Geometric Methods in the Analysis of Glivenko-Cantelli Classes

Shahar Mendelson

Computer Sciences Laboratory, RSISE, The Australian National University, Canberra 0200, Australia
shahar@csl.anu.edu.au

Abstract. We use geometric methods to investigate several fundamental problems in machine learning. We present a new bound on the L_p covering numbers of Glivenko-Cantelli classes for $1 \leq p < \infty$ in terms of the fat-shattering dimension of the class, which does not depend on the size of the sample. Using the new bound, we improve the known sample complexity estimates and bound the size of the Sufficient Statistics needed for Glivenko-Cantelli classes.

1 Introduction

Estimating the covering numbers of a class of functions has always been important in the context of Learning Theory. This is due to the fact that almost all the sample complexity estimates and the characterization of Glivenko-Cantelli classes are based on the growth rate of the covering numbers as a function of the size of the sample. In fact, the strength of the combinatorial learning parameters (the VC dimension and the fat-shattering dimension) is that they enable one to bound the covering numbers of the class ([3, 17]), usually with respect to the empirical L_∞ norms.

Let n be an integer and set S_n to be a sample which consists of at most n points. Sauer's lemma for VC classes and its real valued counterpart for classes with a finite fat-shattering dimension [3] provide a bound on the covering numbers of the set $F/S_n = \left\{ \left(f(\omega_1), ..., f(\omega_n)\right) \middle| f \in F, \omega_i \in S_n \right\}$ with respect to the L_∞ norm. These bounds imply that as the size of the sample is increased, the L_∞ covering numbers increase by a logarithmic factor in n.

For VC classes, it is possible to estimate the covering numbers with respect to other L_p norms and probability measures which are not necessarily supported on a finite set [18]. In particular, for $1 \leq p < \infty$ the covering numbers of F/S_n are uniformly bounded as a function of n. In the real valued case some progress was made, but only for empirical L_1 norms; in [4] it was shown that if F has a finite fat-shattering dimension then for every empirical measure μ_n its $L_1(\mu_n)$ covering numbers are linear in the fat-shattering dimension (up to a logarithmic factor). In this article we prove similar bounds for $1 < p < \infty$ and present several applications of the estimates. The most important application is an estimate on the so-called ℓ-norm of the sets F/S_n (defined below) when viewed as subsets

of L_2. The first application of the ℓ-norm estimates is a significant improvement in the known sample complexity estimates for Glivenko-Cantelli classes. The second application is a new bound on size of the *Sufficient Statistics* of Glivenko-Cantelli classes, which is the number of linear combinations of point evaluation functionals needed to identify a member of the given class.

We end this introduction with some definitions, notation and basic results we require in the sequel.

Recall that a class of functions is a Glivenko-Cantelli (GC) class if it satisfies the *uniform law of large numbers*. Formally, if F is a class of functions on a measurable set (Ω, Σ), it is a GC class if for every $\varepsilon > 0$,

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

where the supremum is taken with respect to all probability measures μ, $(X_i)_{i=1}^{\infty}$ are independently sampled according to μ and $\mathbb{E}_\mu$ is the expectation with respect to μ.

We investigate the *GC sample complexity* of the class F which is a quantified version of (1); F is a GC class if and only if for every $\varepsilon > 0$ and $0 < \delta < 1$, there exists some integer n_0, such that for every probability measure μ and every $n \geq n_0$,

$$\mu\left\{ \sup_{f\in F} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}_\mu f \right| \geq \varepsilon \right\} \leq \delta \ . \tag{2}$$

For every $\varepsilon > 0$ and $0 \leq \delta < 1$, the smallest possible integer n_0 such that (2) is satisfied is called the Glivenko-Cantelli sample complexity associated with the pair ε, δ. If $\mathcal{H}$ is a class of functions on Ω, $g : \Omega \to [0,1]$ and $1 \leq q < \infty$, let $\{|h-g|^q \,|h \in \mathcal{H}\}$ be the q-loss class given by $\mathcal{H}$, g and q. The sample complexity the q-loss class is denoted by $S_q(\varepsilon, \delta, g, \mathcal{H})$.

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.

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 for every $\varepsilon > 0$, $N(\varepsilon, F) \leq D(\varepsilon, F) \leq N(\varepsilon/2, F)$.

Given a Banach space X, denote its unit ball by $B(X)$. For any probability measure μ on the measurable space (Ω, Σ), let $\mathbb{E}_\mu$ be 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, i.e., for every function f on Ω, $\delta_\omega(f) = f(\omega)$. We denote by μ_n an empirical measure supported on a set of n points, hence, $\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\omega_i}$. 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. Finally, 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.

The first combinatorial parameter used in learning theory was introduced by Vapnik and Chervonenkis for classes of $\{0,1\}$-valued functions [17]. Later, this parameter was generalized in various fashions. The parameter which we focus on is the *fat shattering dimension*.

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

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

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

The following result, due to Alon, Ben-David, Cesa-Bianchi and Haussler [3], enables one to estimate the $L_\infty(\mu_n)$ covering numbers of classes in terms of the fat shattering dimension.

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

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

In particular, the same estimate holds in $L_p(\mu_n)$ for any $1 \leq p < \infty$.

Although $\log D\big(\varepsilon, F, L_p(\mu_n)\big)$ is "almost linear" in $\mathrm{fat}_\varepsilon(F)$, this estimate is not dimension free, since it carries a factor of $\log^2 n$.

Let ℓ_2^n be a real n-dimensional inner product space, and denote the inner product by $\langle\ ,\ \rangle$. Given a set F, the symmetric convex hull of F is $\mathrm{absconv}(F) = \{\sum_{i=1}^n a_i f_i | n \in \mathbb{N}, f_i \in F, \sum_{i=1}^n |a_i| = 1\}$. The convex hull of F is $\mathrm{conv}(F) = \{\sum_{i=1}^n a_i f_i | n \in \mathbb{N}, f_i \in F,\ a_i \geq 0,\ \sum_{i=1}^n a_i = 1\}$.

If F is a class and μ_n is an empirical measure, we endow $\mathbb{R}^n$ with the Euclidean structure of $L_2(\mu_n)$, which is isometric to ℓ_2^n. If μ_n is the empirical measure supported on $\{\omega_1, ..., \omega_n\}$, let $F/\mu_n = \Big\{ \sum_{i=1}^n f(\omega_i)\chi_{\{\omega_i\}} | f \in F \Big\} \subset L_2(\mu_n)$.

