Learnability in Hilbert spaces with Reproducing Kernels

Shahar Mendelson
Institute of Computer Science
Hebrew University, Jerusalem 91904
Israel
E-mail shahar@tx.technion.ac.il

Abstract

We explore the question of learnability of classes of functions contained in a Hilbert space which has a reproducing kernel. We show that if the evaluation functionals are uniformly bounded and if the class is norm bounded then it is learnable. We formulate a learning procedure related to the well known Support Vector Machine, which requires solving a system of *linear* equations, rather than the quadratic programming needed for the SVM. As a part of our discussion, we estimate the fat shattering dimension of the unit ball of the dual of a Banach space when considered as a set of functions on the unit ball of the space itself. Our estimate is based on a geometric property of the Banach space called *type*.

In this paper we investigate the following question: assume that a function f is arbitrarily selected from a given class of functions $\mathcal{F}$ on $\Omega \subset \mathbb{R}^d$. We wish to identify this function using only its values on samples drawn according to an unknown probability measure on Ω. Since we can not hope for a complete identification of f using this partial information, we try to approximate f in the following sense: given a sample S, $S = \{\omega_1, ..., \omega_n\}$, $\{f(\omega_1), ..., f(\omega_n)\}$, we search for functions $g_n = g_n(S)$ such that (g_n) tends to f in some sense as we increase to size of the sample. Since the measure μ according to which the sample is selected is unknown and since f is unknown too, the convergence must be in the worst case scenario, i.e., it must be uniform both in f and in μ. Hence, our aim is to show that

$$\sup_{\mu} Pr\{\mathbb{E}_\mu (g_n - f)^2 \geq \varepsilon\} \to 0$$

uniformly in f, where $\mathbb{E}_\mu$ is the expectation with respect to μ.

For practical purposes, it is necessary to estimate the *sample complexity*, which is the size of the sample required to ensure that $\sup_\mu Pr\{\mathbb{E}_\mu (g_n - f)^2 \geq \varepsilon\}$ does not exceed a given $\delta > 0$.

Intuitively, the "smaller" $\mathcal{F}$ is, the easier it is to find the desired f. One property which can be interpreted to mean that $\mathcal{F}$ is "small" is that it satisfies the law of large numbers uniformly in both f and in μ. Classes of functions with this property are called *uniform Glivenko–Cantelli classes* ([4], [15]).

It is possible to show (see [2]) that if one can define a learning rule which produces a sequence of functions (g_n) that "almost" agree with f on a given sequence of samples, and if $\mathcal{F}$ is uniform Glivenko–Cantelli, then (g_n) approximates f in the sense that $\sup_\mu Pr\{\mathbb{E}_\mu (g_n - f)^2 \geq \varepsilon\} \to 0$.
We shall advance in two directions: one is to show that the classes we are interested in are uniform Glivenko–Cantelli classes, and the other is to find a learning rule which produces a function that "almost" agrees with f on a given sample.

To show that a class is uniform Glivenko–Cantelli, we use the scale sensitive dimensions of $\mathcal{F}$: the *fat shattering dimension* $VC_\varepsilon(\Omega, \mathcal{F})$ and the *Pollard dimension* $P_\varepsilon(\Omega, \mathcal{F})$ (see [2], [10]), which indicate how "large" $\mathcal{F}$ is. Indeed, it was shown in [2] that if $\mathcal{F}$ consists of functions with a uniformly bounded range, then $\mathcal{F}$ is uniform Glivenko–Cantelli if and only if $VC_\varepsilon(\Omega, \mathcal{F})$ (resp. $P_\varepsilon(\Omega, \mathcal{F})$) is finite for every $\varepsilon > 0$.

We focus on classes of functions on some set Ω which are contained in the unit ball of a Banach space in which the evaluation functionals δ_ω are uniformly bounded. We give an upper estimate for $VC_\varepsilon(\Omega, \mathcal{F})$ which depends on a geometric property of the Banach space X called *type*. This upper estimate is obtained by using a tight bound for $VC_\varepsilon(B(X), B(X^*))$, where $B(X)$ is the unit ball of X and $B(X^*)$ is the unit ball of the dual of X.

The first section of this paper is devoted to the proof of the bound on $VC_\varepsilon(\Omega, \mathcal{F})$. In the second section we narrow the discussion to classes of functions contained in Hilbert spaces with reproducing kernels and introduce a learning process which enables us to find functions g_n for which $\sup_\mu Pr\{\mathbb{E}_\mu (g_n - f)^2 \geq$

$\varepsilon\} \to 0$ uniformly in f. The idea behind this learning rule is to embed the samples in a high dimensional space and find functionals which agree with f on the given samples. This idea is similar to the one used in the *Support Vector Machine* ([16]) process. Due to the fact that the space in question has a reproducing kernel, it is possible to embed the samples in the dual of the given space, and the desired functionals on the dual will be members of our class. Moreover, in order to find each functional, one only needs to solve a *linear* system of equations, which is much simpler that the quadratic programming needed for the SVM.

We end the second section by showing that under additional mild assumptions on the space X, we can estimate the sample complexity of the learning process. The third section consists of concluding remarks. In it, we discuss a different and more direct approach to the problem of evaluating the fat shattering dimension and compare it to the one presented in the first section. We show that in many important cases the "soft" approach presented in the first section gives a much better bound than the direct one.

We end this introduction by recalling a few standard definitions and some notation. For a Banach space X, the *dual* of X (denoted by X^*) consists of all the bounded linear functionals on X, with the norm $\|x^*\|_{X^*} = \sup_{\|x\|_X=1} |x^*(x)|$. We denote by $B(X)$ the unit ball of X, i.e., $B(X) = \{x| \, \|x\| \leq 1\}$. For every $r > 0$ $rB(X) = \{x| \, \|x\| \leq r\}$. For every $A \subset X$, let $\operatorname{conv} A$ be the convex hull of A. A Banach space is called *reflexive* if X is isometric to X^{**} via the *duality map* $x \to x^{**}$ given by $x^{**}(x^*) = x^*(x)$. If $1 \leq p < \infty$, let ℓ_p^n be $\mathbb{R}^n$ equipped with the norm $\|x\|_p = \left(\sum_1^n |x_i|^p\right)^{\frac{1}{p}}$. Throughout this paper Ω will denote a compact subset of $\mathbb{R}^d$. $C(\Omega)$ is the Banach space of continuous functions on Ω, with respect to the norm $\|f\|_\infty = \sup\limits_{\omega\in\Omega} |f(\omega)|$. 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. Finally, if x is some point in some metric space, then $B(x, r)$ is the open ball centered at x with radius r.

