Statistical Sufficiency for Classes in Empirical L_2 Spaces

Shahar Mendelson and **Naftali Tishby**
School of Computer Science and Engineering
The Hebrew University, Jerusalem 91904, Israel

Abstract

We explore the notion of ε sufficient linear statistics for a class of real valued functions. We show that for function classes with a polynomial rate of the Parametric Pollard dimension one can find a set of linear empirical functionals of polynomial size in the dimension that are sufficient for ε approximation of any function in the class. We also present a probabilistic scheme for producing those functionals.

1 Introduction

A fundamental problem in statistical estimation theory is the availability of a set of empirical functions that capture the information on the parameters of the underlying distribution. For parametric distributions such functions were called "sufficient statistics" and the question of their possible existence was fully answered by the celebrated theorem of Koopman [9], Pitman [13], and Darmois [3], who restricted it to exponential families. A direct corollary of these results is that when sufficient statistics exist they can always be expressed as empirical means (for i.i.d. samples), or as linear functionals of the sample points. This fundamental results makes the parameter estimation for exponential families very efficient, both in terms of the estimate variance and computational complexity.

In this paper we address the question of the availability of a similar notion for the learnability of functions. We consider the number of empirical functionals (defined below) that capture the information needed for approximating a function, based on its values on a given random sample. We define the notion of ε-sufficient statistics as a set of linear functionals of the sample points, whose values suffice to obtain an L_2 ε-approximation for any target function in a class $\mathcal{F}$.

We begin with a few definitions and some notation. Throughout, μ will denote a probability measure on a set $\Omega \subset \mathbb{R}^d$. To avoid measurability problems, we will assume that all the measures are Borel measures. For every measure μ, $\mathbb{E}_\mu$ is the expectation with respect to μ, and $L_2(\mu)$ is the set of all measurable functions on Ω such that $\mathbb{E}_\mu |f|^2 < \infty$. This space is a Hilbert space with respect to the norm $\|f\|_{L_2(\mu)} = \left(\mathbb{E}_\mu |f|^2\right)^{1/2}$.

For every set $\mathcal{S}_n = \{\omega_1, ..., \omega_n\} \subset \Omega$, let μ_n be an empirical measure supported on $\mathcal{S}_n$. Thus, $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{\omega_i}$, where δ_{ω_i} is the evaluation functional at ω_i (i.e., $\delta_{\omega_i}(f) = f(\omega_i)$). Ω^∞ is the infinite product of the set Ω. Each $\vec{\omega} \in \Omega^\infty$ is of the form $(\omega_1, \omega_2, ...)$, where each $\omega_i \in \Omega$. For every probability measure μ on Ω we endow Ω^∞ with the infinite product measure μ^∞, which is also a probability measure.

Definition 1.1 *A linear functional x^* is called empirical if it is a linear combination of point evaluation functionals, i.e., $x^* = \sum_{i=1}^m a_i \delta_{\omega_i}$. We say that x^* is supported on the set $\{\omega_1, ..., \omega_n\}$ if it has a representation as a linear combination of $\{\delta_{\omega_1}, ..., \delta_{\omega_n}\}$.*

Definition 1.2 *Let $\mathcal{F}$ 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 $(S_i)_1^m$ is called ε-sufficient statistics with respect to $L_2(\mu)$ if, for every $f, g \in \mathcal{F}$ such that for every $1 \le i \le m$, $S_i(g) = S_i(f)$, then $\|f - g\|^2_{L_2(\mu)} < \varepsilon$. The infimum on the number of the ε sufficient statistics of $\mathcal{F}$ in $L_2(\mu)$ is denoted by $S_{\mathcal{F},\mu}(\varepsilon)$.*

Note that by this definition, the functionals (S_i) capture the structure of the class $\mathcal{F}$ up to a small permitted error. For example, for any $f \in \mathcal{F}$ the data $\left(S_i(f)\right)$ is enough to characterize f up to an accuracy of ε. Also, it is important to emphasize that the selection of ε sufficient statistics must have a random element when the measure μ is unknown. Hence, unless prior information on the measure is given, one has to involve sampling according to μ in the selection process of the sufficient statistics.