Throughout this article we make extensive use of probabilistic averaging techniques. To that end, we define the Gaussian averages of F/μ_n.

Definition 2. *Let F be a class of functions on Ω. Let $\{\omega_1, ..., \omega_n\} \subset \Omega$ be a fixed sample and let μ_n be the empirical measure supported on the sample. Set*

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

where $(g_i)_{i=1}^n$ are independent standard Gaussian random variables.

First, note that $\ell\big(\mathrm{absconv}(F)/\mu_n\big) = \ell(F/\mu_n)$. Much less obvious and considerably more important is the following result, which provides an upper bound on $\ell(F/\mu_n)$ in terms of the covering numbers of F in $L_2(\mu_n)$. This bound was demonstrated by Dudley [6]

Theorem 2. *There is an absolute constant C such that for every integer n and every $F \subset L_2(\mu_n)$,*

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

In a similar fashion, it is possible to define the Rademacher averages associated with a class F and a sample $\{\omega_1, ..., \omega_n\}$.

Definition 3. *Let F be a class of functions on Ω. Let $\{\omega_1, ..., \omega_n\} \subset \Omega$ be a fixed sample and let μ_n be the empirical measure supported on the sample. Set*

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

where $(\varepsilon_i)_{i=1}^n$ are independent Rademacher (i.e. symmetric $\{-1,1\}$-valued) random variables.

Remark 1. It is possible to show [16] that there is an absolute constant C such that for every class F, every integer n and every sample $\{\omega_1, ..., \omega_n\}$, $R(F/\mu_n) \leq C\ell(F/\mu_n)$.

2 The Covering Theorem and its Applications

The main result presented in this section is an estimate on the covering numbers of a class, when considered as a subset of $L_p(\mu_n)$, for an empirical measure μ_n. The estimate is expressed in terms of the fat shattering dimension of the class. Controlling the covering numbers is important in this case, since this quantity appears in most deviation results used for sample complexity estimates. We present a bound which is dimension free (i.e., does not depend on the cardinality of the set on which the measure is supported) and is "almost" linear in $fat_\varepsilon(F)$. Thus far, the only way to obtain such a result in every L_p space was through the L_∞ estimates (Theorem 1). Unfortunately, those estimates carry a factor of $\log^2 |I|$, where I is the set on which the measure is supported. Hence, the estimate is not dimension free.

The proof is based on a idea which is due to Pajor [12] and is divided to two steps. We begin by showing that if μ_n is supported on $\{\omega_1, ..., \omega_n\}$ and if a set $F \subset B\big(L_\infty(\Omega)\big)$ is "well separated" in $L_2(\mu_n)$, there is a "small" subset $I \subset \{\omega_1, ..., \omega_n\}$ 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 1. *Let $F \subset B\big(L_\infty(\Omega)\big)$ and suppose that μ_n is the empirical measure supported on $A = \{\omega_1, ..., \omega_n\}$. Fix $\varepsilon > 0$ and $p \geq 1$, set $d_p = D\big(\varepsilon, F, L_p(\mu_n)\big)$ and assume that $d_p > 1$. Then, there is a constant c_p, which depends only on p, and a subset $I \subset A$, such that $|I| \leq c_p \varepsilon^{-p} \log d_p$, and $\log D\big(\varepsilon, F, L_p(\mu_n)\big) \leq \log D\big(\varepsilon/2, F, L_\infty(I)\big)$.*

Proof. Fix any integer n and $p \geq 1$ and let $\{f_1, ..., f_{d_p}\} \subset F$ be ε-separated in $L_p(\mu_n)$. Hence, for every $i \neq j$, $\varepsilon^p < n^{-1} \sum_{k=1}^n |f_i(\omega_k) - f_j(\omega_k)|^p$. Let $L(i,j)$ be the set of indices on which $|f_i(\omega_k) - f_j(\omega_k)| \leq \varepsilon/2$. Note that for every $i \neq j$,

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

Thus,

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

Let $(X_i)_{1 \leq i \leq t}$ be t independent random variables, uniformly distributed on $\{1, ..., n\}$. 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 $\Big(1 - \Big(\frac{2^p-1}{4^p}\Big)\varepsilon^p\Big)^t$. 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

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

If $\Theta < 1$, there is a set $I \subset \{\omega_1, ..., \omega_n\}$ 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 $t \geq c_p \varepsilon^{-p} \log d_p$ where c_p is a constant which depends only on p.

Theorem 3. *If $F \subset B\big(L_\infty(\Omega)\big)$, 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 $\varepsilon > 0$,*

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

Proof. Fix $\varepsilon > 0$. By Lemma 1 and Theorem 1 there is a subset $I \subset \{\omega_1, ..., \omega_n\}$ such that

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

and

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

Therefore,

$$\log D\big(\varepsilon, F, L_p(\mu_n)\big) \leq c_p \mathrm{f}at_{\frac{\varepsilon}{8}}(F) \log^2\Big(\frac{2\mathrm{f}at_{\frac{\varepsilon}{8}}(F)}{\varepsilon}\Big) \ .$$

as claimed.

Note that the estimate we presented is optimal up to a logarithmic factor. This follows from the next general lower bound (see [2] for further details).

Theorem 4. *Let F be a class of functions. Then, for any $\varepsilon > 0$, every integer $n \geq \mathrm{f}at_{16\varepsilon}(F)$ and any $1 \leq p < \infty$,*

$$\sup_{\mu_n} N\big(\varepsilon, F, L_p(\mu_n)\big) \geq e^{\mathrm{f}at_{16\varepsilon}(F)/8} \ .$$

Corollary 1. *Let $F \subset B\big(L_\infty(\Omega)\big)$. Then, for every $\varepsilon > 0$ and $1 \leq p < \infty$,*

$$\frac{\mathrm{f}at_{16\varepsilon}(F)}{8} \leq \sup_n \sup_{\mu_n} \log N\big(\varepsilon, F, L_p(\mu_n)\big) \leq c_p \mathrm{f}at_{\frac{\varepsilon}{4}}(F) \log^2\Big(\frac{2\mathrm{f}at_{\frac{\varepsilon}{4}}(F)}{\varepsilon}\Big) \ ,$$

where c_p is a constant which depends only on p.