1 Fat Shattering dimension and Type

In this section we investigate the fat shattering dimension (defined below) of $\mathcal{F} = B(X^*)$ which is the unit ball of the dual of a Banach space X, where the elements of $\mathcal{F}$ are viewed as functions on $B(X)$.

Definition 1.1 *Let $\mathcal{F}$ be a class of functions on a space Ω. We say that $\mathcal{F}$ ε–shatters $\omega_1, ..., \omega_n$ if there is some $a \in \mathbb{R}$ such that for every $I \subset \{1, ..., n\}$ there is a function $f \in \mathcal{F}$ for which $f(\omega_i) \geq a + \varepsilon/2$ if $i \in I$ and $f(\omega_j) \leq a - \varepsilon/2$ if $j \notin I$. Let $VC_\varepsilon(\Omega, \mathcal{F})$ be the largest integer N such that there exists a set of N elements of Ω which is ε–shattered by $\mathcal{F}$. We set $VC_\varepsilon(\Omega, \mathcal{F}) = \infty$ if there exist such integers N which are arbitrarily large. The fat shattering co-dimension $COVC_\varepsilon(\Omega, \mathcal{F})$ is defined to be $VC_\varepsilon(\mathcal{F}, \Omega)$ in the sense that the "base" space is $\mathcal{F}$ and each $\omega \in \Omega$ is identified with the evaluation functional δ_ω, i.e., each ω is*

identified with a function on $\mathcal{F}$ defined by $\omega(f) \equiv f(\omega)$.

First, we study the case where $\Omega = B(X)$ and $\mathcal{F} = B(X^*)$. Thus, the set $\{x_1, ..., x_n\}$ is ε–shattered if for every $I \subset \{1, ..., n\}$ there exists $x^* \in X^*$ with $\|x^*\| \leq 1$ such that $x^*(x_i) \geq a + \varepsilon/2$ if $i \in I$, while $x^*(x_j) \leq a - \varepsilon/2$ otherwise.

Definition 1.2 *A set $A = \{\omega_1, ..., \omega_n\}$ is said to be ε–shattered in the Pollard sense by $\mathcal{F}$ if there is some function $s : A \to \mathbb{R}$, such that for every $I \subset \{1, ..., n\}$ there is some $f \in \mathcal{F}$ for which $f(\omega_i) \geq s(\omega_i) + \varepsilon/2$ if $i \in I$, and $f(\omega_j) \leq s(\omega_j) - \varepsilon/2$ if $j \notin I$. We define the Pollard dimension $P_\varepsilon(\Omega, \mathcal{F})$ as the largest integer N such that there exists a set of N elements of Ω which is ε–shattered in the Pollard sense. Again, $P_\varepsilon(\Omega, \mathcal{F}) = \infty$ if such integers N can be arbitrarily large.*

Throughout this paper we will be investigating classes of functions which have a uniformly bounded range. Therefore, we assume that there is some M such that for every $\omega \in \Omega$, $\sup_{f \in \mathcal{F}} |f(\omega)| \leq M$. By the pigeonhole principle it is easy to see that the Pollard dimension and the fat shattering dimension are related for classes of functions which have a uniformly bounded range. In this case, for every $\gamma > 0$ and if $\sup_{\omega \in \Omega} \sup_{f \in \mathcal{F}} |f(\omega)| \leq M$, then

$$VC_\gamma(\Omega, \mathcal{F}) \leq P_\gamma(\Omega, \mathcal{F}) \leq C \frac{VC_{\frac{\gamma}{2}}(\Omega, \mathcal{F})}{\gamma}, \tag{1.1}$$

where C depends only on M.

Recall that for every integer k, the k-th Rademacher random variable $r_k(t)$ is defined on the interval $[0, 1]$ by $r_k(t) = sign\big(\sin(2^k \pi t)\big)$. Thus, (r_k) are independent $\{-1, 1\}$–valued functions and for every k

$$|\{t | r_k(t) = 1\}| = |\{t | r_k(t) = -1\}| = \frac{1}{2},$$

where $|\ |$ is the Lebesgue measure on $[0, 1]$.

Definition 1.3 *A Banach space X has type p, if there is some C such that for every $x_1, ..., x_n \in X$,*

$$\mathbb{E} \left\| \sum_1^n r_i(t) x_i \right\| \leq C \Big(\sum_1^n \|x_i\|^p \Big)^{1/p} \tag{1.2}$$

where $r_i(t)$ are i.i.d. Rademacher random variables on $[0, 1]$. The best constant for which (1.2) holds is called the p-type constant of X and denoted by $T_p(X)$.

The basic facts concerning the concept of type may be found, for example, in [11] or in [12]. Clearly, for every Banach space (1.2) holds in the case $p = 1$ with $T_1(X) = 1$. If we set

$$T(X) = \sup\{p | X \text{ has type } p\}$$

then it follows that $T(X) \in [1,2]$. For example, Hilbert spaces and L_p spaces for $2 \leq p < \infty$ have type 2, thus, if X is a Hilbert space or if $X = L_p$ for $2 \leq p < \infty$ then $T(X) = 2$ and the supremum is attained. Also, one can show that X has a nontrivial type (i.e. $p > 1$) if and only if X^* has a nontrivial type.

Definition 1.4 *The Banach–Mazur distance between two isomorphic Banach spaces X and Y, denoted by $d(X,Y)$, is given by:*

$$d(X,Y) = inf\{\|T\| \left\|T^{-1}\right\| \, | T : X \to Y \text{ is an isomorphism from } X \text{ to } Y\}$$

Clearly, if X, Y and Z are isomorphic then $d(X,Y) \leq d(X,Z)d(Y,Z)$.
We say that an infinite dimensional space X contains ℓ_p^n λ-uniformly, if for every n, X has an n–dimensional subspace X_n such that $d(X_n, \ell_p^n) \leq (1+\lambda)$. Note that $T(X) = 1$ if and only if X contains ℓ_1^n λ-uniformly for every $\lambda > 0$ (see [12]).

Theorem 1.5 *For every infinite dimensional Banach space X, the fat shattering dimension $VC_\varepsilon\big(B(X), B(X^*)\big)$ is finite if and only if $T(X) > 1$. If $T(X) = p$ and if X has type p', then for every $\varepsilon > 0$,*

$$\left(\frac{2}{\varepsilon}\right)^{\frac{p}{p-1}} - 1 \leq VC_\varepsilon\big(B(X), B(X^*)\big) \leq 2\left(\frac{2T_{p'}(X)}{\varepsilon}\right)^{\frac{p'}{p'-1}} + 1.$$

In particular, if X has type $p = T(X)$ then for every $\varepsilon > 0$,

$$\left(\frac{2}{\varepsilon}\right)^{\frac{p}{p-1}} - 1 \leq VC_\varepsilon\big(B(X), B(X^*)\big) \leq 2\left(\frac{2T_{p}(X)}{\varepsilon}\right)^{\frac{p}{p-1}} + 1.$$

The idea of connecting $VC_\varepsilon\big(B(X), B(X^*)\big)$ with the type of X first appeared in [8], where it was shown that if X has type p, then $VC_\varepsilon\big(B(X), B(X^*)\big) = O\big(\varepsilon^{-p/(p-1)}\big)$ – without an estimate on the constant. Our proof of the upper bound is slightly simpler than the proof in [8], and bypasses a gap in the original proof.