The problem we wish to investigate is how to estimate the number of linear empirical functionals needed to ensure ε statistical sufficiency.

This problem has two aspects. First, one has to bound the number of statistics needed for ε-sufficiency. Second, (though important), one has to estimate the size of the sample on which the set of statistics is supported.

One example which we shall focus on is that of uniform Glivenko–Cantelli classes:

Definition 1.3 *A class of functions $\mathcal{F}$ is called a uniform Glivenko Cantelli class (GC class) if for every $\varepsilon > 0$*

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

where μ_m is the empirical measure supported the first m coordinates of $\vec{\omega} = (\omega_1, ...) \in \Omega^\infty$, and for every measure μ, Pr is the infinite product measure μ^∞. The supremum is taken with respect to all the (Borel) probability measures on Ω.

For every $\varepsilon > 0$ and $0 < \delta \leq 1$, set $n_{\mathcal{F}}(\varepsilon, \delta)$ to be the sample complexity of $\mathcal{F}$, i.e., the minimal n for which

$$\sup_{\mu\in\Lambda} Pr\Big\{ \sup_{m>n} \sup_{f\in\mathcal{F}} |\mathbb{E}_\mu f - \mathbb{E}_{\mu_m} f| \geq \varepsilon \Big\} \leq \delta.$$

A trivial solution to the two parts of our puzzle may be found through the Glivenko-Cantelli condition. Indeed, it is possible to show that if $\mathcal{F}$ is a GC class of functions into $[0,1]$, then $(\mathcal{F}-\mathcal{F})^2 = \{(f-g)^2 | f, g \in \mathcal{F}\}$ is also GC. Hence, for every $\varepsilon > 0$ and $\delta \in (0,1)$, there is a set of samples $\mathcal{U}_{\varepsilon,\delta} \subset \Omega^\infty$ such that $Pr(\mathcal{U}_{\varepsilon,\delta}) \geq 1-\delta$ and for every $n \geq n_{(\mathcal{F}-\mathcal{F})^2}(\varepsilon,\delta)$ and every $\vec{\omega} \in \mathcal{U}_{\varepsilon,\delta}$,

$$\sup_{f,g\in\mathcal{F}} \left| \mathbb{E}_\mu (f-g)^2 - \mathbb{E}_{\mu_n} (f-g)^2 \right| < \varepsilon, \qquad (1.1)$$

Thus, for every $\vec{\omega} = \{\omega_1, ..., \omega_n, ...\} \in \mathcal{U}_{\varepsilon,\delta}$ and $n = n_{(\mathcal{F}-\mathcal{F})^2}(\varepsilon,\delta)$ the statistics $S_i = \delta_{\omega_i}$, $1 \leq i \leq n$ are ε-sufficient.

Therefore, for every $\varepsilon, \delta \in (0,1)$, $n_{(\mathcal{F}-\mathcal{F})^2}(\varepsilon,\delta)$ yields an upper bound to the number of statistics $S_{\mathcal{F},\mu}(\varepsilon)$, as well as to the size of the sample on which the statistics are supported.

It is important to note that in order to apply this bound one needs to use the Glivenko-Cantelli condition for the class $(\mathcal{F}-\mathcal{F})^2$, and not with respect to any specific loss-function class $\mathcal{L}_h = \{(f-h)^2 | f \in \mathcal{F}\}$. If one were to use any specific loss-function class associated with some $h \in \mathcal{F}$, the set of statistics (δ_{ω_i}) would be ε sufficient for that particular target concept. For example, if $\mathcal{F}$ is a GC class and $f, g \in \mathcal{F}$ have disjoint supports, then a sample which yields a good approximation for f may be contained in the support of f, and thus may prove to be a "bad" sample of g. However, we need the sufficient statistics to apply to every $h \in \mathcal{F}$, hence one has to use the Glivenko-Cantelli condition for $(\mathcal{F}-\mathcal{F})^2$.

