 5.2: Let F and F^a be as in the above paragraph. If $\varepsilon > 0$ and n are such that

$$n^{\frac{1}{2}} \geq Ka^{-\frac{1}{2}}\varepsilon^{-1}\overline{R}_{n,\mu} \tag{5.1}$$

then

$$\Pr\left\{\sup_{f \in F} |\mathbb{E}_{\mu_n} f - \mathbb{E}_\mu f| \geq \varepsilon\right\} \leq K \exp(-11an\varepsilon^2).$$

Proof: Clearly,

$$\Pr\left\{\sup_{f \in F} |\mathbb{E}_{\mu_n} f - \mathbb{E}_\mu f| \geq \varepsilon\right\} = \Pr\left\{\sup_{f \in F^a} |\mathbb{E}_{\mu_n} f - \mathbb{E}_\mu f| \geq \sqrt{a}\varepsilon\right\}.$$

Let $M = a^{1/2}n^{1/2}\varepsilon$. Since $a, \varepsilon \leq 1$ then if n satisfies (5.1), both conditions of Theorem 5.1 are automatically satisfied. The assertion follows directly from that theorem. □

We can apply Lemma 5.2 and obtain the desired sample complexity estimate. We first present a general complexity estimate in terms of the parameter $\mathrm{rav}_\varepsilon(F)$.

Theorem 5.3: Let F be a class of functions into $[0, 1]$. Then, there is an absolute constant C such that for every $0 < \varepsilon, \delta < 1$, and every probability measure μ

$$\Pr\left\{\sup_{f \in F} |\mathbb{E}_{\mu_n} f - \mathbb{E}_\mu f| \geq \varepsilon\right\} \leq \delta$$

provided that

$$n \geq C \max\left\{\mathrm{rav}_{C\varepsilon}(F), \frac{1}{\varepsilon^2}\log\frac{1}{\delta}\right\}.$$

Proof: Let $C_1 = 2Ka^{-1}$ and M as in Lemma 5.2. Note that in order to ensure that $n^{1/2} \geq 2Ka^{-1}\varepsilon^{-1}\overline{R}_{n,\mu}$, it is enough that $R_n \leq C\varepsilon n^{1/2}$. This will hold if $n \geq \mathrm{rav}_{C\varepsilon}(F)$. Thus, by Lemma 5.2

$$\Pr\left\{\sup_{f \in F} |\mathbb{E}_{\mu_n} f - \mathbb{E}_\mu f| \geq \varepsilon\right\} \leq K \exp(-11M^2) < \delta$$

where the last inequality is valid provided that $n \geq C\varepsilon^{-2}\log\frac{1}{\delta}$. □

Corollary 5.4: Let G be a class of functions whose range is contained in $[0, 1]$, such that $\mathrm{fat}_\varepsilon(G) \leq \gamma\varepsilon^{-p}$ for some $p > 0$. Then, for every $1 \leq q < \infty$ there are constants $C(p, q, \gamma)$ such that for any g: $\Omega \to [0, 1]$

$$S_q(\varepsilon, \delta, g) \leq C(p, q, \gamma)\begin{cases} \frac{1}{\varepsilon^2}\log\frac{1}{\delta}, & \text{if } 0<p<2 \\ \frac{1}{\varepsilon^2}\left(\log^4\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right), & \text{if } p=2 \\ \frac{1}{\varepsilon^p}\log\frac{1}{\delta}, & \text{if } p>2. \end{cases}$$

Proof: We shall prove the claim for $p > 2$. The assertion in the other two cases follows in a similar fashion.