Proof: Let $(e_i)_{i=1}^n$ denote the standard basis in ℓ_1^n. We claim that $\{e_1, ..., e_n\}$ is 2–shattered by $B(\ell_1^{n\,*})$. To see this, let $I \subset \{1, ..., n\}$ and set $y_I^* \in (\ell_1^n)^*$ by $y_I^*(e_i) = 1$ if $i \in I$, and $y_I^*(e_j) = -1$ if $j \notin I$. By the definition of the dual norm

$$\|y_I^*\| = \sup_{\sum_1^n |\lambda_i| = 1} \left| y_I^*\big(\sum_{i=1}^n \lambda_i e_i\big) \right| \leq \sup_{\sum_1^n |\lambda_i| = 1} \sum_{i=1}^n \lambda_i \, |y_I^*(e_i)| \leq 1$$

implying that indeed $\{e_1, ..., e_n\}$ is 2–shattered by $B(\ell_1^{n\,*})$.

Next, if $T(X) = p > 1$ then for every $\lambda > 0$ and every integer n, there is a subspace $X_n \subset X$ such that $\dim X_n = n$ and $d(\ell_p^n, X_n) \leq 1 + \lambda$ (see [12]). Also, recall that $d(\ell_1^n, \ell_p^n) = n^{1-\frac{1}{p}}$ (see [14]), hence, $d(X_n, \ell_1^n) \leq (1+\lambda)n^{1-\frac{1}{p}}$. Set $T_n : \ell_1^n \to X_n$ such that T_n is an isomorphism, $\|T_n\| = 1$ and $\left\|T_n^{-1}\right\| \leq (1+\lambda)n^{1-\frac{1}{p}}$.

We will show that the set $\{x_1,...,x_n\}$ where $x_i = T_n e_i$ is $2(1+\lambda)^{-1}/n^{\frac{p-1}{p}}$ shattered by $B(X^*)$, implying that

$$VC_\varepsilon(B(X),B(X^*)) \geq \left(\frac{2}{\varepsilon}\right)^{\frac{p}{p-1}} - 1.$$

Indeed, if $I \subset \{1,...,n\}$, put $x_I^* = \frac{y_I^*(T_n^{-1})}{(1+\lambda)n^{1-\frac{1}{p}}} \in X_n^*$. Clearly, if $i \in I$ then $x_I^*(x_i) = \frac{1}{(1+\lambda)n^{1-\frac{1}{p}}}$ and if $j \notin I$ then $x_I^*(x_j) = \frac{-1}{(1+\lambda)n^{1-\frac{1}{p}}}$. Since $\|x_I^*\|_{X_n^*} \leq 1$, then by the Hahn-Banach theorem x_I^* may be extended to an element of $B(X^*)$. Our claim follows by taking λ to 0.

Turning to the reverse inequality, if $VC_\varepsilon\big(B(X),B(X^*)\big) = m$, set $n = [m/2]$. Thus, there exists a set $\{x_1,...x_{2n}\} \subset B(X)$ which is ε-shattered by $B(X^*)$, implying that for every $I \subset \{1,...,2n\}$ there is some $x^* \in X^*$ with $\|x^*\| \leq 1$ such that for $i \in I$, $x^*(x_i) \geq a + \varepsilon/2$ and for $j \notin I$, $x^*(x_j) \leq a - \varepsilon/2$. Set $A_I = \mathrm{conv}\{x_i \mid i \in I\}$ and $B_I = \mathrm{conv}\{x_j \mid j \notin I\}$. Note that if $y \in A_I$ and $z \in B_I$, then writing $y = \sum_{i\in I} \lambda_i x_i$ and $z = \sum_{i\notin I} \mu_i x_i$, we have

$$\|z-y\| \geq x^*(y-z) = \sum_{i\in I} \lambda_i x^*(x_i) - \sum_{j\notin I} \mu_j x^*(x_j) \geq a + \varepsilon/2 - a + \varepsilon/2 = \varepsilon.$$

For every $1 \leq i \leq n$, put $y_i = x_{2i} - x_{2i-1}$. If X has type p' then by the type estimate and since $\|y_i\| \leq 2$, then

$$\mathbb{E}\left\|\sum_1^n r_i(t) y_i\right\| \leq T_{p'}(X)\big(\sum_1^n \|y_i\|^{p'}\big)^{1/p'} \leq 2T_{p'}(X) n^{1/p'}.$$

Thus, for some choice of the numbers $\varepsilon_i \in \{+1,-1\}$ we have $\|\sum_1^n \varepsilon_i y_i\| \leq 2T_{p'}(X)n^{1/p'}$. On the other hand, there exists a set $J \subset \{1,...,2n\}$ such that $|J| = n$ and $\sum_{i=1}^n \varepsilon_i y_i = \sum_{j\in J} x_j - \sum_{j\notin J} x_j$. Hence, $\frac{1}{n}\sum_{j\in J} x_j \in A_J$ and $\frac{1}{n}\sum_{j\notin J} x_j \in B_J$, from which it follows that $\frac{1}{n}\left\|\sum_1^n \varepsilon_i y_i\right\| \geq \varepsilon$. Therefore $n\varepsilon \leq 2T_{p'}(X)n^{1/p'}$, implying that

$$VC_\varepsilon\big(B(X),B(X^*)\big) \leq 2n+1 \leq 2\left(\frac{2T_{p'}(X)}{\varepsilon}\right)^{\frac{p'}{p'-1}} + 1.$$

Finally, recall that if $T(X) = 1$ then X contains ℓ_1^n λ-uniformly for every $\lambda > 0$. Fix $\lambda > 0$ and let $T_n : \ell_1^n \to X$ such that T_n is an isomorphism, $\|T_n\| = 1$ and $\left\|T_n^{-1}\right\| \leq 1+\lambda$. Since the standard basis in ℓ_1^n is 2–shattered, then by the same argument as above, the set $\{T_n e_1,...,T_n e_n\}$ is $\frac{1}{1+\lambda}$–shattered in X by $\mathcal{F} = B(X^*)$. Hence, for every $\varepsilon < 2$, $VC_\varepsilon\big(B(X),B(X^*)\big) = \infty$. ■

Note that X may not have type p for $p = T(X)$, but does have type p' for every $1 \leq p' < T(X)$. Since in the proof of the upper bound one uses (1.2), then in the general case this bound can be established only for $p' < T(X)$.