We show that the size of a set of ε sufficient statistics may be significantly improved. The improved bound is established using a parameter originating from the local theory of Banach spaces called the ℓ-norm. It was shown by Dudley and Sudakov that the ℓ-norm of a set $\mathcal{F}$ is related to the covering number and entropy integral of $\mathcal{F}$. The improvement in the upper bound becomes more significant as the size of the class increases (e.g. when the class has a larger parametric Pollard dimension). This is done without increasing the size of the sample on which the functionals are supported.

Another application which may be derived from the theory we develop here is a learning process for a target concept which belongs to a GC class.

In the usual context of a learning problem, one tries to estimate a function based on its values on a given sample. This is usually done by selecting a sample $\{\omega_1, ..., \omega_n\}$ according to the given measure μ, and finding some function f from the class which agrees with the target concept h on that sample. If the class is a Glivenko-Cantelli class and assuming that the sample is large enough, it follows that $\|f-h\|_{L_2(\mu)}$ is small with high probability.

In the language we wish to introduce, one can say that for every target concept $h \in \mathcal{F}$ and for n large enough the point evaluation functionals $(\delta_{\omega_i})_1^n$ are with high probability "ε-sufficient statistics" in the following sense: given $\delta_{\omega_i}(h) = h(\omega_i)$, then for every $f \in \mathcal{F}$ such that $\delta_{\omega_i}(f) = \delta_{\omega_i}(h)$ for every $1 \leq i \leq n$, $\|f-h\|^2_{L_2(\mu)} < \varepsilon$. Thus, learning problem is reduced to finding some $f \in \mathcal{F}$ which satisfies the set of linear empirical constraints $\delta_{\omega_i}(f) = \delta_{\omega_i}(h)$.

Note that the functionals (δ_{ω_i}) are not ε sufficient in the "usual sense", since the fact that a set $(\delta_{\omega_i})_1^n$ captures almost all the information regarding a one target concept in $\mathcal{F}$ does not guarantee it will do the job for other target concepts in $\mathcal{F}$. In other words, in the context of a learning problem the statistics depend on the target concept.

We show that for any function in the given class it is possible to reduce the number of linear constraints (viewed as linear equations) that the sample induces on the function class from the sample size, to approximately its square-root, without losing any information on the target. This suggests a more computationally efficient algorithm for learning a concept from a class with a finite VC dimension or a "small" parametric Pollard dimension. (see definition below).

The results we present are in two generic cases. The first in when $\mathcal{F}$ is viewed as a subset of $L_2(\mu_n)$ for some empirical measure μ_n. In this case the statistics are supported on the same points as the empirical measure. The second case we explore is when one views $\mathcal{F}$ as a subset of $L_2(\mu)$ for a general probability measure μ.

2 Theoretical Background

This section is devoted to several well know definitions which will be used in the sequel. The following are definitions of well known combinatorial parameters which are used to characterize GC classes.

Definition 2.1 *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 $j \notin I$. Let* $VC(\mathcal{F}, \Omega) = \sup\Big\{ |A| \,\Big|\, A \subset \Omega,\ A \text{ is shattered by } \mathcal{F} \Big\}$.

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

Definition 2.2 *Let $\mathcal{F}$ be a class of functions on a space Ω and let $\varepsilon > 0$. We say that $\mathcal{F}$ ε–shatters $\{\omega_1, ..., \omega_n\} \subset \Omega$ if there is some $a \in \mathbb{R}$ such that for every $I \subset$*

$\{1,...,n\}$ *there is a function* $f_I \in \mathcal{F}$ *for which* $f(\omega_i) \geq a+\varepsilon/2$ *if* $i \in I$ *and* $f_I(\omega_j) \leq a-\varepsilon/2$ *if* $j \notin I$. *Let*

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

$VC_\varepsilon(\mathcal{F},\Omega)$ *is called the fat shattering dimension of* $\mathcal{F}$.

The parametric Pollard dimension (defined below) may serve the same purposes as the fat shattering dimension.

