ℓ–norm and its Application to Learning Theory

S. Mendelson [1]

AMS classification numbers: 41B46, 46B07, 68T05

Abstract

We investigate connections between an important parameter in the theory of Banach spaces called the ℓ-norm, and two properties of classes of functions which are essential in Learning theory - the uniform law of large numbers and the Vapnik–Chervonenkis (VC) dimension. We show that if the ℓ-norm of a set of functions is bounded in some sense, then the set satisfies the uniform law of large numbers. Applying this result, we show that if X is a Banach space which has a nontrivial type, then the unit ball of its dual satisfies the uniform law of large numbers. Next, we estimate the ℓ-norm of a set of $\{0,1\}$-functions in terms of its VC dimension. Finally, we present a "Gelfand number" like estimate to certain classes of functions. We use this estimate to formulate a learning rule, which may be used to approximate functions from the unit balls of several Banach spaces.

Key words: Glivenko-Cantelli classes, VC dimension, ℓ-norm

[1] Institute of Computer Science, Hebrew University, Jerusalem, Israel

1 Introduction

One of the fundamental tools in the local theory of Banach spaces is the so–called ℓ-norm. In its usual context, the ℓ-norm is a norm on the space of operators from ℓ_2^n to a Banach space X.

Definition 1.1 *Let $u : \ell_2^n \to X$ then*

$$\ell(u) = \Big(\int_{\mathbb{R}^n} \|ux\|_X^2 \, d\gamma_n(x) \Big)^{1/2}$$

where γ_n is the Gaussian measure on $\mathbb{R}^n$. If $u : L_2 \to X$ then $\ell(u) = \sup \ell(u_{|H})$ and the supremum is taken with respect to all finite dimensional subspaces of L_2.

Recall that if F is a convex symmetric subset of $\mathbb{R}^n$ which has a nonempty interior, then F is the unit ball of some norm denoted by $\| \ \|_F$. Set $\| \ \|_{F^*}$ to be the dual norm to F and define

$$\ell(F) = \Big(\int_{\mathbb{R}^n} \|x\|_{F^*}^2 \, d\gamma_n(x) \Big)^{1/2}.$$

Thus, if $I : \ell_2^n \to F^*$ is the formal identity operator, then $\ell(F) \equiv \ell(I)$.

In this paper we explore possible connections between the ℓ-norm and other parameters in the realm of learning theory.

The basic question in Learning Theory is that of generalization. Assume that $g : \Omega \to \mathbb{R}$ is an unknown function selected from a class of functions $\mathcal{F}$. Assume further that μ is an (unknown) probability measure on Ω, and that $\omega_1, ..., \omega_n$ are selected independently according to μ. The "student" in the learning game receives as data a set s_n which consists of the sample (ω_i) and the values of the unknown function g on the sample. In return, he has to provide some function $A_{s_n} \in \mathcal{F}$, which is determined by s_n. The learning rule A_{s_n} has to be efficient in the sense that for the majority of the samples, A_{s_n} is close to g in $L_2(\mu)$.

Usually, this learning rule is established in two steps. The first step is based on the empirical data, namely, if μ_n is the empirical measure supported on $(\omega_1, ..., \omega_n)$, the first step is to find an element of $\mathcal{F}$ which approximates g in $L_2(\mu_n)$. The second step is to show that if n is large enough, then with high probability, $\|f - g\|_{L_2(\mu)}$ is close to $\|f - g\|_{L_2(\mu_n)}$ for every $f \in \mathcal{F}$.
As shown in the text, a sufficient condition which ensures that with high probability

$$\lim_{n \to \infty} \sup_{f \in \mathcal{F}} \Big| \|f - g\|_{L_2(\mu)} - \|f - g\|_{L_2(\mu_n)} \Big| = 0$$

is that the members of $\mathcal{F}$ satisfy the law of large numbers uniformly. Classes of functions which satisfy the uniform law of large numbers are called Glivenko–Cantelli classes.

From a quantitative point of view, it is important to estimate the size of the sample $(\omega_1, ..., \omega_n)$ needed to ensure that

$$Pr\Big\{ \sup_{f \in \mathcal{F}} \Big| \|f - g\|_{L_2(\mu)} - \|f - g\|_{L_2(\mu_n)} \Big| \geq \varepsilon \Big\} \leq \delta.$$

In the case of $\{0,1\}$ functions, this estimate is usually obtained using the so-called Vapnik–Chervonenkis (VC) dimension of the class (see a definition in the next section). In the general case (i.e., for classes of functions which are not $\{0,1\}$), one uses the parametric version of the VC dimension, called the fat-shattering dimension (see [15], [1]). Thus, the ability to estimate the VC dimension of a class is of crucial importance in Learning Theory, since this is a possible way to obtain complexity estimates.

The goal of this paper is to demonstrate that the ℓ-norm may serve as an additional parameter, which, in some cases, may replace the VC dimension or the fat-shattering dimension, and assist in establishing complexity estimates.

To see how the ℓ-norm may take part in this game, assume that $\mathcal{F}$ is a set of functions on Ω, all of which are bounded by 1. Let K be the symmetric convex hull of $\mathcal{F}$, i.e.,

$$K = \Big\{ \sum_{i=1}^{n} \lambda_i f_i \Big| \sum_{i=1}^{n} |\lambda_i| = 1,\ f_i \in \mathcal{F},\ n \in \mathbb{N} \Big\}.$$

Let μ be a probability measure on Ω and set μ_n to be an empirical measure supported on an i.i.d. sample selected according to μ. Let $\chi_{\omega_i} \in L_2(\mu_n)$ be the characteristic function of $\{\omega_i\}$, and put

$$K/\mu_n = \Big\{ \sum_{i=1}^{n} k(\omega_i)\chi_{\omega_i} | k \in K \Big\} \subset L_2(\mu_n).$$

We will show that if

$$\big(\mathbb{E}\ell^2(K/\mu_n)\big) = o(\sqrt{n})$$

then K is a Glivenko–Cantelli class. Moreover, it is possible to provide complexity bounds in terms of the empirical ℓ-norms. As an example, we estimate the empirical ℓ-norms of the dual unit ball of spaces X which have a nontrivial type. We use this result to prove that if X is a Banach space with type $p > 1$ then the unit ball of the dual is a Glivenko–Cantelli class of functions on the unit ball of X. Using the estimate on the empirical ℓ-norms combined with a famous result due to Pajor and Tomczak-Jaegermann (see [8]), we present an upper bound on the Gelfand numbers of classes of functions on a set Ω which are contained in the unit ball of a Banach space X, provided that the point evaluation functionals δ_ω are uniformly bounded and that X^* has type 2.