Corollary 1.6 *If X is an infinite dimensional Hilbert space then*

$$\frac{4}{\varepsilon^2} - 1 \leq VC_\varepsilon\big(B(X), B(X^*)\big) = COVC_\varepsilon\big(B(X), B(X^*)\big) \leq \frac{8}{\varepsilon^2} + 1.$$

The proof of Corollary 1.6 follows since a Hilbert space is reflexive, implying that

$$VC_\varepsilon\big(B(X), B(X^*)\big) = COVC_\varepsilon\big(B(X), B(X^*)\big).$$

Also, note that X has type 2 with $T_2(X) = 1$. Hence, our claim follows by Theorem 1.5.

Corollary 1.7 *If X is an infinite dimensional Banach space and $T(X) \neq T(X^*)$ then $VC_\varepsilon(B(X), B(X^*))$ and $COVC_\varepsilon(B(X), B(X^*))$ are not of the same order of magnitude.*

Before proving this Corollary, we need two preliminary results. First, note that by linearity, for every Banach space X and every $r_1, r_2 > 0$,

$$VC_\varepsilon\big(r_1 B(X), r_2 B(X^*)\big) = VC_{\varepsilon/r_1 r_2}\big(B(X), B(X^*)\big),$$

and

$$COVC_\varepsilon\big(r_1 B(X), r_2 B(X^*)\big) = COVC_{\varepsilon/r_1 r_2}\big(B(X), B(X^*)\big).$$

Second, is a classical result from Banach space theory, which is called the *principle of local reflexivity* (see [9]).

Lemma 1.8 *Let X be a Banach space, which is identified with its canonical image in X^{**}. For every finite dimensional subspaces $G \subset X^{**}$ and $F \subset X^*$ and every $\delta > 0$, there is a map $T : G \to X$ such that*

1. *For every $x^{**} \in G \cap X$, $Tx^{**} = x^{**}$.*
2. *For every $x^{**} \in G$, $(1-\delta)\,\|x^{**}\| \leq \|Tx^{**}\| \leq (1+\delta)\,\|x^{**}\|$.*
3. *For every $x^* \in F$ and every $x^{**} \in G$, $x^*(Tx^{**}) = x^{**}(x^*)$.*

Proof of Corollary 1.7: We will show that for every $\varepsilon > 0$,

$$\begin{aligned} COVC_\varepsilon\big(B(X), B(X^*)\big) &\leq VC_\varepsilon\big(B(X^*), B(X^{**})\big) \leq \\ &\leq \liminf_{a \to 0^+} COVC_{\varepsilon - a}\big(B(X), B(X^*)\big) + 1. \end{aligned}$$

Since X is isometrically embedded in X^{**}, then for every $\varepsilon > 0$,

$$COVC_\varepsilon\big(B(X), B(X^*)\big) \leq VC_\varepsilon\big(B(X^*), B(X^{**})\big).$$

To prove the reverse inequality, set $\varepsilon > 0$ and let $\{x_1^*, ..., x_n^*\} \subset B(X^*)$ be ε-shattered by $B(X^{**})$. Thus, for every $I \subset \{1, ..., n\}$ there is a "shattering"

functional $x_I^{**} \in B(X^{**})$. Let $G = \text{span}\{x_I^{**} | I \subset \{1,...,n\}\}$ and set $F = \text{span}\{x_1^*,...,x_n^*\}$. For every $\delta > 0$ let $T : G \to X$ be as in Lemma 1.8. Hence, $\{x_1^*,...,x_n^*\}$ are ε-shattered by the set $\{Tx_I^{**}\} \subset (1+\delta)B(X)$. Therefore,

$$VC_\varepsilon\big(B(X^*), B(X^{**})\big) \leq COVC_\varepsilon\big((1+\delta)B(X), B(X^*)\big) \leq$$
$$\leq COVC_{\varepsilon/(1+\delta)}\big(B(X), B(X^*)\big).$$

Our assertion follows by taking $\delta \to 0$.

Finally, by Theorem 1.5 and since $T(X) \neq T(X^*)$, $VC_\varepsilon\big(B(X), B(X^*)\big)$ and $COVC_\varepsilon\big(B(X), B(X^*)\big)$ are not of the same order of magnitude.

■

Theorem 1.9 *Let Ω be a compact subset of $\mathbb{R}^d$ and suppose that X is a Banach space whose elements are Borel measurable functions on Ω. Assume that for every $\omega \in \Omega$ the evaluation functional $\delta_\omega(f) = f(\omega)$ is continuous and that $\sup_{\omega\in\Omega} \|\delta_\omega\| \leq M < \infty$. Assume further that X^* has type $p > 1$. Then*

$$VC_\varepsilon(\Omega, \mathcal{F}) \leq 2\left(\frac{2MT_p(X^*)}{\varepsilon}\right)^{\frac{p}{p-1}} + 1$$

for every $\mathcal{F} \subset B(X)$.

Proof: Let us fix some $\mathcal{F} \subset B(X)$. Note that $\mathcal{F}$ is isometrically embedded into $B(X^{**})$ using the duality mapping $f \to f^{**}$ defined by $f^{**}(f^*) = f^*(f)$. Denote the image of $\mathcal{F}$ in $B(X^{**})$ by $\mathcal{F}^{**}$. Clearly $f(\omega) = f^{**}(\delta_\omega)$ for every $\omega \in \Omega$ and $f \in \mathcal{F}$. Hence, if $\{\omega_1,...,\omega_n\}$ is ε–shattered by $\mathcal{F}$ then the set $\{\delta_{\omega_1},...,\delta_{\omega_n}\}$ is ε–shattered by $\mathcal{F}^{**}$. Since $\sup_\omega \|\delta_\omega\| \leq M$ then

$$VC_\varepsilon(\Omega, \mathcal{F}) \leq VC_\varepsilon\big(MB(X^*), \mathcal{F}^{**}\big) \leq VC_\varepsilon\big(MB(X^*), B(X^{**})\big) =$$
$$= VC_{\varepsilon/M}\big(B(X^*), B(X^{**})\big) \leq 2\left(\frac{2MT_p(X^*)}{\varepsilon}\right)^{\frac{p}{p-1}} + 1,$$

where the last inequality follows from the upper bound in Theorem 1.5.

■

2 Hilbert spaces with reproducing Kernels and Learning Procedures

We begin this section with the definition of *learnability*. For every integer n, let S_n be the set of all the samples $\{\omega_1,...,\omega_n\}$, $\{f(\omega_1),...f(\omega_n)\}$ of length n, where $\omega_i \in \Omega$ and $f \in \mathcal{F}$. A *learning procedure* is a mapping A which assigns a function in $\mathcal{F}$, denoted by A_S, to each sample $S \in \bigcup_n S_n$.