Definition 2.3 *For every* $\varepsilon > 0$, *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* $j \notin I$. *Let*

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

By the pigeonhole principle it is easy to see that the parametric Pollard dimension and the fat shattering dimension are related for classes of functions which have a uniformly bounded range (i.e., if there is some $M \in \mathbb{R}$ such that $\sup_{f\in\mathcal{F}}\sup_{\omega\in\Omega}|f(\omega)| \leq M$). In this case, there is some constant $C > 0$ such that for every $\varepsilon > 0$,

$$VC_\varepsilon(\mathcal{F},\Omega) \leq P_\varepsilon(\mathcal{F},\Omega) \leq C\frac{VC_{\varepsilon/2}(\mathcal{F},\Omega)}{\varepsilon},$$

where C depends only on the uniform bound on the members of $\mathcal{F}$.

The connection between Glivenko-Cantelli classes and the combinatorial parameters defined above is the following fundamental result:

Theorem 2.4 *A class of* $\{0,1\}$ *valued functions is a Glivenko–Cantelli class if and only if it has a finite VC dimension. A class of uniformly bounded real-valued functions is a Glivenko–Cantelli class if and only if it has a finite parametric Pollard dimension for every* $\varepsilon > 0$.

The "if" part in the first claim is due to Vapnik and Chervonenkis (see [16]) while the "only if" is due to Assouad and Dudley ([2]). The second claim was established by Alon, Ben-David, Cesa-Bianchi and Haussler ([1]).

A key tool in the analysis of GC classes are covering number estimates (defined below). It turns out that the covering numbers of a given class not only determines whether it is a GC class or not, but, in fact, enable one to estimate the sample complexity (see [6], [1]).

If (X,d) is a metric space and if $\mathcal{F} \subset X$, denote by $N(\varepsilon,\mathcal{F},d)$ the minimal number of open 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 metric is clear we shall denote the covering numbers by $N(\varepsilon,\mathcal{F})$.

The course of action we take is as follows: we use a well know geometric parameter from the local theory of Banach spaces called the ℓ-norm. This parameter measures how "large" a given set is. It is possible to establish both upper and a lower bounds on $\ell(\mathcal{F})$ by the L_2-log covering numbers of the set $\mathcal{F}$. This important fact is due to Dudley ([4]) and Sudakov ([14]). In section 3 and in the Appendix we investigate the ℓ-norms of GC classes in empirical L_2 spaces. We provide an upper bound to $\ell(\mathcal{F})$ in terms of $VC(\mathcal{F})$ or $P_\varepsilon(\mathcal{F})$.

The ℓ-norm estimates enable us to bound the number of the statistics required for ε sufficiency.

Recall that a set K is said to be symmetric if the fact that $x \in K$ implies that $-x \in K$. Consider a convex symmetric set $K \subset L_2(\mu_n)$. If $x_1^*,...,x_k^*$ are linear functionals on $L_2(\mu_n)$ then they induce a k-codimensional section of K. Indeed, $V = \cap_i ker(x_i^*)$ is a k-codimensional subspace of $L_2(\mu_n)$, thus $V \cap K$ is a k-codimensional section of K. If the diameter of this section is small, then the functionals $x_1^*,...,x_k^*$ may be used to identify any member of K: if $g,f \in K$ such that $x_i^*(g) = x_i^*(f)$, then $g/2 - f/2 \in V$. Moreover, since K is convex and symmetric then $g/2 - f/2 \in K$, implying that the $\|g/2-f/2\|_{L_2(\mu_n)}$ is bounded by the diameter of $V \cap K$, which was assumed to be small. Intuitively, for every vector of k real numbers $\vec{\alpha} = (\alpha_1,...,\alpha_k)$, the affine space $V(\vec{\alpha}) = \{x \in X | x_i^*(x) = \alpha_i,\ i = 1,...n\}$ which is a translation of V, slices K to disjoint slices $K \cap V(\vec{\alpha})$. Each slice has a diameter which is smaller than $diam(V \cap K)$. Hence, if one wishes to locate an unknown element $h \in K$, it is enough to find some $f \in K$ which is on the same translation of V as h. Such an f will automatically be "close" to h, since their distance is bounded by the diameter of $K \cap V$.