Next, we show that if $\mathcal{F}$ is a VC class (i.e., has a finite VC dimension) with $VC(\mathcal{F}) = d$, then there is an absolute constant C such that for every empirical measure μ_n, $\ell(K/\mu_n) \leq Cd^{1/2}$, where K is the symmetric convex hull of $\mathcal{F}$. We use this result to prove that a class of $\{0,1\}$ functions is Glivenko–Cantelli if and only if the empirical ℓ-norms of $\mathcal{F}$ are uniformly bounded.

Finally, we show that if the empirical ℓ-norms of a class are bounded, then its symmetric convex hull has a section of codimension k which has a "small" diameter (i.e., we obtain a "Gelfand number" like estimate). Moreover, the functionals which determine this section may be calculated empirically. Since

the essence of Learning Theory is to approximate an unknown function using empirical data, such a result may come in handy. As an example, we present a complete learning scheme to an unknown function on Ω which belongs to the unit ball of a Hilbert space in which the point evaluation functionals δ_ω are bounded by 1.

2 Preliminary Results

This section is devoted to basic definitions, some notation and results which will be used in the sequel. Given a Banach space X, the *dual* of X is denoted by X^* and B_X is the unit ball of X. Let ℓ_2^n be $\mathbb{R}^n$ equipped with the norm $\|x\|_2 = \left(\sum_1^n |x_i|^2\right)^{\frac{1}{2}}$. For every probability measure μ on Ω, let $\mathbb{E}_\mu$ denote the expectation with respect to μ. In fact, we shall always assume that μ is a Borel measure. For every $\omega \in \Omega$, let δ_ω be the point evaluation functional. Thus, for every function f on Ω, $\delta_\omega(f) = f(\omega)$. Denote by μ_n an empirical measure supported on a set of n points, i.e., $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{\omega_i}$. Given a set A, let $|A|$ be its cardinality. Throughout, all absolute constants are denoted by C or c. Their values may change from line to line.

Definition 2.1 *Let $F \subset \ell_2^n$ be convex and symmetric. Set*

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

F^{o} is called the polar of F.

Note that if F has a nonempty interior, then F^{o} is the unit ball of the norm $\|\ \|_{F^*}$, which is the dual norm to the norm defined by F.

We wish to extend the definition of the ℓ-norm to a set F which is not necessarily convex:

Definition 2.2 *For $F \subset \ell_2^n$, let*

$$\ell(F) = \left(\int_{\mathbb{R}^n} \sup_{f\in F} \left|\langle f, x\rangle\right|^2 d\gamma_n\right)^{\frac{1}{2}} \tag{2.1}$$

where γ_n is the Gaussian measure on $\mathbb{R}^n$. If $F \subset L_2$ then $\ell(F) = \sup_H \ell(F\cap H)$, where the supremum is taken with respect to all finite dimensional subspaces of L_2 and $F \cap H$ is identified as ℓ_2^n by the natural isometry.

Let K be the symmetric convex hull of $F \subset \ell_2^n$ and assume it has a nonempty interior. It is easy to see that if $g_1, ..., g_n$ are independent standard Gaussian random variables on some probability space and if $e_1, ..., e_n$ is an orthonormal basis in ℓ_2^n then $\ell(F) = (\mathbb{E}\|\sum_{i=1}^n g_i e_i\|_{K^*}^2)^{1/2}$. Indeed, since the dual norm is determined by the extreme points of K, which all belong to $F \cup -F$, then

$$\mathbb{E}\left\|\sum_{i=1}^n g_i e_i\right\|_{K^*}^2 = \int_{\mathbb{R}^n} \sup_{f\in F\cup -F} \langle x, f\rangle^2 d\gamma_n,$$

implying that $\ell(K) = \ell(F)$.

If (X, d) is a metric space and if $\mathcal{F} \subset X$, denote by $N(\varepsilon, \mathcal{F}, d)$ the minimal number of balls with radius ε (with respect to the metric d) needed to cover $\mathcal{F}$. $N(\varepsilon, \mathcal{F}, d)$ are called the covering numbers of $\mathcal{F}$. In cases where the subset $\mathcal{F}$ is obvious, we shall use the notation $N(\varepsilon, d)$, and in cases where the metric is clear we shall denote the covering numbers by $N(\varepsilon, \mathcal{F})$.

The next theorem demonstrates the connection between the covering numbers of a set F and $\ell(F)$. The upper bound was established by Dudley in [2], while the lower one is due to Sudakov (see [13]). A proof of both bounds may be found in [9].

Theorem 2.3 *Let $F \subset \ell_2^n$. Then, there are absolute positive constants c and C such that*

$$c \sup_{\varepsilon > 0} \varepsilon \sqrt{\log(N(\varepsilon, F))} \leq \ell(F) \leq C \int_0^\infty \sqrt{\log(N(\varepsilon, F))} d\varepsilon \tag{2.2}$$

Next, we examine some basic definitions and results from learning theory.

Definition 2.4 *Let $\mathcal{F}$ be a class of $\{0,1\}$ functions on a space Ω. We say that $\mathcal{F}$ shatters $\omega_1, ..., \omega_n$, if for every $I \subset \{1, ..., n\}$ there is a function $f \in \mathcal{F}$ for which $f(\omega_i) = 1$ if $i \in I$ and $f(\omega_j) = 0$ if $j \notin I$. Define the Vapnik-Chervonenkis dimension of $\mathcal{F}$ by*

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

Given a probability measure μ on Ω, denote by Pr the infinite product measure μ^∞. If $(\omega_1, ..., \omega_n, ...)$ is an infinite i.i.d. sample selected according to μ, let μ_n be an empirical measure supported on $(\omega_1, ..., \omega_n)$. Glivenko–Cantelli classes (defined below) are classes of functions in which, with high probability, random empirical measures μ_n approximate the measure μ uniformly on the elements of the class.