Our goal in a learning process is to approximate an unknown function $f \in \mathcal{F}$ with respect to the $L_2(\mu)$ norm. Recall that we denote by $\mathbb{E}_\mu$ the expectation

with respect to μ. Thus, by the definition of the $L_2(\mu)$ norm, we need to find a sequence (g_n) for which $\|f-g_n\|^2_{L_2(\mu)} = \mathbb{E}_\mu(f-g_n)^2 \to 0$. The functions (g_n) will be determined by the samples $\{\omega_1,...,\omega_n\}$, $\{f(\omega_1),...,f(\omega_n)\}$ selected according to the measure μ. Hence, a learning process is useful if it can find (with high probability with respect to the induced measure on the samples) an "almost" minimizer to $\mathbb{E}_\mu(f-h)^2$ using data derived from the samples. Having this in mind, for every $h \in \mathcal{F}$, let $\mathcal{L}_h = (h(x)-f(x))^2$ be the loss function associated with $h \in \mathcal{F}$. Given a Borel probability measure μ on Ω, denote by Pr the product measure μ^∞, on the product space $\mathcal{S} = \Omega^\infty$.

Definition 2.1 *We say that $\mathcal{F}$ is learnable if there is a learning procedure A such that for every $\varepsilon > 0$*

$$\lim_{n\to\infty} \sup_{\mu} Pr\{\mathbb{E}_\mu(\mathcal{L}_{A_{S_n}}) > \varepsilon\} = 0.$$

Note that if a class is learnable then it is possible to approximate its members by a sequence $(g_n) \subset \mathcal{F}$. We will be equally interested in cases where the approximating sequence (g_n) is not necessarily contained in $\mathcal{F}$.

One of the main goals of this section is to introduce an approximating procedure which, for every given sample, produces an element of X which approximates f on the sample. This learning procedure is based on the properties of Hilbert spaces with reproducing kernels. The setup we focus on is as follows: put Ω to be a compact subset of $\mathbb{R}^d$ and let X be a Hilbert space which consists of Borel functions on Ω with respect to an inner product denoted by $\langle -,- \rangle$. Assume further that the evaluation functionals δ_ω are continuous and uniformly bounded, i.e., that there is some M such that $\sup_\omega \|\delta_\omega\| \leq M$.

By the Riesz representation theorem, for every $\omega \in \Omega$ there is $W_\omega \in X$ such that $\|W_\omega\| = \|\delta_\omega\|$ and for every $f \in X$, $\langle f, W_\omega \rangle = f(\omega)$.

Definition 2.2 *A Hilbert space X which consists of functions on Ω is said to have a reproducing kernel $K : \Omega \times \Omega \to \mathbb{R}$ if for every $\omega \in \Omega$ and $f \in X$, $\langle K(\omega,t), f(t) \rangle = f(\omega)$.*

Note that by the uniqueness of the representation in Riesz's theorem and since $f(\omega) = \langle K(\omega,-), f(-) \rangle = \langle W_\omega(-), f(-) \rangle$ it follows that $K(\omega,-) = W_\omega(-)$.

For examples of Hilbert spaces in which the reproducing kernel has an explicit representation, we refer the reader to [13]. One well known example of a Hilbert space with a reproducing kernel is the following:

Example – Sobolev Spaces: Let Ω be a compact subset of $\mathbb{R}^d$. Given a smooth function $f : \Omega \to \mathbb{R}$ and a multiindex α, denote by $D^\alpha f$ the weak derivative of f with respect to α. Let $W_0^{k,2}(\Omega)$ be the space of all the functions such that for every multi index α, with $|\alpha| = k$, $D^\alpha f \in L_2(\Omega)$ and f vanishes on $\partial\Omega$. (see [7] for the basic facts regarding Sobolev spaces, or [1] for a more detailed survey).

There are two equivalent norms on this space:

$$\|f\|^2 = \sum_{|\alpha|=k} \|D^\alpha f\|_{L_2}^2 \tag{2.1}$$

and

$$\|f\|^2 = \sum_{|\alpha|\le k} \|D^\alpha f\|_{L_2}^2 . \tag{2.2}$$

Under both these norms, $W_0^{k,2}$ is a separable Hilbert space with the inner product

$$\langle f, g\rangle = \sum_{|\alpha|=k} \langle D^\alpha f, D^\alpha g\rangle_{L_2(\Omega)}$$

which induces (2.1), and

$$\langle f, g\rangle = \sum_{|\alpha|\le k} \langle D^\alpha f, D^\alpha g\rangle_{L_2(\Omega)}$$

which induces (2.2). By the Sobolev inequalities ([1], [7]) it follows that if $k > \frac{d}{2}$ then X is a closed subspace of $C(\Omega)$. In particular, the evaluation functionals δ_ω are continuous functionals and uniformly bounded.

In many cases one can find an explicit representation for the reproducing kernel. For example, let $X = W_0^{1,2}[0,1]$. It can be shown that X is the space of the absolutely continuous functions on $[0,1]$ such that $f(0) = f(1) = 0$. Also, X is continuously embedded in $C(0,1)$ and its reproducing kernel with respect to the norm (2.2) is

$$K(x,y) = \begin{cases} \frac{1}{2(e^2-1)}(e^x + e^{-x})(e^y + e^{2-y}) & \text{if } 0 \le x \le y, \\ \frac{1}{2(e^2-1)}(e^y + e^{-y})(e^x + e^{2-x}) & \text{if } y \le x \le 1. \end{cases}$$

In the general case, even when one does not have an explicit representation for the reproducing kernel, one can approximate it using the following computation.

Let (u_n) be a complete orthonormal basis of X, then

$$f(\omega) = \langle K(\omega,-), f(-)\rangle = \sum_{n=1}^{\infty} \langle f, u_n\rangle \langle K(\omega,-), u_n(-)\rangle = \sum_{n=1}^{\infty} \langle f, u_n\rangle u_n(\omega).$$

Therefore, for $f = K(-,\omega_2)$ and since K is symmetric,

$$K(\omega_1,\omega_2) = \sum_{n=1}^{\infty} u_n(\omega_1)u_n(\omega_2).$$

Hence, for every ω_1, ω_2

$$\langle W_{\omega_1}, W_{\omega_2}\rangle = \sum_{n=1}^{\infty} u_n(\omega_1)u_n(\omega_2). \tag{2.3}$$

Recall that for every Borel probability measure μ on Ω, Pr is the product measure μ^∞ on the product space $\mathcal{S} = \Omega^\infty$. If $\vec{\omega} \in \mathcal{S}$, let μ_m be the empirical measure supported on the first m coordinates of $\vec{\omega}$. For $f : \Omega \to \mathbb{R}$, $\mathbb{E}_{\mu_m}(f)$ denotes the empirical mean of f, i.e.,

$$\mathbb{E}_{\mu_m}(f) = \frac{1}{m}\sum_{i=1}^{m} f(\omega_i).$$

Definition 2.3 *A class of functions $\mathcal{F}$ satisfies the ε uniform Glivenko–Cantelli condition on Ω if*

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

To avoid the measurability problems that might be caused by the supremum, one usually uses an outer measure in the definition of a uniform Glivenko–Cantelli class (see [6]). Actually, only a rather weak assumption (called "image admissibility Suslin") is needed to avoid the above mentioned measurability problem. (see [5] for more details).