Let $F = |G - g|^q$. By Theorem 3.11, there are constants $C = C(p, q, \gamma)$ such that for every integer n and every empirical measure μ_n, $R_n \leq Cn^{1/2-1/p}$. Hence, for every $\varepsilon > 0$, $\mathrm{rav}_\varepsilon(F) \leq C\varepsilon^{-p}$. Our result follows from Theorem 5.3. □

Remark 5.5: Using a simple scaling argument, if $\sup_{f \in F} |f| \leq M$ then sample complexity will be bounded by

$$C \max\left\{\mathrm{rav}_{C\varepsilon/M}, \frac{M^2}{\varepsilon^2}\log\frac{1}{\delta}\right\}.$$

Corollary 5.6: Let $F \subset B(L_\infty(\Omega))$, such that $\mathrm{fat}_\varepsilon(F) \leq \gamma\varepsilon^{-p}$ for some $p > 0$. Then, for every $1 \leq q < \infty$ and every $M > 0$ there are constants $C(p, q, M, \gamma)$, such that for every $0 < \varepsilon, \delta < 1$

$$\sup_{\|g\|_\infty \leq M} S_q(\varepsilon, \delta, g) \leq C(p, q, M, \gamma)\begin{cases} \frac{1}{\varepsilon^2}\log\frac{1}{\delta}, & \text{if } 0 < p < 2 \\ \frac{1}{\varepsilon^2}\left(\log^4\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right), & \text{if } p = 2 \\ \frac{1}{\varepsilon^p}\log\frac{1}{\delta}, & \text{if } p > 2. \end{cases}$$

A. GC Complexity Versus Learning Complexity

The term "sample complexity" is often used in a slightly different way than the one we use here. Normally, when one talks about the sample complexity of a learning problem, the meaning is the following, more general setup. For every $1 \leq q < \infty$, let $\ell_q^f(x, y) = |f(x) - y|^q$. Let $\mathcal{Y}$ be a bounded subset of R. A *learning rule* is a mapping which assigns to each sample of arbitrary length $z_n = (x_i, y_i)_{i=1}^n$, some $f \in F$. For every class F and $\mathcal{Y} \subset \mathbb{R}$, let the *learning sample complexity* be the smallest integer n_0 such that for every $n \geq n_0$ the following holds: there exists a learning rule A such that for every probability measure P on $\Omega \times \mathcal{Y}$

$$\Pr\left\{\mathbb{E}|A_{Z_n} - Y|^q \geq \inf_{f \in F} \mathbb{E}\ell_q^f(X, Y) + \varepsilon\right\} < \delta$$

where Z_n are n independent samples of (X, Y), sampled according to P. We denote the learning sample complexity associated with the range $\mathcal{Y}$ and the class F by $C_q(\varepsilon, \delta, \mathcal{Y}, F)$.

It is possible to show that if $\mathcal{Y} \subset [-M, M]$ then

$$C_2(\varepsilon, \delta, \mathcal{Y}, F) \leq \sup_{\|g\|_\infty \leq M} S_2(\varepsilon, \delta, g, F).$$

Note that S_q is monotone with respect to inclusion: if $F \subset G$ then for every ε, δ, q, and g

$$S_q(\varepsilon, \delta, g, F) \leq S_q(\varepsilon, \delta, g, G).$$

On the other hand, the same does not hold for $\mathcal{C}_q$, since learning rules may use particular geometric features of the class F. For example, improved learning complexity estimates for convex classes are indicated in the next result.

Theorem 5.7: Let F be a convex class of functions into $[0, 1]$.

1) For every $M > 0$ there is a constant $c(M)$ such that for every $0 < \varepsilon, \delta < 1$ and every $\mathcal{Y} \subset [-M, M]$

$$\mathcal{C}_2(\varepsilon, \delta, \mathcal{Y}, F) \leq \frac{c(M)}{\varepsilon}\left(\mathrm{fat}_{c\varepsilon}(F) + \log\frac{1}{\delta}\right)$$

where c is some absolute constant.