Definition 2.5 *$\mathcal{F}$ is called a Glivenko Cantelli class (GC class) with respect to a family of measures Λ if for every $\varepsilon > 0$*

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

We end this section with several results which connect the Glivenko–Cantelli condition to estimates on the covering numbers. All of the results are stated for classes of functions which are bounded by 1. The results remain valid for classes of functions with a uniformly bounded range - up to an constant which depends only on that bound.

We start with an estimate which is at the heart of the proof of the Glivenko-Cantelli Theorem (see [11]). It provides a sufficient condition for GC in terms of the $L_1(\mu_n)$-covering numbers of the class.

Theorem 2.6 *Let $\mathcal{F}$ be a class of functions on Ω, all of which are bounded by 1, and let μ be a probability measure on Ω. Then, there is an absolute constant C such that for every $\varepsilon > 0$,*

$$Pr\{\sup_{f\in\mathcal{F}} |\mathbb{E}_\mu f - \mathbb{E}_{\mu_n} f| > \varepsilon\} \leq \tag{2.3}$$

$$\leq 8\Big(exp(-n\varepsilon^2/C) + Pr\Big\{\log N\big(\varepsilon/8, \mathcal{F}, L_1(\mu_n)\big) > \frac{n\varepsilon^2}{C}\Big\}\Big). \tag{2.4}$$

In particular, if for every $\varepsilon > 0$,

$$\sup_{\mu_n} \log N\big(\varepsilon, \mathcal{F}, L_1(\mu_n)\big) = o(n)$$

then $\mathcal{F}$ is a GC class with respect to the measure μ.

Remark 1 *Note that for every probability measure μ and every $\varepsilon > 0$,*

$$N\big(\varepsilon, F, L_1(\mu)\big) \leq N\big(\varepsilon, F, L_2(\mu)\big).$$

Thus, to prove that $\mathcal{F}$ is GC with respect to μ, it suffices to show that

$$\sup_{\mu_n} \log N\big(\varepsilon, \mathcal{F}, L_2(\mu_n)\big) = o(n).$$

This next result is due to Dudley, Giné and Zinn (see [4]).

Theorem 2.7 *Let $\mathcal{F}$ be a class of functions on Ω which are all bounded by 1. Then, $\mathcal{F}$ is a GC class with respect to all Borel probability measures if and only if for every $\varepsilon > 0$,*

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

Corollary 2.8 *Let $\mathcal{F}$ be a GC class of functions on Ω which are all bounded by 1. Then, for every $h \in \mathcal{F}$, the class $\mathcal{F}_h = \{(f-h)^2 | f \in \mathcal{F}\}$ is a GC class too.*

The proof of this claim is based on the fact that the $L_\infty(\mu_n)$–covering numbers of $\mathcal{F}_h$ may be estimates by the $L_\infty(\mu_n)$–covering numbers of $\mathcal{F}$.

3 ℓ-Norm and Glivenko–Cantelli Classes

Let $\mathcal{F}$ be a class of uniformly bounded functions on a domain Ω. We will show that if $\ell(\mathcal{F})$ is bounded in some sense, then $\mathcal{F}$ is a GC class. We use this result to show that if X has type $p > 1$ (defined below), then B_{X^*} is a GC class of functions on B_X with respect to a certain family of measures.

Let $\{\omega_1, ..., \omega_n\} \subset \Omega$ be an i.i.d. sample selected according to the measure μ and denote the empirical measure associated with this sample by μ_n. Set $F/\mu_n = \{\sum_{i=1}^n f(\omega_i)\chi_{\omega_i} | f \in \mathcal{F}\}$ and put $\ell_n = \big(\mathbb{E}\ell^2(\mathcal{F}/\mu_n)\big)^{1/2}$, where the expectation is with respect to the product measure μ^∞.

Theorem 3.1 *If $\ell_n = o(\sqrt{n})$ then $\mathcal{F}$ is a GC class with respect to the measure μ. Moreover, if $\ell = \sup_n \ell_n$ is finite, then there is some absolute constant C such that for every $\varepsilon > 0$ and $0 < \delta \leq 1$,*

$$Pr\Big\{ \sup_{f \in \mathcal{F}} |\mathbb{E}_\mu f - \mathbb{E}_{\mu_n} f| \geq \varepsilon \Big\} < \delta$$

provided that

$$n \geq C \max\Big\{ \frac{\log \frac{1}{\delta}}{\varepsilon^2}, \frac{\ell^2}{\varepsilon^4 \delta} \Big\}.$$

Proof: First, note that $L_2(\mu_n)$ is isometric to ℓ_2^n. Thus, by Theorem 2.3, there is some absolute constant C such that for every empirical measure μ_n and every $\varepsilon > 0$,

$$\log N\big(\varepsilon, \mathcal{F}/\mu_n, L_2(\mu_n)\big) \leq C \frac{\ell^2(\mathcal{F}/\mu_n)}{\varepsilon^2}.$$

Hence, by Chebyshev's inequality and since $N\big(\varepsilon, L_1(\mu_n)\big) \leq N\big(\varepsilon, L_2(\mu_n)\big)$ it follows that

$$\begin{aligned} Pr\Big\{ \log N\big(\varepsilon/8, \mathcal{F}/\mu_n, L_1(\mu_n)\big) \geq n\varepsilon^2/C \Big\} \leq & \frac{\mathbb{E} \log N\big(\varepsilon/8, \mathcal{F}/\mu_n, L_2(\mu_n)\big)}{n\varepsilon^2/C} \leq \\ \leq & C \frac{\ell_n^2}{n\varepsilon^4}. \end{aligned}$$

Now, the claim follows by Theorem 2.6.

■

One class of functions which is particularly interesting is the unit ball of a dual space. In the finite dimensional setting, B_{X^*} is a GC class of functions on any bounded subset of X. On the other hand, it is far from obvious whether the same holds in the infinite dimensional case. The infinite dimensional case is important because linear separators on an infinite dimensional space are at the heart of the Support Vector Machine procedure, which is highly regarded by the Learning Theory community (see [15]).