It was shown in [2] that a class of bounded functions $\mathcal{F}$ is ε uniform Glivenko–Cantelli for every $\varepsilon > 0$ if and only if $VC_\varepsilon(\Omega, \mathcal{F})$ is finite for every $\varepsilon > 0$. Also, by [3], if for some $\tau > 0$, $k = P_{(1/4-\tau)\varepsilon}$ is finite, then

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

for

$$n = O\Big(\frac{1}{\varepsilon^2}\Big(k\log^2\frac{1}{\varepsilon} + \log\frac{1}{\delta}\Big)\Big). \tag{2.4}$$

Theorem 2.4 *Let $\Omega \subset \mathbb{R}^d$ be a compact set, and let X be a Hilbert space of Borel functions on Ω such that the evaluation functionals are uniformly bounded. Then $\mathcal{F} = B(X)$ is learnable as a class of functions on Ω.*

Proof: First, note that by Theorem 1.9 $\mathcal{F}$ is ε uniform Glivenko–Cantelli for every $\varepsilon > 0$. Fix $f \in \mathcal{F}$ and a sample $\{\omega_1, ..., \omega_m\}$, $\{f(\omega_1), ..., f(\omega_m)\}$. By re-ordering the sample, select $\{\omega_1, ..., \omega_n\}$ such that $W_i \equiv W_{\omega_i}$ are independent as elements of X and span the set $\{W_i | 1 \leq i \leq m\}$. Let $E_n = span\{W_1, ..., W_n\} \subset X$ and note that E_n is a closed subspace of a Hilbert space, implying that it is also a Hilbert space and is isometric to its dual. Define a functional e^* on E_n by $e^*(W_i) = f(\omega_i)$. Therefore, there are $\alpha_1, ..., \alpha_n$ such that for every $x \in E_n$, $e^*(x) = \big\langle \sum_{i=1}^n \alpha_i W_i, x \big\rangle$. Thus, for every $1 \leq j \leq m$

$$f(\omega_j) = e^*(W_j) = \sum_{i=1}^{n} \alpha_i \big\langle W_i, W_j \big\rangle. \tag{2.5}$$

This *linear* equation system in the variables $(\alpha_1, ..., \alpha_n)$ has a *unique* solution. Moreover, if $(\alpha_1, ..., \alpha_n)$ is the solution then $\sum_{i=1}^n \alpha_i W_i \in B(X)$. Indeed,

let $P_{E_n} : X \to E_n$ be an orthogonal projection. Thus, if $E_n^{\perp}$ is the ortho–complement of E_n, then $f = P_{E_n} f + P_{E_n^{\perp}} f$. If $g = P_{E_n} f$, then $\|g\| \leq \|f\| \leq 1$ and for every $1 \leq i \leq m$,

$$g(\omega_i) = \langle P_{E_n} f, W_i \rangle = \langle P_{E_n} f + P_{E_n^{\perp}} f, W_i \rangle = \langle f, W_i \rangle = f(\omega_i).$$

The fact that this solution is unique follows since the matrix $A = (\langle W_i, W_j \rangle)$ is the matrix representation of the bilinear form on $E_n \times E_n$ given by $D(x, y) = \langle x, y \rangle$ with respect to the basis $\{W_1, ..., W_n\}$. Since D is positive definite then A is invertible, and the uniqueness follows.

Set g_n to be the solution of (2.5). We will show that (g_n) approximates f in the appropriate sense.

For any $h \in \mathcal{F}$, let $\mathcal{L}_h : \Omega \to \mathbb{R}^+$ be the loss function associated with h, i.e., $\mathcal{L}_h(x) = \big(h(x) - f(x)\big)^2$. It is easy to see (e.g., [2]) that if $\mathcal{F}$ is a class of uniformly bounded functions which is ε uniform Glivenko–Cantelli for every $\varepsilon > 0$, then $\mathcal{G} = \{\mathcal{L}_h \mid h \in B(X)\}$ also satisfies the uniform Glivenko–Cantelli condition for every $\varepsilon > 0$.

Given a probability measure μ_n on Ω, denote by $\mathbb{E}_{\mu_n}(\mathcal{L}_h)$ the empirical loss given by

$$\mathbb{E}_{\mu_n}(\mathcal{L}_h) = \frac{1}{n} \sum_{i=1}^{n} \big(h(\omega_i) - f(\omega_i)\big)^2.$$

Since $\mathcal{G}$ is uniform Glivenko–Cantelli, then for every $\varepsilon > 0$ and every $\delta > 0$ there is some N such that

$$Pr\{\sup_{n>N} \sup_{h \in \mathcal{F}} |\mathbb{E}_{\mu_n}(\mathcal{L}_h) - \mathbb{E}_{\mu}(\mathcal{L}_h)| \geq \varepsilon\} \leq \delta$$

for every probability measure μ on Ω. Since g_n solves (2.5), then for every $1 \leq i \leq n$, $g_n(\omega_i) = f(\omega_i)$. Thus, $\mathbb{E}_{\mu_n}(\mathcal{L}_{g_n}) = 0$, implying that for every μ,

$$Pr\{\sup_{n>N} \mathbb{E}_{\mu}(\mathcal{L}_{g_n}) \geq \varepsilon\} \leq \delta.$$

Hence, $\lim_{n \to \infty} \sup_{\mu} Pr\{\mathbb{E}_{\mu}(g_n - f)^2 \geq \varepsilon\} = 0$, as required.

■

Corollary 2.5 *Let X be as in Theorem 2.4, and assume that $\mathcal{F} \subset X$ such that there is some M for which $\sup_{f \in \mathcal{F}} \|f\| \leq M$. Then, if $f \in \mathcal{F}$, there is a map $A : \mathcal{S} \to X$ such that*

$$\lim_{N \to \infty} \sup_{\mu} Pr\{\sup_{n>N} \mathbb{E}_{\mu}(A_{S_n} - f)^2 \geq \varepsilon\} = 0$$

where $\mathcal{S} = \Omega^{\infty}$ and $S_n = S_n(f, \mu)$ are samples of f of length n drawn independently according to μ.