The behaviour of the supremum of the log-covering numbers plays an important part in the analysis of empirical processes. For example, Pollard's entropy condition, which is a condition sufficient to ensure that a class satisfies the Universal Central Limit Theorem, is given in terms of the integral of this supremum [18].

3 ℓ-Norm Estimates

The ℓ-norm of subsets of $L_2(\mu_n)$ spaces play an important part in our discussion. We will show that they may be used to estimate the number of Sufficient Statistics needed for a given class. Moreover, bounding the ℓ-norm yields a bound on the Rademacher averages associated with the class, which plays a major role in the analysis of the sample complexity of Glivenko-Cantelli classes.

We begin with the following lemma, which is based on the proof of the upper bound in Theorem 2 (see [14]).

Lemma 2. *Let μ_n be an empirical measure on Ω, put $F \subset B\big(L_\infty(\Omega)\big)$ 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,*

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

In particular,

$$\ell(F/\mu_n) \leq C \sum_{k=1}^N \varepsilon_{k-1} \mathrm{f}at^{\frac{1}{2}}_{\frac{\varepsilon_k}{8}}(F) \log\Big(\frac{2\mathrm{f}at_{\frac{\varepsilon_k}{8}}(F)}{\varepsilon}\Big) + 2\varepsilon_N n^{\frac{1}{2}} \ . \tag{3}$$

The latter part of Lemma 2 follows from its first part and Theorem 3. Before presenting the proof of Lemma 2, we require the following lemma, which is based on the classical inequality due to Slepian [14].

Lemma 3. *Let $(Z_i)_{i=1}^N$ be Gaussian random variables (i.e., $Z_i = \sum_{j=1}^m a_j g_j$ where (g_i) are independent standard Gaussian random variables). Then, there is some absolute constant C such that $\mathbb{E}\sup_i Z_i \leq C \sup_{i,j} \|Z_i - Z_j\|_2 \log^{\frac{1}{2}} N$.*

Proof (of Lemma 2). We may assume that F is symmetric and contains 0. The proof in the non-symmetric case follows the same path. Let μ_n be an empirical measure supported on $\{\omega_1, ...\omega_n\}$. For every $f \in F$, let $Z_f = n^{-1/2} \sum_{i=1}^n g_i f(\omega_i)$, where $(g_i)_{i=1}^n$ are independent standard Gaussian random variables on the probability space (Y, P). Set $\mathcal{Z}_F = \{Z_f | f \in F\}$ and note that since $V : L_2(\mu_n) \to L_2(Y, P)$ is an isometry for which $V(F/\mu_n) = \mathcal{Z}_F$ then

$$N\big(\varepsilon, F/\mu_n, L_2(\mu_n)\big) = N\big(\varepsilon, \mathcal{Z}_F, L_2(P)\big) \ .$$

Let $(\varepsilon_k)_{k=0}^\infty$ be a monotone sequence decreasing to 0 such that $\varepsilon_0 = 1$, and set $H_k \subset \mathcal{Z}_F$ to be a $2\varepsilon_k$ cover of $\mathcal{Z}_F$. Thus, for every $k \in \mathbb{Z}$ and every $Z_f \in \mathcal{Z}_F$ there is some $Z_f^k \in H_k$ such that $\left\|Z_f - Z_f^k\right\|_2 \leq 2\varepsilon_k$, and we select $Z_f^0 = 0$. Writing $Z_f = \sum_{k=1}^N (Z_f^k - Z_f^{k-1}) + Z_f - Z_f^N$ it follows that

$$\mathbb{E}\sup_{f\in F} Z_f \leq \sum_{k=1}^N \mathbb{E}\sup_{f\in F}(Z_f^k - Z_f^{k-1}) + \mathbb{E}\sup_{f\in F}(Z_f - Z_f^N) \ .$$

By the definition of Z_f^k and Lemma 3, there is an absolute constant C for which

$$\begin{aligned}\mathbb{E}\sup_{f\in F}(Z_f^k - Z_f^{k-1}) \leq& \mathbb{E}\sup\big\{Z_i - Z_j | Z_i \in H_k, Z_j \in H_{k-1}, \|Z_i - Z_j\|_2 \leq 4\varepsilon_{k-1}\big\} \\ \leq& C \sup_{i,j} \|Z_i - Z_j\|_2 \log^{\frac{1}{2}} |H_k| \, |H_{k-1}| \\ \leq& C\varepsilon_{k-1} \log^{\frac{1}{2}} N\big(\varepsilon_k, F, L_2(\mu_n)\big) \ .\end{aligned}$$

Since $Z_f^N \in \mathcal{Z}_F$, there is some $f' \in F$ such that $Z_f^N = Z_{f'}$. Hence,

$$\left(\sum_{i=1}^n \Big(\frac{f(\omega_i) - f'(\omega_i)}{\sqrt{n}}\Big)^2\right)^{\frac{1}{2}} = \|Z_f - Z_{f'}\|_2 \leq 2\varepsilon_N \ ,$$

which implies that for every $f \in F$ and every $y \in Y$,

$$\big|Z_f(y) - Z_f^N(y)\big| \leq \sum_{i=1}^n \left|\frac{f(\omega_i) - f'(\omega_i)}{\sqrt{n}} g_i(y)\right| \leq 2\varepsilon_N \Big(\sum_{i=1}^n g_i^2(y)\Big)^{\frac{1}{2}} \ .$$

Therefore, $\mathbb{E}\sup_{f\in F} Z_f - Z_f^N \leq \varepsilon_N \mathbb{E}\big(\sum_{i=1}^n g_i^2\big)^{\frac{1}{2}} = 2\varepsilon_N\sqrt{n}$, and the claim follows.

Using this result it is possible to estimate the ℓ-norm of classes with a polynomial fat-shattering dimension.

Theorem 5. *Let $F \subset B\big(L_\infty(\Omega)\big)$ and assume that there is some $\gamma > 1$ such that for any $\varepsilon > 0$, $\mathrm{fat}_\varepsilon(F) \leq \gamma\varepsilon^{-p}$. Then, there are absolute constants C_p, which depends only on p, such that for any empirical measure μ_n,*

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

Proof. Let μ_n be an empirical measure on Ω. If $p < 2$ then by Theorem 3,

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

and the bound on the ℓ-norm follows from the upper bound in Theorem 2.