The connection to our problem is simple. Given the class $\mathcal{F}$ and an empirical measure μ_n supported on $\{\omega_1,...,\omega_n\}$, set K/μ_n to be the symmetric convex hull of $\mathcal{F}$, viewed as subset of $L_2(\mu_n)$. Because of the definition of the empirical L_2 space, each linear functional on $L_2(\mu_n)$ is a linear combination of point evaluation functionals δ_{ω_i}, $1 \leq i \leq n$. Hence, if $(x_i^*)_1^k$ are such that $diam\big(\cap_i\big(ker(x_i^*) \cap K/\mu_n\big)\big) < \sqrt{\varepsilon}$, then $x_1^*,...,x_k^*$ are ε sufficient statistics for the convex hull of $\mathcal{F}$ in $L_2(\mu_n)$, since they are empirical and capture almost all the relevant information regarding the convex hull of $\mathcal{F}$. Thus, they are ε sufficient for $\mathcal{F}$ itself.

It is possible to show that the ℓ-norm may be used to connect the number of functionals selected with the diameter of an "almost optimal" slice of the given set of that codimension. This celebrated result is due to Pajor and Tomczak-Jaegermann (see [11]). Moreover, they were able to show that an "almost optimal" section may be obtained using random selection process, and, in fact, most of the k-codimensional sections of the set are "almost optimal". In section 4 we combine their result with the ℓ-norm estimates discussed in section 3 and provide a bound on the number of statistics needed to ensure ε-sufficiency in empirical L_2 spaces. We then use the Glivenko-Cantelli condition to pass from empirical L_2 spaces to general L_2 spaces and establish a bound on the number of ε sufficient statistics in general $L_2(\mu)$ spaces.

3 ℓ-norm estimates in $L_2(\mu_n)$

We begin this section with a several standard definitions from the theory of Banach spaces.

Given a Banach space X, the *dual* of X (denoted by X^*) consists of all the bounded linear functionals on X, with the norm $\|x^*\|_{X^*} = \sup_{\|x\|_X=1} |x^*(x)|$. Let ℓ_2^n be a real n-dimensional inner product space with respect to the inner product $\langle\ ,\ \rangle$ and let K be a bounded convex symmetric subset of ℓ_2^n which has a nonempty interior. It follows that K is the unit ball of some norm denoted by $\|\ \|_K$. Set $\|\ \|_{K^*}$ to be the dual norm to $\|\ \|_K$.

Recall the definition of the ℓ-norm of a set F:

Definition 3.1 *For every set $F \subset \ell_2^n$, let*

$$\ell(F) = \Big(\int_{\mathbb{R}^n} \sup_{f \in F} \big|\langle f, x\rangle\big|^2 d\gamma_n \Big)^{\frac{1}{2}}, \qquad (3.1)$$

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

Denote by K the symmetric convex hull of a bounded set $F \subset \ell_2^n$ 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 the closure of $F \cup -F$, 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 $\ell(K) = \ell(F)$.

The following deep result provides a connection between the ℓ-norm of a set and its covering numbers in ℓ_2^n. The upper bound was established by Dudley in [4] while the lower one is due to Sudakov (see [14]). A proof of both bounds may be found in [12].

Theorem 3.2 *Let $F \subset \ell_2^n$. Then there are absolute positive constants c and C such that*

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

If $\mathcal{F}$ is a class of functions on Ω, then for every empirical measure μ_n, $\mathcal{F}$ may be viewed as a subset of $L_2(\mu_n)$ – which is isometric to ℓ_2^n. Indeed, if χ_{ω_i} is the characteristic function of the set $\{\omega_i\}$ then

$$\mathcal{F}/\mu_n = \Big\{ \sum_{i=1}^n f(\omega_i)\chi_{\omega_i} | f \in \mathcal{F} \Big\} = \Big\{ \sum_{i=1}^n n^{-\frac{1}{2}} f(\omega_i) e_i | f \in \mathcal{F} \Big\},$$

where e_i is an orthonormal basis of $L_2(\mu_n)$.