2) If there is a constant C such that $\mathrm{fat}_\varepsilon(F) \leq C\varepsilon^{-p}$ for some $0 < p < 2$, then for every $M > 0$ there is a constant $c(M)$ such that for every $0 < \varepsilon, \delta < 1$, and every $\mathcal{Y} \subset [-M, M]$

$$\mathcal{C}_2(\varepsilon, \delta, \mathcal{Y}, F) \leq \frac{c(M)}{\varepsilon^{1+\frac{p}{2}}}\left(\log^2\frac{2}{\varepsilon} + \log\frac{2}{\delta}\right).$$

The first part of the claim is due to Lee, Bartlett, and Williamson [11], [12], while the second is presented in [15].

It is worthwhile to compare the estimates obtained in Corollary 5.4 with previous GC sample complexity estimates. The following result is due to Bartlett and Long [4].

Theorem 5.8: Let F be a class of functions into $[0, 1]$. Assume that for every $\varepsilon > 0$, $\mathrm{fat}_\varepsilon(F) < \infty$. Then, there is some $0 < \tau < 1/4$ such that for every $M > 0$ and every $0 < \varepsilon, \delta < 1$

$$\sup_{\|g\|_\infty \leq M} S_1(\varepsilon, \delta, g, F) \leq c(M)\left(\frac{1}{\varepsilon^2}\left(d\log^2\frac{2}{\varepsilon} + \log\frac{2}{\delta}\right)\right)$$

where $d = \mathrm{fat}_{(\frac{1}{4}-\tau)\varepsilon}(F)$ and $c(M)$ is a constant which depends only on M.

Thus, if $\mathrm{fat}_\varepsilon(F) = O(\varepsilon^{-p})$ then the bound which follows from Corollary 5.4 is considerably better. The general bound, obtained by combining Corollary 4.5 and Theorem 5.3 is essentially the same as in Theorem 5.8.

On the polynomial scale, the GC sample complexity results we obtain are optimal (with respect to rates), at least for quadratic loss and $p > 2$. Indeed, note that the GC complexity estimates remain true even if $|F - g|^q$ is not bounded by 1, but rather by some other constant. Hence, the asymptotics of these estimates hold even if g is not into $[0, 1]$. It is possible to show [1] that if the range of g exceeds $[0, 1]$ (for example, may be taken to be $[-1, 2]$), then the learning sample complexity is $\Omega(\mathrm{fat}_{4\varepsilon}(F))$. Therefore, the bound found in Corollary 5.4 is optimal.

VI. Conclusion

The common view is that the fat-shattering dimension is the "correct" way of measuring certain properties of the given class, mainly its covering numbers in empirical L_p spaces. Theorem 3.2 seems to strengthen this opinion. However, when it comes to complexity estimates and other geometric properties, the fact that the covering numbers change "smoothly" with the fat-shattering dimension hides a phase transition which occurs on the polynomial scale when $\mathrm{fat}_\varepsilon(F) = O(\varepsilon^{-p})$ at $p = 2$. This phase transition is evident, for example, when considering the Rademacher averages (resp., ℓ-norms). Indeed, when $p < 2$, the averages are uniformly bounded, and when $p > 2$, they are polynomial in n. As for GC complexity estimates, the smooth change which appears in Theorem 5.8 is due to a loose upper bound. The optimal result which we were able to obtain reveals the phase transition: if $p < 2$, the estimate is $O(\varepsilon^{-2})$, and for $p > 2$, it is $O(\varepsilon^{-p})$.

These facts seem to indicate that the "correct" parameter which measures the GC sample complexity is $\mathrm{rav}_\varepsilon(F)$ and not the fat-shattering dimension.

Other advantages in using $\mathrm{rav}_\varepsilon(F)$ are the following: first, the Rademacher and Gaussian averages remain unchanged when passing to the convex hull of the class. This may be exploited because it implies that in many cases one may solve the learning problem within the convex hull of the original class rather than in the class itself, without having to pay a significant price. Second, in many cases one may compute $\sup_{f\in F}|\sum_{i=1}^n r_i f(\omega_i)|$ for a realization of (r_i). Since this random variable is concentrated near its mean, it is possible to estimate Rademacher averages by sampling. Finally, and in our opinion most importantly, Rademacher and Gaussian averages are closely linked to the geometric structure of the class. They can be used to estimate not only covering numbers but approximation numbers as well (see, for example, [16]), which serves as a good indication of the size of the class and may be used to formulate alternative learning procedures.