Assume that $p \geq 2$ and, using the notation of Lemma 2, select $\varepsilon_k = 2^{-k}$ and $N = p^{-1}\log n$. By (3),

$$\begin{aligned} \ell(F/\mu_n) &\leq C_p(\gamma^{\frac{1}{2}}\log\gamma)\sum_{i=1}^N \varepsilon_k^{1-\frac{p}{2}}\log\frac{2}{\varepsilon_k} + 2\varepsilon_N n^{\frac{1}{2}} \\ &\leq C_p(\gamma^{\frac{1}{2}}\log\gamma)\sum_{i=1}^N k2^{k(\frac{p}{2}-1)} + 2n^{\frac{1}{2}-\frac{1}{p}} \ . \end{aligned}$$

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

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

whereas is $p > 2$ it is bounded by $C_p(\gamma^{\frac{1}{2}}\log\gamma)n^{\frac{1}{2}-\frac{1}{p}}$.

4 Complexity Estimates

Next, we investigate the sample complexity of Glivenko-Cantelli classes. The term "sample complexity" is often used in a slightly different way than the one we use in this article. Normally, when one talks about the sample complexity of a learning problem the meaning is the following: given a class $\mathcal{H}$ and some $1 \leq q < \infty$, let $\ell_q^h(x,y) = |h(x) - y|^q$ and set $\mathcal{Y}$ to be a bounded subset of $\mathbb{R}$. A *learning rule* is a mapping which assigns to each sample of arbitrary length $z_n = (x_i, y_i)_{i=1}^n$, some $A_{z_n} \in \mathcal{H}$. Given a class $\mathcal{H}$ 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 $\Omega \times \mathcal{Y}$,

$$P\left\{\mathbb{E}\,|A_{Z_n} - Y|^q \geq \inf_{h\in\mathcal{H}} \mathbb{E}\ell_q^h(X,Y) + \varepsilon\right\} < \delta \ ,$$

where Z_n is the sample $(X_i, Y_i)_{i=1}^n$ selected according to P. We denote the learning sample complexity associated with the range $\mathcal{Y}$ and the class $\mathcal{H}$ by $\mathcal{C}_q(\varepsilon, \delta, \mathcal{Y}, \mathcal{H})$.

It is possible to show [2] that if $\mathcal{Y} \subset [-M, M]$ then

$$\mathcal{C}_q(\varepsilon, \delta, \mathcal{Y}, \mathcal{H}) \leq \sup_{\|g\|_\infty \leq M} S_q(\varepsilon, \delta, g, \mathcal{H}) \ .$$

Hence, the GC sample complexity may be used to establish upper bounds on the sample complexity of learning problems.

We introduce a new parameter and show that it governs the GC sample complexity. This parameter is determined by the growth rate of the Rademacher averages associated with the class (see further details below), hence, by Theorem 2 and the fact that the Rademacher averages may be bounded using the ℓ-norm, it is possible to estimate the new parameter in terms of the covering numbers of the class.

4.1 Averaging Techniques

Let us start with a modified definition of the Rademacher averages associated with the class:

Definition 4. *Let F be a class of functions on Ω and let μ be a probability measure. Set*

$$R_n(F) = \sup_{(\omega_i)_{i=1}^n} \frac{1}{\sqrt{n}} \mathbb{E}_\varepsilon \sup_{f \in F} \left| \sum_{i=1}^n \varepsilon_i f(\omega_i) \right|, \quad \bar{R}_{n,\mu} = \frac{1}{\sqrt{n}} \mathbb{E}_\mu \mathbb{E}_\varepsilon \sup_{f \in F} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \ ,$$

where $(\varepsilon_i)_{i=1}^n$ are independent Rademacher random variables, μ is a probability measure and $(X_i)_{i=1}^n$ are independent, distributed according to μ.

Note that $R_n(F) = \sup_{\mu_n} R(F/\mu_n)$. Also, the relations between R_n and $\bar{R}_{n,\mu}$ are analogous to those between the VC dimension and the VC entropy; R_n is a "worst case" parameter, while $\bar{R}_{n,\mu}$ is an averaged version which takes into account the particular measure according to which one is sampling.

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]:

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

The parameter we wish to introduce measures the growth rate of the Rademacher averages as a function of the size of the sample. We show that it may be used instead of the usual combinatorial parameters, e.g, the fat-shattering dimension, to obtain deviation estimates.

Definition 5. *Let $F \subset B\big(L_\infty(\Omega)\big)$. For every $\varepsilon > 0$, let*

$$\mathrm{rav}_\varepsilon(F) = \sup\big\{n \in \mathbb{N} | R_n(F) \geq \varepsilon\sqrt{n}\big\} \ .$$

To see the connection between $fat_\varepsilon(F)$ and $\mathrm{rav}_\varepsilon(F)$, assume that $\{\omega_1, ..., \omega_n\}$ is ε-shattered. Let $(\varepsilon_1, ..., \varepsilon_n) \in \{-1, 1\}^n$ and set $I = \{i | \varepsilon_i = 1\}$. For every $J \subset \{\omega_1, ..., \omega_n\}$, let f_J be the function shattering J. Then, by the triangle inequality and selecting $f = f_I$, $f' = f_{I^c}$, it follows that

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

Thus, if $\{\omega_1, ..., \omega_n\}$ is ε-shattered, then for every realization of the Rademacher random variables, $n^{-1/2} \sup_{f \in F} |\sum_{i=1}^n \varepsilon_i f(\omega_i)| \geq \sqrt{n}\varepsilon$, while $\mathrm{rav}_\varepsilon(F)$ is determined by averaging such realizations. Hence, $\mathrm{rav}_\varepsilon(F) \geq fat_\varepsilon(F)$. It turn out that it is possible to find a general lower bound on $\mathrm{rav}_\varepsilon(F)$ in terms of $fat_\varepsilon(F)$ (see [10]).

The appeal in using the Rademacher averages instead of the fat-shattering dimension or other learning parameters is that the Rademacher averages remain unchanged if one takes the convex hull of the class, whereas $fat_\varepsilon(F)$ or the covering numbers may change dramatically by taking the convex hull. We shall use this property of the Rademacher averages to show that when assessing the GC sample complexity, there is no computational price to pay for taking the convex hull of the class.