The difference between this Corollary and Theorem 2.4 is that here we do not impose that the approximating functions A_{S_n} belong to $\mathcal{F}$.

In this learning process we use a mechanism similar to the well known Support Vector Machine ([16]): first, we embed the sample in a high dimensional space and then we produce an approximating functional. In this case, since the sample is not Boolean, the functional produced is not a separating functional. Rather, it agrees with the unknown function on the sample. Since in this case the evaluation functionals are uniformly bounded, all the samples may be embedded in the dual of X.

Our procedure is easier than the Support Vector Machine since solving (2.5) and finding the desired g_n is obtained in two simple steps:
1) To calculate the coefficients in (2.5) which are $\langle W_{\omega_i}, W_{\omega_j} \rangle$, one can either use the reproducing kernel (since $\langle W_x, W_y \rangle = K(x,y)$), or, if one does not have an explicit formula for $K(x,y)$, one may use the fact that $\langle W_x, W_y \rangle = \sum_{n=1}^{\infty} u_n(x)u_n(y)$, where (u_n) is a complete orthonormal basis of X.
2) Once the coefficients are discovered, one needs to solve the *linear* equation system (2.5). The solution is unique, and automatically satisfies the norm constraint.
This procedure is much simpler than the quadratic programming minimization problem needed in the Support Vector Machine procedure. The computational price is paid in cases where one does not have an explicit formula for the reproducing kernel.

Note that we use the fact that the samples are not Boolean. For a Boolean sample, (2.5) becomes a system of inequalities, for which there may be many solutions, and one has to find a solution which satisfies the norm constraint. This problem requires quadratic programming, hence gives no advantage compared to the SVM.

One may be tempted to think of an even simpler learning procedure, which is to define the approximating functions (g_n) as a linear interpolation of the given sample. Appealing as it is, this method will not be useful since such functions g_n may not belong to X. Even if $(g_n) \subset X$, one does not have an a–priori bound on $\|g_n\|$, without which the proof that g_n tends to f is no longer true.
Thanks to the estimate for $VC_\varepsilon(\Omega, \mathcal{F})$, we can estimate the sample complexity in the Hilbert space setup. To that end, we shall make an additional assumption on the structure of our space. We impose that the family of functions $(B(X) - B(X))^2 = \{(f-h)^2 |\ f, h \in B(X)\}$ is norm bounded in X (i.e., there is some M such that $\sup_{f,g \in B(X)} \left\|(f-g)^2\right\| \leq M$). By Theorem 1.9 it follows that $VC_\varepsilon\big(\Omega, (B(X) - B(X))^2\big) = O(1/\varepsilon^2)$.

In many cases this assumption is satisfied automatically. For example, let $X = W_0^{1,2}(a,b)$ and note that by the Sobolev inequality, there is some constant $C > 0$ such that for every $f \in B(X)$, $\|f\|_\infty \leq C$, hence

$$\left\|(f-h)^2\right\|^2 = 4\int_a^b (f(t)-h(t))^2(f'(t)-h'(t))^2 dt \leq C' \|f-h\|^2 \leq C''.$$

Corollary 2.6 *Let X and $\mathcal{F}$ be as in Theorem 2.4. Assume that $(B(X) -$*

$B(X))^2$ *is norm bounded in* X. *Then,* $Pr\{\mathbb{E}_\mu(f-g_n)^2 \geq \varepsilon\} \leq \delta$ *for*

$$n = O\Big(\frac{1}{\varepsilon^2}\big(\frac{1}{\varepsilon^3}\log^2\frac{1}{\varepsilon} + \log\frac{1}{\delta}\big)\Big).$$

Proof: Since the evaluation functionals on X are uniformly bounded, then $\sup_{\omega\in\Omega}|f(w)| = \sup_{w\in\Omega}|\delta_\omega(f)| \leq \sup_{\omega}\|\delta_\omega\|\,\|f\|$. Hence a norm bounded class of functions $\mathcal{F} \subset X$ consists of functions with a uniformly bounded range. Thus, as in (1.1), $P_\gamma(\Omega,\mathcal{F}) \leq C\frac{VC_\gamma(\Omega,\mathcal{F})}{\gamma}$. We may assume that $\mathcal{F} \subset B(X)$ and put $\mathcal{G} = \{\mathcal{L}_h \,|\, h \in B(X)\}$. Since $\mathcal{G}$ is norm bounded in X then $VC_\varepsilon(\Omega,\mathcal{G}) = O(1/\varepsilon^2)$. Therefore, be setting $\tau = 1/8$ and applying (2.4), it follows that for every probability measure μ on Ω,

$$Pr\{\sup_{h\in\mathcal{G}}|\mathbb{E}_{\mu_n}(\mathcal{L}_h) - \mathbb{E}_\mu(\mathcal{L}_h)| \geq \varepsilon\} \leq \delta$$

for

$$n = O\Big(\frac{1}{\varepsilon^2}\big(\frac{1}{\varepsilon^3}\log^2\frac{1}{\varepsilon} + \log\frac{1}{\delta}\big)\Big).$$

The corollary follows since $\mathbb{E}_{\mu_n}(\mathcal{L}_{g_n}) = 0$.

■

Remark 1 *Here, the assumption that* $\big((B(X)-B(X)\big)^2$ *is bounded is used to bound the fat shattering dimension of* $\mathcal{G}$. *An alternative approach is to estimate the covering numbers of* $\mathcal{G}$ *in* $L_\infty(\mu_n)$, *and thus to provide complexity estimates (see [2] for further details). This approach yields the same complexity estimate without assuming that* $\big(B(X)-B(X)\big)^2$ *is norm bounded.*

3 Concluding Remarks

In this final section we discuss an alternative approach to the problem of estimating the fat shattering dimension of a class $\mathcal{F}$. A possible source of information regarding the fat shattering dimension may be derived once we know that $\mathcal{F}$ is a compact subset of $C(\Omega)$. It is tempting to think that in our setup (Banach spaces in which the evaluation functionals are uniformly bounded) $\mathcal{F}$ is compactly embedded in $C(\Omega)$, since this is the case in all the "classical" spaces, for example, Sobolev spaces or Bergman spaces (see [13]). Therefore, a valid question might have been whether the estimate shown in section 1 using a "soft" approach may be improved using a more direct method.

First, we will describe the alternative and more direct method mentioned above. Then, we shall give two examples: in the first one the estimates obtained using the two approaches coincide, while in the second one, the soft approach yields a better bound. Finally, we will construct a Hilbert space of continuous functions on $[0,1]$ in which the evaluation functionals are bounded by 1, but its unit ball is not a compact subset of $C(0,1)$. Hence, the direct approach does not apply to this case.