It is possible to obtain upper bounds on $\ell(\mathcal{F}/\mu_n)$ based on the entropy of the class $\mathcal{F}$. We shall focus on two cases. The first, is when $\mathcal{F}$ is the a class of $\{0,1\}$ functions with a finite VC dimension. The second case is when $\mathcal{F}$ a class which consists of functions bounded by 1, such that the parametric Pollard dimension is $O(\varepsilon^{-p})$ for some $p > 0$.

Below are the ℓ-norm estimates we were able to establish. The proof of this claim may be found in the Appendix.

Theorem 3.3 *Let $\mathcal{F}$ be a class of functions whose range is a subset of $[0,1]$.*

1. *If $\mathcal{F}$ is a $\{0,1\}$ class such that $VC(\mathcal{F}) = d$, then there is some absolute constant C such that for every empirical measure μ_n, $\ell(\mathcal{F}/\mu_n) \le Cd^{1/2}$.*
2. *Assume that for every $\varepsilon > 0$, $P_\varepsilon(\mathcal{F}) \le \gamma\varepsilon^{-p}$ for some $\gamma \ge 1$. Then, there are constants C_p which depend only on p such that for every $n > 1$ and every empirical measure μ_n,*

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

4 Application of the ℓ-norm estimates

In this section we show how the ℓ-norm estimates assist us in estimating the number of statistics needed to ensure ε sufficiency.

From the geometric point of view, we focus our attention to the possibility of constructing a subspace $V \subset L_2(\mu_n)$ which has a "small" codimension such that the diameter of its intersection with K/μ_n is also small, where K is the symmetric convex hull of $\mathcal{F}$.

It turns out that the diameter of an "almost optimal" k-codimensional section may be estimated in terms of the ℓ-norm. This important result is due to Pajor and Tomczak Jaegermann (see [12]). Moreover, it follows that the desired subspace may be selected randomly in some sense. Indeed, let (g_{ij}) be standard independent Gaussian random variables on some probability space Y. Set $G : \ell_2^n \to \ell_2^m$ to be an operator whose matrix representation with respect to an orthonormal basis is (g_{ij}).

Theorem 4.1 *Let $K \subset \ell_2^n$ be convex, bounded and symmetric with a nonempty interior. There is an absolute constant C_1 and a set $Y_1 \subset Y$, such that $Pr(Y_1) \ge 1/3$ and for every $y \in Y_1$*

$$diam\big(ker G(y) \cap K\big) \le C_1 m^{-1/2} \ell(K).$$

Also, there is some absolute constant C_2 such that if $n > m > C_2 \log(1/\delta)$ then Y_1 may be chosen so that $Pr(Y_1) \ge 1 - \delta$.

The proof of the first part of Theorem 4.1 appears in [12]. The estimate on the measure the set Y_1 may be found in [7].

In our case, the n dimensional Hilbert space is $L_2(\mu_n)$ and the set we wish to investigate is

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

Note that the assumption that K/μ_n has a nonempty interior poses no obstacle. Due to the structure of $L_2(\mu_n)$, the set K/μ_n has an empty interior in $L_2(\mu_n)$ if and only if there is some ω_i on which all the elements of $\mathcal{F}$ vanish. Thus, by removing such points from Ω, we may assume that K/μ_n has a nonempty interior. Recall that in $L_2(\mu_n)$ the set $(\sqrt{n}\chi_{\omega_i})_{i=1}^n$ is an orthonormal basis. Therefore, the functionals (x_i^*) for which $diam\big(ker(x_i^*) \cap K/\mu_n\big)$ is small are given by

$$\sqrt{n}\sum_{j=1}^n g_{ij}(y)\chi_{\omega_j}.$$

Thus, if $y \in Y_1$ and $f, g \in \mathcal{F}$ such that for every $1 \leq j \leq m$

$$\sum_{j=1}^n g_{ij}(y)f(\omega_j) = \sum_{j=1}^n g_{ij}(y)h(\omega_j) \qquad (4.1)$$

then $\|f-h\|_{L_2(\mu_n)} \leq C_1 m^{-1/2}\ell(\mathcal{F}/\mu_n)$.