Our first goal is to estimate the Rademacher averages of a loss class in terms of the Rademacher averages of the original class. The proof of the following claim is standard and is omitted.

Theorem 6. *Let F be a class of functions on Ω. Then,*

1. $R_n\big(\mathrm{absconv}(F)\big) = R_n(F)$.
2. *If Φ is a Lipschitz function such that $\Phi(0) = 0$ and L_Φ is its Lipschitz constant, then $R_n\big(\Phi(F)\big) \leq 2L_\phi R_n(F)$.*
3. *For any uniformly bounded function g let $F + g = \{f + g | f \in F\}$. There is an absolute constant C such that $R_n(F + g) \leq R_n(F) + C\|g\|_\infty$.*

From Theorem 6 we can derive the next corollary:

Corollary 2. *Let $\mathcal{H} \subset B\big(L_\infty(\Omega)\big)$ and $g \in B\big(L_\infty(\Omega)\big)$. If F_q is the q-loss function associated with the target g and* $\mathrm{conv}(\mathcal{H})$, *(i.e., each $f \in F_q$ is given by $f = |h - g|^q$, where $h \in \mathrm{conv}(\mathcal{H})$) then there is an absolute constant C such that*

$$R_n(F_q) \leq Cq\big(1 + R_n(\mathcal{H})\big) \ .$$

4.2 The Direct Approach

The first example of the direct approach we use to bound the rav, is for classes of linear functionals. This example is interesting in the context of Machine Learning since linear functionals are at the heart of the theory of Kernel Machines.

Recall that a Banach space has type p if there is a constant C such that for every integer n and every $x_1, ..., x_n \in X$,

$$\mathbb{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i x_i \right\| \leq C \Big(\sum_{i=1}^n \|x_i\|^p \Big)^{\frac{1}{p}} \quad , \tag{4}$$

where, as always, $(\varepsilon_i)_{i=1}^n$ are independent symmetric $\{-1, 1\}$-valued random variables. The smallest constant for which (4) holds is called the p type constant of X and is denoted by $T_p(X)$. Clearly, every Banach space has type $p = 1$, and it is possible to show that X can not have type p for $p > 2$. For example, Hilbert spaces have type 2 and $T_2 = 1$. For further details about the *type* of Banach spaces we refer the reader to [9, 13].

Let X be a Banach space which has a nontrivial type $1 < p \leq 2$ with a type constant $T_p(X)$. Let $\mathcal{H} \subset B(X^*)$ when considered as a class of functions on $B(X)$. If $x_1, ..., x_n \in B(X)$ then

$$\begin{aligned} \mathbb{E}_\varepsilon \sup_{h \in \mathcal{H}} \left| \sum_{i=1}^n \varepsilon_i h(x_i) \right| &\leq \mathbb{E}_\varepsilon \sup_{x^* \in B(X^*)} \left| \sum_{i=1}^n \varepsilon_i x^*(x_i) \right| = \mathbb{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i x_i \right\| \\ &\leq T_p(X) \Big(\sum_{i=1}^n \|x_i\|^p \Big)^{\frac{1}{p}} \quad . \end{aligned}$$

Since $\|x_i\| \leq 1$, $R_n(\mathcal{H}) \leq T_p(X) n^{\frac{1}{p} - \frac{1}{2}}$. Therefore, it follows that if $\|g\|_\infty \leq 1$ and if F_q is the q-loss associated with absconv($\mathcal{H}$) and a target g then

$$\mathrm{rav}_\varepsilon(F_q) \leq C \big(q T_p(X) \big)^{\frac{p}{p-1}} \Big(\frac{1}{\varepsilon} \Big)^{\frac{p}{p-1}} \quad .$$

The second example we present (which is very close in its nature to the first one) is that of Sobolev spaces.

Given a set $\Omega \subset \mathbb{R}^d$, let $W^{k,2}(\Omega)$ be the space of all the functions for which all the weak derivatives up to order k exist and belong to $L_2(\Omega)$.

Define $\|f\|_{W^{k,2}(\Omega)} = \sum_{|\alpha| \leq k} \|D^\alpha f\|_{L_2(\Omega)}$ and set $X = W_0^{k,2}(\Omega)$ to be the closure (with respect to the $\| \ \|_{W^{k,2}(\Omega)}$ norm) of the set of functions which vanish on the boundary of Ω and are continuously differentiable k times (see [1] for more details on Sobolev space).

It is possible to show that X is a Hilbert space, and therefore it has type 2 with $T_2(X) = 1$. Moreover, if $k > d/2$ then X is compactly embedded in the space of continuous functions on Ω. Hence, there is some $M > 0$ which depends on Ω such that for every $\omega \in \Omega$, $\|\delta_\omega\|_{X^*} \leq M$. Since X is a Hilbert space, each bounded linear functional is represented by an element of X, also denoted by δ_ω, which has the same norm as the functional.

Note that if $\mathcal{H}$ is a subset of unit ball of X then $R_n(\mathcal{H}) \le M$. Indeed, for any sample $\{\omega_1, ..., \omega_n\}$,

$$\frac{1}{\sqrt{n}}\mathbb{E}_\varepsilon \sup_{h\in\mathcal{H}} \sum_{i=1}^n \varepsilon_i h(\omega_i) = \frac{1}{\sqrt{n}}\mathbb{E}_\varepsilon \sup_{h\in\mathcal{H}} \sum_{i=1}^n \varepsilon_i \langle h, \delta_{\omega_i}\rangle = \frac{1}{\sqrt{n}}\mathbb{E}_\varepsilon \sup_{h\in\mathcal{H}} \langle \sum_{i=1}^n \varepsilon_i \delta_{\omega_i}, h\rangle \tag{5}$$

$$\le \frac{1}{\sqrt{n}}\mathbb{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i \delta_{\omega_i} \right\|_X \le \frac{1}{\sqrt{n}} \Big(\sum_{i=1}^n \|\delta_{\omega_i}\|_X^2 \Big)^{\frac{1}{2}} \le M \ .$$

In particular,

$$rav_\varepsilon(F_q) \le \frac{Cq^2(1+M)^2}{\varepsilon^2} \ .$$

Remark 2. Using a similar method one can estimate R_n of bounded subsets in the more general Sobolev spaces $W_0^{k,p}(\Omega)$.