Formally, the question we investigate is as follows: let X be a Banach space. Is B_{X^*} a GC class of functions on B_X? It was shown in [7] that the answer is determined by a geometric property of the space X called type.

Definition 3.2 *A Banach space X has Gaussian type p if there is some constant C such that for every $x_1, ..., x_n \in X$,*

$$\Big(\mathbb{E} \Big\| \sum_1^n g_i x_i \Big\|^2 \Big)^{1/2} \leq C \Big(\sum_1^n \|x_i\|^p \Big)^{1/p}. \tag{3.1}$$

The basic facts concerning the concept of type may be found, for example, in [6] or in [14]. Clearly, for every Banach space (3.1) holds in the case $p = 1$. The best constant C which satisfies (3.1) is called the type constant and is denoted by $T_p(X)$. If $K \subset \ell_2^n$ is convex, symmetric with a nonempty interior, set $T_p(K)$ to be the type constant of the norm induced by K.

A similar definition of type may be given using Rademacher random variables instead of Gaussian variables, but the two definitions are equivalent (see [14]).

In [7] it was shown that B_{X^*} is a GC class of functions on B_X if and only if X has type $p > 1$. The idea behind the proof in [7] was to find a tight bound on the so–called fat shattering dimension of B_{X^*}, which is a parametric version of the VC dimension. Here, we present a completely different approach to this question. We show that if X has type $p > 1$ then B_{X^*} is a GC class with respect to a certain family of measures using an estimate on the ℓ-norms of the class. It turns out that the sample complexity estimates obtained using this method are better than those found in [7].

Note that there is an easier proof to the fact that if X has type $p > 1$ then B_{X^*} is a GC class. The proof is based on an idea which is due to Maurey (see [10]). However, the estimates for $\ell(B_{X^*}/\mu_n)$ which are at the heart of the proof we present, serve other purposes in the sequel.

Lemma 3.3 *Let X be a Banach space which has type $p > 1$. Then, for every empirical measure μ_n on B_X which is supported on a linearly independent sample, $\ell(B_{X^*}/\mu_n) \leq T_p(X) n^{1/p-1/2}$.*

Proof: Note that if X has type p then X^{**} has type p and $T_p(X) = T_p(X^{**})$ (see [14], pg. 83). Set $T_p = T_p(X)$ and let $x_1, ..., x_n \in B_X$ be linearly independent. Set $V = span\{x_1, ..., x_n\}$ and let $Y = (V, \| \ \|_Y)$, where $\| \ \|_Y$ is given by $\|\sum \lambda_i x_i\|_Y = (\sum \lambda_i^2)^{1/2}$. Put $E = \{\sum x^*(x_i)x_i \,|\, \|x^*\| \leq 1\} \subset Y$. Clearly, Y is an inner product space and (x_i) is an orthonormal basis in Y. Also, $E \subset Y$ is symmetric, convex and has a nonempty interior.

Note that $T_p(E^o) \leq T_p$. Indeed, define $Q : X^* \to (V, \| \ \|_E)$ by $Qx = \big(x^*(x_1), ..., x^*(x_n)\big)$. Since $Q(B_{X^*}) = E$ then E is isometric to a quotient of X^* and therefore $(V, \| \ \|_{E^o})$ is isometric to a subspace of X^{**}. Thus, $T_p(E^o) \leq T_p$ as claimed.

Let μ_n be the empirical measure supported on the sample $x_1, ..., x_n \in X$. Since $\|x^*\|_{L_2(\mu_n)} = \Big(\frac{1}{n}\sum_{i=1}^n \big(x^*(x_i)\big)^2\Big)^{1/2}$ then $\|x^*\|_{L_2(\mu_n)} = \|Q(x^*/\sqrt{n})\|_Y$, and the map $T : L_2(\mu_n) \to Y$ which is given by $Tf = 1/\sqrt{n}\sum_1^n f(x_i)x_i$ is an isometry. Clearly, $T(B_{X^*}/\mu_n) = E/\sqrt{n}$ and we shall denote this set by F. From Definition 3.2 it is easy to see that $T_p(F^o) = T_p(E^o)$ and that with respect to Y,

$$\ell(F) = \Big(\mathbb{E}\Big\|\sum_{i=1}^n g_i x_i\Big\|_{F^o}^2\Big)^{1/2} \leq T_p(E^o)\Big(\sum_{i=1}^n \|x_i\|_{F^o}^p\Big)^{1/p}. \tag{3.2}$$

But

$$\|x_i\|_{F^o} = \sup_{f\in F}\langle f, x_i\rangle = \sup_{\|x^*\|\leq 1/\sqrt{n}}\langle\sum_{j=1}^n x^*(x_j)x_j, x_i\rangle \leq \frac{1}{\sqrt{n}},$$

and therefore, $\ell(F) \leq T_p n^{1/p-1/2}$ in Y, implying that $\ell(B_{X^*}/\mu_n) \leq T_p n^{1/p-1/2}$ in $L_2(\mu_n)$.

■

Theorem 3.4 *Let X be a Banach space which has type $p > 1$ and let μ be a probability measure on B_X such that for every n,*

$$Pr\Big\{ s_n = (x_1, ..., x_n) | x_1, ..., x_n \text{ are linearly dependent} \Big\} = 0.$$

Then, B_{X^*} is a GC class on the set B_X with respect to μ. Moreover, there is some constant C which depends only on $T_p(X)$, such that for every $\varepsilon, \delta \in (0,1)$ and for $n \geq (C\varepsilon^4\delta)^{-p/(2p-2)}$,

$$\sup_{\mu} Pr\Big\{ \sup_{\|x^*\| \leq 1} |\mathbb{E}_\mu x^* - \mathbb{E}_{\mu_n} x^*| \geq \varepsilon \Big\} \leq \delta.$$

Proof: First, recall that by Lemma 3.1, it is enough to show that $\ell_n = \big(\mathbb{E}\ell^2(B_{X^*}/\mu_n)\big)^{1/2}$ is $o(\sqrt{n})$. By the assumption on the measure μ, a random sample is supported on a linearly independent set $\{x_1, ..., x_n\}$ almost surely. Thus, by Lemma 3.3, for an empirical measure supported on such a sample, $\ell(B_{X^*}/\mu_n) \leq T_p(X) n^{1/p - 1/2}$. Therefore, if $P > 1$ then $\ell_n = o(\sqrt{n})$.