Assume that $\mathcal{F}$ is a compact subset of $C(\Omega)$ where Ω is the unit ball of $\mathbb{R}^d$ and put

$$osc_{\mathcal{F}}(\delta) = \sup_{f\in\mathcal{F}} \sup_{\|x-y\|\le\delta} |f(x)-f(y)|\,.$$

A standard compactness argument shows that $\lim_{\delta\to 0} osc_{\mathcal{F}}(\delta) = 0$, which ensures the existence of an upper bound on $osc_{\mathcal{F}}(\delta)$.

Let $\{\omega_1,...,\omega_n\}$ be ε–shattered in Ω, and set $r = \min_{i\neq j}\|\omega_i-\omega_j\|$. Thus, there are $x,y\in\Omega$ and $f\in\mathcal{F}$ such that $\|x-y\| = r$ and $|f(x)-f(y)|\ge\varepsilon$, implying that $osc_{\mathcal{F}}(r)\ge\varepsilon$.

On the other hand, for $i\neq j$, $B(\omega_i, r/2)$ and $B(\omega_j, r/2)$ are disjoint, and

$$(1+\frac{r}{2})\Omega \supset \bigcup_{i=1}^{n} B(\omega_i, \frac{r}{2}).$$

By a volume estimate, $(1+r/2)^d \ge n(r/2)^d$. Therefore,

$$n \le \big(1+\frac{2}{r}\big)^d.$$

To see how the two estimates combine when one has an upper estimate for $osc_{\mathcal{F}}$, consider for example, $\Omega = [0,1]$ and $X = W_0^{1,2}(\Omega)$. Note that there is some constant $C>0$ such that for every $f\in X$,

$$|f(x)-f(y)| \le C\,\|f\|\,|x-y|^{1/2}$$

(see [1], pg 110). Hence, if $\mathcal{F}\subset B(X)$ and $\{\omega_1,...,\omega_n\}$ is ε–shattered, then $\varepsilon \le osc_{\mathcal{F}}(r) \le C\sqrt{r}$. Thus, by the volume estimate, $n\le O(1/\varepsilon)^2$, which is the same as the estimate established in Theorem 1.9.

Turning to the second example, let $X = W_0^{1,p}(\Omega)$ which is the closure of $C_0^1(\Omega) = \{f | Df \text{ is continuous and } f(\partial\Omega)=0\}$ with respect to the norm

$$\|f\| = \Big(\int_\Omega \sum_{i=1}^{d} \Big|\frac{\partial f}{\partial x_i}\Big|^p\Big)^{\frac{1}{p}}.$$

Thus, X is a Banach space which is isometrically embedded in L_p. Moreover, by the Morrey embedding theorem ([1], [7]), if $p>d$ there is some constant $C = C(p,d)$ such that

$$osc_{\mathcal{F}}(\delta) \le C\delta^{\left(1-\frac{d}{p}\right)} \sup_{f\in\mathcal{F}} \|f\|\,.$$

Therefore, by the same argument presented above,

$$VC_\varepsilon(\Omega,\mathcal{F}) \le C(p,d)\big(\frac{1}{\varepsilon}\big)^{\frac{pd}{p-d}}.$$

For example, if $d\ge 2$ and $p>d$ then this estimate is much worse than the estimate obtained in Theorem 1.9. Indeed, for $p>2$, $W_0^{1,p}$ has type 2 as a subspace of L_p, hence $VC_\varepsilon(\Omega,\mathcal{F}) = O(1/\varepsilon^2)$.

Finally, we construct a Hilbert space of continuous functions on $[0,1]$ which is not compactly embedded in $C(0,1)$, yet the evaluation functionals are all bounded by 1. Let (A_n) be disjoint intervals on $[0,1]$. For every n, set f_n to be supported on $A_n = [a_n, b_n]$, such that f_n is the piecewise–linear interpolation of $f_n(a_n) = f_n(b_n) = 0$ and $f_n(\frac{a_n + b_n}{2}) = 1$. Put

$$X = \{\sum_{n=1}^{\infty} \lambda_n f_n \mid \sum_{n=1}^{\infty} \lambda_n^2 < \infty\}$$

with $\left\| \sum_{n=1}^{\infty} \lambda_n f_n \right\| = \left(\sum_{i=1}^{n} \lambda_n^2\right)^{1/2}$. It is easy to see that X is a Hilbert space consisting of continuous functions on $[0,1]$. Also, for every $x \in [0,1]$ and every $f \in X$,

$$|\delta_x(f)| = |f(x)| \leq \max_n |\lambda_n| \leq \left(\sum_{n=1}^{\infty} \lambda_n^2\right)^{1/2} = \|f\|.$$

Hence, the evaluation functionals are uniformly bounded by 1. On the other hand, (f_n) in not a compact set in $C(0,1)$ since $\|f_n - f_m\|_\infty = 1$.

Thus, the "soft" approach may succeed in cases where the direct one fails. However, we do not rule out the possibility that if one imposes strict conditions on $\mathcal{F}$, the direct method may yield a better bound than the bound established in Theorem 1.9.

References

[1] R.A. Adams: Sobolev Spaces, Pure and Applied Mathematics series 69, Academic Press 1975

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

[3] P.L. Bartlett, P.M. Long: More theorems about scale sensitive dimensions and learnability. *Proceedings of the 8th annual conference on Computational learning theory*, 392–401, ACM

[4] R.M. Dudley: A course on Empirical Processes, *Lecture notes in Mathematics* 1097, 1–143, Springer–Verlag 1984

[5] R.M. Dudley: *Uniform Central Limit Theorems*, Cambridge Studies in advanced mathematics, 63, Cambridge University Press, 1999

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

[7] D.Gilbarg, N.S. Trudinger: *Elliptic Partial Differential Equations of Second Order* Springer 1998

[8] L. Gurvits: A note on a scale sensitive dimension of bounded linear functionals in Banach spaces, Preprint

[9] P. Habala, P. Hajek, V. Zizler: *Introduction to Banach spaces [II]*, matfyzpress, 1996

[10] M. Kearns, R. Schapire: Efficient distribution–free learning of probabilistic concepts, J. Comput. Syst. Sci. 48, 3 (1994) 464–497

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

[12] G. Pisier: Probabilistic methods in the geometry of Banach spaces, *Probability and Analysis*, Lecture notes in Mathematics 1206, 167–241, Springer Verlag 1986

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

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

[15] A.W. Van–der–Vaart, J.A. Wellner: *Weak convergence and Empirical Processes*, Springer series in Statistics, 1996

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