4.3 Estimating rav Using Covering Numbers

We now present bounds on rav which are demonstrated via an indirect route, using the estimate on the ℓ-norm.

Recall that there is an absolute constant C such that for every F and every empirical measure μ_n supported on the sample $\{\omega_1, ..., \omega_n\}$, $R_n(F/\mu_n) \le C\ell(F/\mu_n)$. Combining this fact with the ℓ-norm estimate in Theorem 5, one can prove the following result on $\mathrm{rav}_\varepsilon(\mathcal{H})$ when $\mathrm{fat}_\varepsilon(\mathcal{H})$ is polynomial in ε^{-1}.

Theorem 7. *Let $\mathcal{H} \subset B\big(L_\infty(\Omega)\big)$ and assume that $\mathrm{fat}_\varepsilon(\mathcal{H}) \le \gamma\varepsilon^{-p}$ for some $\gamma > 1$ and every $\varepsilon > 0$. Then, there are constants C_p which depend only on p, such that for every $\varepsilon > 0$,*

$$\mathrm{rav}_\varepsilon(\mathcal{H}) \le C_p \begin{cases} (\gamma \log^2 \gamma)\varepsilon^{-2} & \text{if } 0 < p < 2 \\ \gamma\varepsilon^{-2}\log^4 \gamma\varepsilon^{-1} & \text{if } p = 2 \\ (\gamma^{\frac{p}{2}} \log^p \gamma)\varepsilon^{-p} & \text{if } p > 2 \ . \end{cases} \tag{6}$$

4.4 GC Sample Complexity

Here, we present the GC sample complexity estimates in terms of $\mathrm{rav}_\varepsilon(F)$. 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 show that $\mathrm{rav}_\varepsilon(F)$ measures the sample complexity of the class F.

Recall the following concentration result, which is due to Talagrand [15]:

Theorem 8. *There are two absolute constants K and $a \le 1$ with the following property: consider a class of functions F whose range is a subset of $[0,1]$, such that $\sup_{f\in F} \mathbb{E}(f - \mathbb{E}f)^2 \le a$. If μ is any probability measure on Ω and*

$$\sqrt{n} \ge K\bar{R}_{n,\mu}, \qquad M \ge K\bar{R}_{n,\mu}$$

then

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

Assume that members of $\mathcal{H}$ and g map Ω into $[0,1]$, and fix some $1\leq q<\infty$. Clearly, $F\equiv F_q$ is also a class of function whose range is a subset of $[0,1]$. Let a be as in Theorem 8, put $F^a=\{\sqrt{a}f|f\in F\}$ and note that $\sup_{f\in F^a}\mathbb{E}(f-\mathbb{E}f)^2\leq a$.

Lemma 4. *Let F and F^a be as in the above paragraph. There is an absolute constant C such that if $\varepsilon>0$ and n satisfy that*

$$n^{\frac{1}{2}}\geq KCa^{-\frac{1}{2}}\varepsilon^{-1}q\big(\bar{R}_{n,\mu}(\mathcal{H})+1\big)\ , \tag{7}$$

then

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

Proof. Clearly,

$$Pr\left\{\sup_{f\in F}|\mathbb{E}_{\mu_n}f-\mathbb{E}_\mu f|\geq\varepsilon\right\}=Pr\left\{\sup_{f\in F^a}|\mathbb{E}_{\mu_n}f-\mathbb{E}_\mu f|\geq\frac{\sqrt{a}\varepsilon}{2}\right\}\ .$$

Let $M=a^{1/2}n^{1/2}\varepsilon$. Since $a,\varepsilon\leq 1$, then if n satisfies (7) both conditions of Theorem 8 are automatically satisfies. The assertion follows directly from that theorem and the estimate on $\bar{R}_{n,\mu}(F_q)$ given in Corollary 2.

Theorem 9. *Assume that $\mathcal{H},g$ and q are as above and set F to be a q-loss class associated with $\mathcal{H}$ and g. 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 $n\geq C\max\big\{\mathrm{rav}_{C\varepsilon/q}(\mathcal{H}),\varepsilon^{-2}\log 1/\delta\big\}$.

4.5 Application: Smooth Functions

Let $\Omega\subset\mathbb{R}^d$ and set X to be the Sobolev space $W_0^{k,2}(\Omega)$. Assume that $k>d/2$ and that $\mathcal{H}\subset B(X)$. By the estimates on the Rademacher averages established in (5) one may obtain the GC sample complexity estimates for F_q.

Theorem 10. *Let $\mathcal{H}\subset B\big(W_0^{k,2}(\Omega)\big)$ and fix some g such that $\|g\|_\infty\leq 1$. Then,*

$$S_q\big(\varepsilon,\delta,g,\mathrm{conv}(\mathcal{H})\big)\leq\frac{C_{q,\Omega}}{\varepsilon^2}\log\frac{1}{\delta}\ ,$$

where $C_{q,\Omega}$ is a constant which depends only on q and Ω.

4.6 Application: Classes with Polynomial Fat-Shattering Dimension

The most important application of the theory presented here is a considerable improvement we in the GC sample complexity for classes with polynomial fat-shattering dimension:

Theorem 11. *Let $\mathcal{H}$ be a class of functions whose range in contained in $[0,1]$, such that $\mathrm{fat}_\varepsilon(\mathcal{H}) \leq \gamma\varepsilon^{-p}$ for some $p > 0$. Then, for every $1 \leq q < \infty$ there are constants $C_{p,q,\gamma}$, which depend only on p, q, γ such that for any $g : \Omega \to [0,1]$,*

$$S_q\big(\varepsilon, \delta, g, \mathrm{conv}(\mathcal{H})\big) \leq C_{p,q,\gamma} \begin{cases} \varepsilon^{-2}\log\delta^{-1} & \text{if } 0 < p < 2 \\ \varepsilon^{-2}(\log^4\varepsilon^{-1} + \log\delta^{-1}) & \text{if } p = 2 \\ \varepsilon^{-p}\log\delta^{-1} & \text{if } p > 2 \ . \end{cases} \tag{8}$$

Note that the best known estimates on the GC sample complexity were demonstrated in [3, 5]. It was shown that if $\mathcal{H}$ is a GC class then

$$S_q(\varepsilon, \delta, g, \mathcal{H}) \leq C\frac{1}{\varepsilon^2}\Big(\mathrm{fat}_{\frac{\varepsilon}{4}}(\mathcal{H})\log^2\frac{1}{\varepsilon} + \log\frac{1}{\delta}\Big) \ . \tag{9}$$

If the fat-shattering dimension is polynomial, this result yields a bound which is $O(\varepsilon^{-(p+2)})$ up to logarithmic factors in ε^{-1} and δ^{-1}. Thus, (8) is a much better bound even regarding the sample complexity of $\mathcal{H}$ itself. If one were to try and estimate the sample complexity of a q-loss class associated with $\mathrm{conv}(\mathcal{H})$ using (9), the difference in the results is even more noticeable, since the fat-shattering dimension of the convex hull may be increased by a factor of ε^{-2}, whereas rav does not change.