To show the complexity estimate, note that by the proof of Lemma 3.1, there is an some absolute constant $C > 0$ such that

$$Pr\Big\{ \sup_{f \in \mathcal{F}} |\mathbb{E}_\mu f - \mathbb{E}_{\mu_n} f| \geq \varepsilon \Big\} \leq$$

$$\leq 8\Big(exp(-n\varepsilon^2/C) + C\frac{\ell_n^2}{n\varepsilon^4} \Big),$$

and our claim follows.

■

Remark 2 *The assertion of Theorem 3.4 remains true if instead of the unit balls B_X and B_{X^*} we take any two balls $rB_X \subset X$ and $RB_{X^*} \subset X^*$ for any $R, r > 0$. In this case, the constant in the complexity estimate $n(\varepsilon, \delta)$ depends on the radii of the two balls.*

Next, we apply Lemma 3.3 and obtain an upper estimate on the Gelfand numbers (defined below) of classes of functions which are subsets of the unit ball of certain Banach spaces.

Definition 3.5 *Let X and Y be Banach spaces and $u : X \to Y$. Denote by $c_k(u)$ the k-th Gelfand number of u, i.e.,*

$$c_k(u) = \inf\{ \big\| u_{|S} \big\| \ \ | S \subset X, \ codim \ S < k \}.$$

If F is a convex symmetric subset of ℓ_2^n which has a nonempty interior and if I is the formal identity $I : l_2^n \to F^*$, then $2c_k(I^*)$ is the smallest possible diameter in ℓ_2^n of a k-codimensional section of F. Denote this quantity by $c_k(F)$.

It turns out that the Gelfand numbers may be estimated in terms of the ℓ-norm. This important result is due to Pajor and Tomczak-Jaegermann (see [8],[9]).

Theorem 3.6 *There is a constant C such that for all Banach spaces X, for all $n \geq 1$ and for every $u : \ell_2^n \to X$*

$$\sup_{k \geq 1} k^{1/2} c_k(u^*) \leq C\ell(u)$$

Let X be a Banach space of functions on Ω. Assume that the point evaluation functionals δ_ω are uniformly bounded (i.e., there is some m such that $\sup_{\omega \in \Omega} \|\delta_\omega\|_{X^*} \leq m$) and that X^* has type 2. For example, X may be a Sobolev Hilbert space $W_0^{k,2}(\Omega)$ for an appropriate selection of k.

We shall use Lemma 3.3 to estimate the Gelfand numbers of $\mathcal{F} = B_X$ in $L_2(\mu)$ for any probability measure on Ω. We begin with the following lemma:

Lemma 3.7 *If $\mathcal{F}$ is a GC class of functions on Ω which are uniformly bounded, then $(\mathcal{F} - \mathcal{F})^2 = \{(f_1 - f_2)^2 | f_i \in \mathcal{F}\}$ is a GC class. Also, for every $\varepsilon > 0$ and every probability measure μ, there is an empirical measure μ_n such that for every $A \subset \mathcal{F}$, $diam_{L_2(\mu)}(A) \leq diam_{L_2(\mu_n)}(A) + \varepsilon$.*

Proof: Clearly, we may assume that all the elements of $\mathcal{F}$ are bounded by 1. First, we shall show that $(\mathcal{F}-\mathcal{F})^2$ is a GC class. Note that if μ_n is any empirical measure then $N\big(\varepsilon, (\mathcal{F}-\mathcal{F})^2, L_\infty(\mu_n)\big) \leq N^2\big(\varepsilon/8, \mathcal{F}, L_\infty(\mu_n)\big)$. Indeed, let $H = (h_i)$ be an ε-cover to $\mathcal{F}$ in L_∞. Since all the elements of $\mathcal{F}$ are bounded by 1 pointwise, we may assume that each h_i is bounded by 1. Thus, it is easy to see that for every $f_1, f_2 \in \mathcal{F}$, $\big|(f_1 - f_2)^2 - (h_1 - h_2)^2\big| \leq 4(|f_1 - h_1| + |f_2 - h_2|)$. Therefore, if $\|f_i - h_i\|_{L_\infty(\mu_n)} \leq \varepsilon$ then

$$\big\|(f_1 - f_2)^2 - (h_1 - h_2)^2\big\|_{L_\infty(\mu_n)} \leq 8\varepsilon$$

and $(H-H)^2$ is an 8ε-net to $(\mathcal{F}-\mathcal{F})^2$. Hence, for every $\varepsilon > 0$ and every integer n,

$$\sup_{\mu_n} \log N\big(\varepsilon, (\mathcal{F} - \mathcal{F})^2, L_\infty(\mu_n)\big) \leq 2 \sup_{\mu_n} \log N\big(\varepsilon/8, \mathcal{F}, L_\infty(\mu_n)\big),$$

which, by Theorem 2.7 is $o(n)$. Therefore, $(\mathcal{F} - \mathcal{F})^2$ is a GC class.