References

[1] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[2] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, "Scale sensitive dimensions, uniform convergence and learnability," *J. Assoc. Comput. Mach.*, vol. 44, no. 4, pp. 615–631, 1997.

[3] P. L. Bartlett, S. R. Kulkarni, and S. E. Posner, "Covering numbers for real valued function classes," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1721–1724, Sept. 1997.

[4] P. L. Bartlett and P. Long, "More theorems about scale sensitive dimensions and learning," in *Proc. 8th Annu. Conf. Computational Learning Theory*, pp. 392–401.

[5] R. M. Dudley, "The sizes of compact subsets of Hilbert space and continuity of Gaussian processes," *J. Functional Anal.*, vol. 1, pp. 290–330, 1967.

[6] ——, "Uniform central limit theorems," in *Cambridge Studies in Advanced Mathematics 63*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[7] R. M. Dudley, E. Giné, and J. Zinn, "Uniform and universal Glivenko–Cantelli classes," *J. Theor. Probab.*, vol. 4, pp. 485–510, 1991.

[8] J. Elton, "Sign embedding of ℓ_1^n," *Trans. Amer. Math. Soc.*, vol. 279, no. 1, pp. 113–124, 1983.

[9] E. Giné, "Empirical processes and applications: An overview," *Bernoulli*, vol. 2, no. 1, pp. 1–38, 1996.

[10] E. Giné and J. Zinn, "Some limit theorems for empirical processes," *Ann. Probab.*, vol. 12, no. 4, pp. 929–989, 1984.

[11] W. S. Lee, "Agnostic learning and single hidden layer neural networks," Ph.D. dissertation, Australian Nat. Univ., Canberra, ACT, 1996.

[12] W. S. Lee, P. L. Bartlett, and R. C. Williamson, "The importance of convexity in learning with squared loss," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1974–1980, Sept. 1998.

[13] J. Lindenstrauss and V. D. Milman, "The local theory of normed spaces and its application to convexity," in *Handbook of Convex Geometry*. Amsterdam, The Netherlands: Elsevier, 1993, pp. 1149–1120.
[14] S. Mendelson, "Geometric methods in the analysis of Glivenko–Cantelli classes," in *Proc. 14th Ann. Conf. Computational Learning Theory*, 2001, pp. 256–272.
[15] ——, "Improving the sample complexity using global data," preprint.
[16] ——, "On the geometry of Glivenko–Cantelli classes," preprint.
[17] A. Pajor, *Sous Espaces ℓ_1^n des Espaces de Banach*. Paris, France: Hermann, 1985.
[18] G. Pisier, *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge, U.K.: Cambridge Univ. Press, 1989.
[19] V. N. Sudakov, "Gaussian processes and measures of solid angles in Hilbert space," *Sov. Math.–Dokl.*, vol. 12, pp. 412–415, 1971.
[20] M. Talagrand, "Type, infratype and the Elton–Pajor theorem," *Inv. Math.*, vol. 107, pp. 41–59, 1992.
[21] ——, "Sharper bounds for Gaussian and empirical processes," *Ann. Probab.*, vol. 22, no. 1, pp. 28–76, 1994.
[22] N. Tomczak-Jaegermann, "Banach–Mazur distance and finite-dimensional operator Ideals," in *Pitman Monographs and Surveys in Pure and Applied Mathematics*. New York: Pitman, 1989, vol. 38.
[23] V. Vapnik and A. Chervonenkis, "Necessary and sufficient conditions for uniform convergence of means to mathematical expectations," *Theory Prob. its Applic.*, vol. 26, no. 3, pp. 532–553, 1971.