5 Gelfand Numbers and Sufficient Statistics of GC Classes

In this final section we present another application of the ℓ-norm estimates. We investigate the number of linear constraints needed to pinpoint a function in a given class, up to a desired accuracy. Thus, the problem we face is as follows: given a class F and some $\varepsilon > 0$, we attempt to find a "small" set $\Gamma(\varepsilon) = (x_i^*)_{i=1}^m$ of linear functionals on F, (which may depend on ε) with the following properties:

1. For every $f \in F$ and every $1 \leq i \leq m$, $x_i^*(f)$ can be computed using empirical data.
2. If two functions $f, g \in F$ agree on every element in $\Gamma(\varepsilon)$ then $\|f - g\|^2_{L_2(\mu)} < \varepsilon$.

The formal definition of ε-Sufficient Statistics is as follows:

Definition 6. *Let F be a class of functions defined on a set Ω and let μ be a probability measure on Ω. A set of linear empirical functionals $(x_i^*)_{i=1}^m$ is called ε-sufficient statistics with respect to $L_2(\mu)$ if for any $f, g \in F$ which satisfy that $x_i^*(g) = x_i^*(f)$ for every $1 \leq i \leq m$, $\|f - g\|^2_{L_2(\mu)} < \varepsilon$.*

If F is a Glivenko-Cantelli class then the set $\Gamma(\varepsilon)$ may be constructed by taking the point evaluation functionals on a large enough sample. Indeed, if F is a Glivenko-Cantelli class, then the set $\{(f-g)^2 | f,g \in F\}$ is also a Glivenko-Cantelli class. Take n to be an integer such that there is an empirical measure μ_n which is supported on a sample S_n, and for every $f,g \in F$,

$$\left|\mathbb{E}_\mu (f-g)^2 - \mathbb{E}_{\mu_n} (f-g)^2\right| < \varepsilon \ . \tag{10}$$

Set $\Gamma(\varepsilon)$ to be the set of point evaluation functionals $\{\delta_{\omega_i} | \omega_i \in S_n\}$. Thus, each element of $\Gamma(\varepsilon)$ is a linear functional on F and if f,g agree on each δ_{ω_i} (i.e., if $f(\omega_i) = g(\omega_i)$) then $\|f-g\|^2_{L_2(\mu)} < \varepsilon$. Hence, $\Gamma(\varepsilon)$ is ε-sufficient statistics for F in $L_2(\mu)$.

Of course, there in no need to merely use point evaluation functionals. As explained in [11], even from the computational point of view it is possible to find a set of linear combinations of point evaluation functionals which are ε-sufficient statistics, using a random selection scheme. This idea is based on the so-called Gelfand numbers.

The Gelfand numbers are parameters which measure the "size" of a symmetric convex set. In some sense, it measures the "minimal width" of the symmetric convex hull of the set. Formally, given some $1 \le k \le n$ let H_k be a k-codimensional subspace of ℓ_2^n. A k-section of F is an intersection of F with some H_k. The k-th Gelfand number of $F \subset \ell_2^n$, denoted by $c_k(F)$, is the "smallest" possible diameter of a k-codimensional section of F. In our case, the Euclidean structure is that of $L_2(\mu_n)$, which is isometric to ℓ_2^n. Note that if the measure is supported on $\{\omega_1, ..., \omega_n\}$, a k-codimensional subspace is the intersection of the kernels of k linear functionals, every one of which is given by $x^* = \sum_{i=1}^n a_i \delta_{\omega_i}$. Assume that we can find a k-codimensional section of $\mathrm{absconv}(F)/\mu_n$ which has a diameter bounded by α, and let $x_1^*, ..., x_k^*$ be the linear functionals which define this section. Thus, each x_i^* is empirical, since $x_i^*(f)$ can be computed using the sample points, and if $x_i^*(f) = x_i^*(g)$ then $(f-g)/2 \in \mathrm{absconv}(F) \cap H_k$, implying that $\|f-g\|_{L_2(\mu_n)} \le \alpha$. Therefore, if μ_n is an empirical measure such that (10) holds, then the set $\Gamma = (x_i^*)_{i=1}^k$ is $(\varepsilon + \alpha^2)$ sufficient statistics for F in $L_2(\mu)$.

It is possible to estimate the Gelfand numbers of a set $F \subset \ell_2^n$ using the ℓ-norm. This fact is due to Pajor and Tomczak-Jaegermann [14].

Theorem 12. *There is an absolute constant C such that for every integer n and every $F \subset \ell_2^n$, $\sup_{k \ge 1} k^{\frac{1}{2}} c_k(F) \le C\ell(F)$.*

Since the ℓ-norm does not change by taking the symmetric convex hull, we can prove the following estimates, improving the bound presented in [11]

Theorem 13. *Let F be a class of functions which map Ω into $[0,1]$ and assume that there are $\gamma \ge 1$ and $p > 0$ such that for every $\varepsilon > 0$, $\mathrm{fat}_\varepsilon(F) \le \gamma \varepsilon^{-p}$. For every probability measure μ and $\varepsilon > 0$ there is a set $\Gamma(\varepsilon)$ of ε-sufficient statistics such that*

$$|\Gamma(\varepsilon)| \le C_{\gamma,p} \begin{cases} \varepsilon^{-1} & \text{if } 0 < p < 2 \\ \varepsilon^{-1} \log^2 \varepsilon^{-1} & \text{if } p = 2 \\ \varepsilon^{-(p-1)} & \text{if } p > 2 \ , \end{cases}$$

where $C_{\gamma,p}$ are constants which depend only on γ and p.