Turning to the second claim, let μ be a probability measure on Ω. Since $(\mathcal{F} - \mathcal{F})^2$ is a GC class, then for every $\varepsilon > 0$ there is some empirical measure μ_n such that

$$\sup_{f,g \in \mathcal{F}} \Big| \|f - g\|^2_{L_2(\mu_n)} - \|f - g\|^2_{L_2(\mu)} \Big| < \varepsilon.$$

Hence, if $A \subset \mathcal{F}$ then $diam^2_{L_2(\mu)}(A) \leq diam^2_{L_2(\mu_n)}(A) + \varepsilon$ and our claim follows. ∎

Corollary 3.8 *Let X be a Banach space of functions on Ω. Assume that the point evaluation functionals δ_ω are uniformly bounded by m and that X^* has type 2. For every probability measure μ on Ω, let $I : X \to L_2(\mu)$ be the inclusion operator and assume that for every $\omega \in \Omega$, $\mu(\{\omega\}) = 0$. Then, there is some constant $C = C(m)$ such that $c_k(I) \leq CT_2(X^*)k^{-1/2}$.*

We shall present a proof to this claim in the case where the point evaluation functionals are bounded by 1. The general case follow by the same argument.
Proof: First, note that for every probability measure μ on Ω, $X \subset L_2(\mu)$. Indeed, since $\sup_{\omega\in\Omega} \|\delta_\omega\|_{X^*} \leq 1$, then for every $x \in B_X$ and every $\omega \in \Omega$ $|x(\omega)| = |\delta_\omega(x)| \leq 1$. Thus, all the elements of B_X bounded by 1 pointwise.

Set $T : \Omega \to B_{X^*}$ by $T(\omega) = \delta_\omega$. For every measure μ on Ω let $\tilde{\mu}$ be a measure on B_{X^*}, given by $\tilde{\mu}(A) = \mu\big(T^{-1}(A)\big)$. Note that if μ_n is an empirical measure on Ω, then T induces an isometry between $L_2(\Omega, \mu_n)$ and $L_2(B_{X^*}, \tilde{\mu}_n)$, under which B_X/μ_n is mapped onto $B_{X^{**}}/\tilde{\mu}_n \subset L_2(B_{X^*}, \tilde{\mu}_n)$. Hence, $c_k(B_{X^{**}}/\tilde{\mu}_n) = c_k(B_X/\mu_n)$. Also, if μ_n is supported on a set of n distinct points (i.e., $\omega_i \neq \omega_j$ if $i \neq j$), then $\delta_{\omega_1}, ..., \delta_{\omega_n}$ are linearly independent. Since X^* has type 2 then by Lemma 3.3 $\ell(B_{X^{**}}/\tilde{\mu}_n) \leq T_2(X^*)$. Thus, by Theorem 3.6, $c_k(B_{X^{**}}/\tilde{\mu}_n) \leq CT_2(X^*)k^{-1/2}$, implying that $c_k(B_X/\mu_n) \leq CT_2(X^*)k^{-1/2}$ in $L_2(\mu_n)$.

Note that for every $\varepsilon > 0$, $\operatorname{ess\,sup}_{\mu_n} \log N\big(\varepsilon, B_X, L_1(\mu_n)\big) = o(n)$. Indeed, by the assumption on μ, almost every empirical measure μ_n is supported on a set of n distinct points. Thus, for almost every empirical selected according to μ,

$$\log N\big(\varepsilon, B_X/\mu_n, L_1(\mu_n)\big) \leq \log N\big(\varepsilon, B_{X^{**}}/\tilde{\mu_n}, L_1(\tilde{\mu}_n)\big) \leq \\ \leq \sup_{\nu_n} \log N\big(\varepsilon, B_{X^{**}}/\nu_n, L_1(\tilde{\nu}_n)\big),$$

where the supremum on the right hand side is taken with respect to all the empirical measures ν_n on B_{X^*} which are supported on a linearly independent set. By Theorem 3.4 the supremum is $o(n)$, therefore, by Theorem 2.6, B_X is a GC class of functions on Ω with respect to μ.

As noted above, for every $x \in B_X$, $\sup_{\omega\in\Omega} |x(\omega)| \leq 1$. Therefore, we may apply Lemma 3.7 to B_X as functions on Ω. Thus, for every $\varepsilon > 0$ and every probability measure μ on Ω, there is some empirical measure μ_n such that for every $A \subset B_X$, $diam_{L_2(\mu)}(A) \leq diam_{L_2(\mu_n)}(A) + \varepsilon$. Again, by the assumption on μ, μ_n may be chosen as an empirical measure supported on n distinct points. Let (z_i^*) be the k functionals on $L_2(\mu_n)$ such that $diam\big(\bigcap(ker(z_i^*) \cap B_X)\big) = c_k(B_X/\mu_n)$. Since the dual of $L_2(\mu_n)$ is spanned by empirical functionals (i.e. linear combinations of point evaluation functionals), then each z_i^* may be represented as $\sum_{i=1}^n a_{ij}\delta_{\omega_i}$. Also, since the point evaluation functionals belong to X^*, then $z_i^* \in X^*$. Hence, $H = \bigcap ker(z_i^*)$ is a k-codimensional subspace of X and

$$diam_{L_2(\mu)}(H \cap B_X) \leq diam_{L_2(\mu_n)}(H \cap B_X) + \varepsilon \leq CT_2(X^*)k^{-1/2} + \varepsilon.$$

Therefore, $c_k(I) \leq CT_2(X^*)k^{-1/2}$ as claimed.

■

We end this section with an estimate on the ℓ-norm of classes which have a finite VC dimension. As a starting point, we require an estimate on the L_2 covering numbers of such classes, which is due to Haussler (see [16]).

Theorem 3.9 *Let $\mathcal{F}$ be class of $\{0,1\}$ functions such that $VC(\mathcal{F}) = d$. Then, there is an absolute constant C such that for every probability measure μ on Ω,*

$$N(\varepsilon, \mathcal{F}, L_2(\mu)) \leq Cd(4e)^d\Big(\frac{1}{\varepsilon}\Big)^{2d}.$$

Theorem 3.10 *Let $\mathcal{F} \subset L_2(\mu)$ which consists of $\{0,1\}$ functions and assume that $VC(\mathcal{F}) = d$. Then, there is some absolute constant C such that $\ell(\mathcal{F}) \leq Cd^{1/2}$.*

Proof: Let H be a finite dimensional subspace of $L_2(\mu)$. Clearly, for every $0 < \varepsilon \leq 1$,

$$\log N\big(\varepsilon, \mathcal{F} \cap H, L_2(\mu)\big) \leq \log N\big(\varepsilon, \mathcal{F}, L_2(\mu)\big) \leq Cd \log \frac{2}{\varepsilon}.$$

If $\varepsilon > 1$ then $\{0\}$ is an ε-cover of $\mathcal{F}$, hence, for such ε, the log-covering numbers of $\mathcal{F}$ vanish. By Theorem 2.3,

$$\ell(F \cap H) \leq \int_0^1 \sqrt{Cd \log \frac{2}{\varepsilon}} d\varepsilon \leq Cd^{1/2}.$$

and our claim follows.