Proof. We present a proof only when $p > 2$. The proof in the other cases follows in a similar fashion.

Fix $\varepsilon > 0$, and let n be an integer such that

$$Pr\Big\{ \sup_{f,g\in F} \left|\mathbb{E}_\mu (f-g)^2 - \mathbb{E}_{\mu_n} (f-g)^2\right| < \frac{\varepsilon}{2} \Big\} \geq \frac{1}{2} \ . \tag{11}$$

To find such an integer n, we use the GC sample complexity estimates. Indeed, let $\mathcal{H} = (F-F)^2 = \{(f-g)^2 | f,g \in F\}$ and note that since each member of F maps Ω into $[0,1]$ then $F - F \subset B\big(L_\infty(\Omega)\big)$. Clearly, $\phi(t) = t^2$ is a Lipshcitz function on $[-1,1]$ with a constant 2. Therefore, by Theorem 6,

$$R_n(\mathcal{H}) \leq 4R_n(F-F) = 8R_n\Big(\frac{1}{2}(F-F)\Big) = 8R_n(F) \ ,$$

where the last inequality holds because $\frac{1}{2}(F-F) \subset \mathrm{absconv}(F)$. Thus, by Theorem 9 we may select $n = C_{p,\gamma}\varepsilon^{-p}$.

Since the set in (11) is nonempty, there is an empirical measure μ_n supported on $\{\omega_1, ..., \omega_n\}$ such that for every $f, g \in F$

$$\left|\mathbb{E}_\mu (f-g)^2 - \mathbb{E}_{\mu_n} (f-g)^2\right| < \frac{\varepsilon}{2} \ .$$

Let $G \subset L_2(\mu_n)$ be the symmetric convex hull of F/μ_n. By the ℓ-norm estimate, for every integer $k \leq n$ there are k linear functionals on $L_2(\mu_n)$, denoted by $(x_i^*)_{i=1}^k$, such that if $f, g \in G$ satisfy that for every $1 \leq i < k$, $x_i^*(f) = x_i^*(g)$ then

$$\|f-g\|_{L_2(\mu_n)} \leq C_{p,\gamma} \frac{n^{\frac{1}{2}-\frac{1}{p}}}{\sqrt{k}} \ .$$

In particular, the same holds if $f, g \in F$. Thus, by the selection of n and μ_n, and for such f and g

$$\|f-g\|_{L_2(\mu)}^2 \leq \|f-g\|_{L_2(\mu_n)}^2 + \frac{\varepsilon}{2} < \frac{C_{p,\gamma}^2}{k}\Big(\frac{1}{\varepsilon}\Big)^{p-2} + \frac{\varepsilon}{2} \ .$$

Note that each $(x_i^*)_{i=1}^k$ is a linear combination of point evaluation functionals δ_{ω_i}. Thus, if $k = \lceil (C_{p,\gamma})^2 \varepsilon^{-(p-1)} \rceil$ then $\|f-g\|_{L_2(\mu)}^2 < \varepsilon$, implying that the set $(x_i^*)_{i=1}^k$ are ε-sufficient statistics, as claimed.

This result implies that the number of sufficient statistics is considerably smaller than the estimate one has by taking the point evaluation functionals. Indeed, if one is restricted to point evaluation functionals, then the number of such functionals needed is given by the sample complexity of the class $\{(f-g)^2 | f,g \in F\}$, which is $O(\varepsilon^{-\max\{2,p\}})$ when $\mathrm{fat}_\varepsilon(F) = O(\varepsilon^{-p})$. Note that when $p > 2$, this bound is optimal [10]. Hence, allowing linear combinations always yields considerably better bounds than those given by using only point evaluation functionals.

References

1. R.A. Adams: *Sobolev Spaces*, Pure and Applied Mathematics series 69, Academic Press 1975.
2. M.Anthony, P.L. Bartlett: *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, 1999.
3. 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.
4. P.L. Bartlett, S.R. Kulkarni, S.E. Posner: Covering numbers for real valued function classes, IEEE transactions on information theory, 43(5), 1721–1724, 1997.
5. P.L. Bartlett, P. Long: More theorems about scale sensitive dimensions and learning, *Proceedings of the 8th annual conference on Computation Learning Theory*, 392-401, 1995.
6. R.M. Dudley: The sizes of compact subsets of Hilbert space and continuity of Gaussian processes, J. of Functional Analysis 1, 290-330, 1967.
7. R.M. Dudley, E. Giné, J. Zinn: Uniform and universal Glivenko–Cantelli classes, J. Theoret. Prob. 4, 485–510, 1991.
8. M. Ledoux, M. Talagrand: *Probability in Banach spaces*, Springer Verlag 1992.
9. J. Lindenstrauss, L. Tzafriri: *Classical Banach Spaces* Vol II, Springer Verlag.
10. S. Mendelson: Rademacher Averages and phase transitions in Glivenko-Cantelli classes, preprint.
11. S. Mendelson, N. Tishby: Statistical Sufficiency for Classes in Empirical L_2 Spaces, Proceedings of the 13th annual conference on Computational Learning Theory, 81-89, 2000.
12. A. Pajor: *Sous espaces ℓ_1^n des espaces de Banach*, 1985
13. G. Pisier: Probabilistic methods in the geometry of Banach spaces, *Probability and Analysis*, Lecture notes in Mathematics 1206, 167–241, Springer Verlag 1986.
14. G. Pisier: *The volume of convex bodies and Banach space geometry*, Cambridge University Press, 1989.
15. M. Talagrand: Sharper bounds for Gaussian and empirical processes, Annals of Probability, 22(1), 28-76, 1994.
16. N. Tomczak–Jaegermann: *Banach–Mazur distance and finite–dimensional operator Ideals*, Pitman monographs and surveys in pure and applied Mathematics 38, 1989
17. V. Vapnik, A. Chervonenkis: Necessary and sufficient conditions for uniform convergence of means to mathematical expectations, Theory Prob. Applic. 26(3), 532-553, 1971
18. A.W. Van–der–Vaart, J.A. Wellner: *Weak convergence and Empirical Processes*, Springer-Verlag, 1996.