■

Corollary 3.11 *Let $\mathcal{F}$ be a class of $\{0,1\}$ functions on Ω. Then, $\mathcal{F}$ is a GC class with respect to all probability measures on Ω if and only if*

$$\sup_n \sup_{\mu_n} \ell(\mathcal{F}/\mu_n) < \infty.$$

Proof: Recall that a class of $\{0,1\}$ functions is GC with respect to all probability measures on Ω if and only if it has a finite VC dimension (see [4]). Thus, one direction of our claim follows from Theorem 3.10. The converse follows from Theorem 3.1.

■

4 Applications to Learning Theory

As stated in the introduction, the basic question in Learning Theory is that of generalization: given an unknown function $g : \Omega \to \mathbb{R}$ which is a member of a class $\mathcal{F}$ and an i.i.d. sample $\big\{(\omega_1, ..., \omega_n),\ (g(\omega_1), ..., g(\omega_n))\big\}$ selected according to an unknown probability measure μ on Ω, is it possible to construct some function f such that $\|f - g\|_{L_2(\mu)}$ is "small"? The law which assigns a function to each sample should be efficient, in the sense that for "most" of the samples, it should produce a function which is indeed close to g. We shall assume that for every $\omega \in \Omega$, $\mu(\{\omega\}) = 0$.

In this section, we formulate such a law using a "Gelfand like" estimate. The basis of our scheme is the following theorem:

Theorem 4.1 *Let $\mathcal{F}$ be a class of functions on Ω which are uniformly bounded by m. Set G to be the symmetric convex hull of $\mathcal{F}$ and assume that $\ell(\mathcal{F}/\mu_n)$ are uniformly bounded by ℓ. Then, there is some constant $C = C(m)$ such that for every $\varepsilon > 0$ there are k empirical linear functionals x_i^*, for which $diam_{L_2(\mu)}\big(\bigcap(ker(x_i^*) \cap G)\big) \leq \varepsilon^{1/2} + Ck^{-1/2}\ell$.*

Moreover, there is some constant $C' = C'(m)$ such that for every $\varepsilon > 0$, $0 < \delta \leq 1$ and

$$n \geq C' \max\Big\{ \frac{\ell^2}{\varepsilon^4 \delta}, \frac{\log \frac{1}{\delta}}{\varepsilon^2} \Big\} \tag{4.1}$$

there is a set S consisting of samples of n elements of Ω, $Pr(S) \geq 1 - \delta$, such that for every sample $(\omega_1, ..., \omega_n) \in S$ the functionals x_i^* can be chosen to be a linear combination of $\{\delta_{\omega_1}, ..., \delta_{\omega_n}\}$.

Proof: Since $\ell(\mathcal{F}/\mu_n) = \ell(G/\mu_n)$ then by Theorem 2.3 there is some constant $C > 0$ such that for every empirical measure μ_n, $\log N\big(\varepsilon, G, L_2(\mu_n)\big) \leq C\ell^2/\varepsilon^2$. Thus, by a similar argument to the one used in the proof of Lemma 3.7,

$$\sup_{\mu_n} \log N\big(\varepsilon, (G-G)^2, L_1(\mu_n)\big) \leq \sup_{\mu_n} \log N\big(\varepsilon/8, G, L_1(\mu_n)\big) \leq C\Big(\frac{\ell^2}{\varepsilon^2}\Big).$$

By the estimate in Theorem 2.6 applied to the class $(G-G)^2$, it follows that if $\varepsilon > 0$, $0 < \delta \leq 1$ and n is as in (4.1), there is a set S of samples $(\omega_1, ..., \omega_n)$ of probability larger than $1 - \delta$, such that for every empirical measure supported on $s_n \in S$, $\sup_{f,g \in G} \Big| \|f-g\|^2_{L_2(\mu)} - \|f-g\|^2_{L_2(\mu_n)} \Big| < \varepsilon$.

Let μ_n be any such empirical measure. Then, for any set $A \subset G$,

$$diam_{L_2(\mu)}(A) \leq diam_{L_2(\mu_n)} + \varepsilon^{1/2}.$$

By Theorem 3.6, there are k linear functionals (x_i^*) on $L_2(\mu_n)$, such that $diam_{L_2(\mu_n)} \bigcap (ker(x_i^*) \cap G) \leq C\ell k^{-1/2}$. Since μ_n is supported on $(\omega_1, ..., \omega_n)$ and since (δ_{ω_i}) is a basis to the dual space of $L_2(\mu_n)$, each functional x_i^* is a linear combination of δ_{ω_i} and our claim follows.

■

As an example of a case in which Theorem 4.1 may be applied is the case of VC classes. Indeed, if $VC(\mathcal{F}) = d$ there is some absolute constant C such that $\ell^2 \leq Cd$.

Using Theorem 4.1 it is possible to construct a learning algorithm which enables one to construct an approximation to the unknown function g, when the data at hand is $\big\{(\omega_1, ..., \omega_n),\ \big(g(\omega_1), ..., g(\omega_n)\big)\big\}$. Indeed, given $\varepsilon > 0$ and $0 < \delta \leq 1$, let S be as in Theorem 4.1. Then, with probability larger than $1 - \delta$, $(\omega_1, ..., \omega_n) \in S$ and the empirical functionals (x_i^*) may be chosen as a linear combination of (δ_{ω_i}). Since a part of the data received are the values $\big(g(\omega_i)\big)$, then for every $1 \leq i \leq k$, $x_i^*(g)$ may be calculated by the empirical data. Hence, if we can find some $f \in G$ such that $x_i^*(g) = x_i^*(f)$, then $\|f-g\|_{L_2(\mu)} \leq 2(\varepsilon^{1/2} + Ck^{-1/2}\ell)$.

In the following example we provide a learning scheme which is based on this idea.

Example: Let X be a Hilbert space of functions on Ω such that the point evaluation functionals δ_ω are uniformly bounded by 1 (e.g. a Sobolev Hilbert space $W_0^{k,2}(\Omega)$ for an appropriate value of k), and set $\mathcal{F}$ to be a subset of B_X. The idea behind the learning scheme is as follows: first, one has to construct a

section of $\mathcal{F}$ which has a small diameter using the empirical functionals (x_i^*). Second, due to the structure of X it is possible to identify a function f such that $x_i^*(f) = x_i^*(g)$.

To construct the functionals, let n be as in Theorem 4.1 and suppose that the sample $\omega_1, ..., \omega_n$ is selected randomly according to the measure μ. Just as in Corollary 3.8, for every empirical measure μ_n, $\ell(B_X/\mu_n) \leq T_2(X^*) = 1$. Fix $\varepsilon > 0$ and $0 < \delta \leq 1$, and let n be as in (4.1). By Theorem 4.1, there is some constant C and a set S such that $Pr(S) \geq 1 - \delta$ and for every sample $s_n \in S$ there are k-empirical functionals $x_i^* = \sum_1^n a_{ij}\delta_{\omega_j}$, for which $diam_{L_2(\mu)}\big(\bigcap(ker(x_i^*) \cap B_X)\big) \leq Ck^{-1/2} + \varepsilon^{1/2}$.

Turning to the second step, since $\delta_\omega \in B_{X^*}$, there is some $W \in B_X$ such that for every $f \in X$, $f(\omega) = \delta_\omega(f) = \langle f, W \rangle$. Given the sample $\{\omega_1, ..., \omega_n\}$, let W_i be the representation of the functional δ_{ω_i}, set $Y = span\{W_1, ..., W_n\}$ and let P_Y be the orthogonal projection onto Y. It is easy to see that for every $x \in X$, $x(\omega_i) = (P_Y x)(\omega_i)$. Moreover, since X is a Hilbert space, then for every $x \in X$, $\|P_Y x\| \leq \|x\|$. Thus, there is some $f \in B_X \cap Y$ such that $x_j^*(f) = x_j^*(g)$. To identify the desired f, note that if $f \in B_X \cap Y$, there are $\lambda_1, ..., \lambda_n$ such that $\sum_{i,j=1}^n \lambda_i \lambda_j \langle W_i, W_j \rangle \leq 1$ and $f = \sum_1^n \lambda_i W_i$. Therefore, an approximating function f will be any solution to the system of k equations with n variables $(\lambda_1, ..., \lambda_n)$

$$\sum_{i=1}^n a_{ij} g(\omega_i) = \sum_{k=1}^n \lambda_k \sum_{i=1}^n a_{ij} \langle W_i, W_k \rangle,$$

(which is simply $x_j^*(g) = x_j^*(f)$), subjected to the constraint that

$$\sum_{i,j=1}^n \lambda_i \lambda_j \langle W_i, W_j \rangle \leq 1.$$

Of course, $f = P_Y g$ is such a solution, but there are many other possible solutions to this system of equations.

From a practical point of view, the only problem in implementing this learning scheme is to determine the functionals (x_i^*). However, it follows from the proof of Theorem 3.6 that the functionals x_i^* may be selected randomly. They are the rows of the random linear operator $G : \mathbb{R}^n \to \mathbb{R}^k$, whose entries are i.i.d. Gaussian random variables. It turns out (see [9], [5]) that for every convex symmetric subset $K \subset \ell_2^n$, with high probability, $diam(ker(G) \cap K) \leq C\ell(K)k^{-1/2}$. Thus, the coefficients (a_{ij}) in the example above may be selected randomly.

Also, note that in many interesting cases, the function $u(\omega_1, \omega_2) = \langle W_1, W_2 \rangle$ can be calculated explicitly. Many such examples of so-called *Reproducing Kernels* may be found in [12].

To sum-up, the ability to estimate the ℓ-norm of classes of functions which are interesting from the perspective of Learning Theory is valuable, since it enables one to construct an approximation scheme based on the empirical data one receives.

The fact that the ℓ-norm provides an estimate on the diameter of an "optimal" k-codimensional section of the class in question has practical implications

thanks to two facts. First, because it is possible to find "almost optimal" sections using a simple random procedure, and second, because the empirical data tells us on what translation of the section the unknown function is located.

Thus, it is possible to use ℓ-norm estimates and construct a learning scheme just as in the example above.

5 Acknowledgments

I would like to thank Gil Luria for many hours of fruitful discussions, and to the anonymous referee for his valuable comments and suggestions.

References

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

[2] R.M. Dudley: The sizes of compact subsets of Hilbert space and continuity of Gaussian processes, J. of Functional Analysis 1, 290-330, 1967.

[3] R.M. Dudley: Universal Donsker classes and metric entropy, Annals of Probability 15, 1306-1326, 1987.

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

[5] Y. Gordon: On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$, *Geometric Aspects of Functional Analysis 1986–1987*, 84–106, Lecture Notes in Mathematics 1317, Springer-Verlag 1980

[6] J. Lindenstrauss, L. Tzafriri: *Classical Banach Spaces* Vol II, Springer Verlag, 1979

[7] S. Mendelson: Learnability in Banach Spaces with Reproducing Kernels, preprint

[8] A. Pajor, N. Tomczak-Jaegermann: Subspaces of small codimension of finite-dimensional Banach spaces, Proceedings of the AMS, 97 (4), 637-642, 1986.

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

[10] G. Pisier: Remarques sur un résultat non publié de B. Maurey, Séminaire d'analyse fonctionelle, 1980-1981, exposé No. 5, École Polytechnique, Palaiseau

[11] D. Pollard: *Convergence of Stochastic Processes*, Springer-Verlag 1984.

[12] S. Saitoh: *Integral Transforms, Reproducing Kernels and their applications*, Pitman research notes in Mathematics 369, Addison Wesley 1997.

[13] V.N. Sudakov: Gaussian processes and measures of solid angles in Hilbert space, Soviet Math. Dokl. 12, 412-415, 1971.

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

[15] V. Vapnik: *Statistical Learning Theory*, Wiley, 1998.

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