4.1 Application for Sufficient Statistics

Here, we show how to construct ε-sufficient statistics for the class $\mathcal{F}$. We begin with the case where the sufficient statistics are constructed in empirical L_2 spaces.

Theorem 4.2 *Let $\mathcal{F}$ be a class of functions into $[0,1]$. Put $0 < \delta < 1$ and let μ_n be an empirical measure on Ω for some $n > 1$.*

1. *If $VC(\mathcal{F}) = d$ then there is some absolute constant C such that for every $\varepsilon > 0$, there exist a system of at most $m = C\frac{d}{\varepsilon}$ linear empirical functionals (x_i^*), such that if f, g satisfy that $x_i^*(f) = x_i^*(g)$, then $\|f-g\|^2_{L_2(\mu_n)} < \varepsilon$.*
2. *If $P_\varepsilon(\mathcal{F}) \leq \frac{\gamma}{\varepsilon^p}$ then there is a set of at most m empirical linear functionals (x_i^*) such that if $f, g \in \mathcal{F}$ satisfy that $x_i^*(f) = x_i^*(g)$, then $\|f-g\|^2_{L_2(\mu_n)} < \varepsilon$. The number of equations required is*

$$m \leq \begin{cases} C_p\frac{\gamma}{\varepsilon}\log^2 n & \text{if } 0 < p < 2, \\ C_2\frac{\gamma}{\varepsilon}\log^4 n & \text{if } p = 2, \\ C_p\frac{\gamma}{\varepsilon}n^{1-\frac{2}{p}}\log^2 n & \text{if } p > 2. \end{cases}$$

where C_p is a constant which depends only on p.

In both cases the selection of the functionals $(x_i^*)_1^m$ which determine the system of equations is random. There is some absolute constant C_1 such that if $m > C_1\log(\frac{1}{\delta})$ then with probability larger than $1-\delta$ the random process provides functionals $(x_i^*)_1^m$ for which our assertion holds.

We shall present a partial proof to this claim by establishing its first part. The remaining assertions follow using similar methods, by applying the ℓ-norm estimates from the previous section.

Proof: Assume that $VC(\mathcal{F}) = d$, let μ_n be an empirical measure and set K/μ_n the symmetric convex hull of $\mathcal{F}/\mu_n$. Thus, by Theorem 3.3, $\ell(K/\mu_n) \leq Cd^{1/2}$. Given $\varepsilon, \delta \in (0,1)$ and $1 \leq i \leq m$, let $x_i^* = \sum_{j=1}^n g_{ij}(y)\delta_{\omega_j}$, where $m = O\big(\log(\frac{1}{\delta})\big)$ and (g_{ij}) are standard independent Gaussian random variables on a space Y. By Theorem 4.1, there is a set $Y_1 \subset Y$ such that $Pr(Y_1) > 1-\delta$, and for every $y \in Y_1$,

$$diam \bigcap_{i=1}^n \big(ker(x_i^*) \cap K/\mu_n\big) \leq C\Big(\frac{d}{m}\Big)^{\frac{1}{2}}.$$

Clearly, for such y, the set $\{x_1^*, ..., x_m^*\}$ are $C\frac{d}{m}$ sufficient statistics for $\mathcal{F}$ in $L_2(\mu_n)$. Indeed, if $x_i^*(f) = x_i^*(g)$ for every $1 \leq i \leq m$ then $\|f-h\|_{L_2(\mu_n)} < C\big(\frac{d}{m}\big)^{1/2}$. Our claim follows by selecting $m = O\big(\max\Big\{\log\frac{1}{\delta}, \frac{d}{\varepsilon}\Big\}\big)$. ∎

By the proof of Theorem 4.2 it follows that there is a random construction algorithm for the sufficient statistics in empirical L_2 spaces, which does not depend on the exact structure of the class $\mathcal{F}$, only on its "size", as captured by the ℓ-norm.

Thus far, we established a bound on the number of ε sufficient statistics in empirical L_2 spaces. When one wishes to pass from empirical L_2 spaces to general L_2 spaces, one has to take advantage of the fact that our class is a GC class. Indeed, if μ_n is an empirical measure such that $\big|\mathbb{E}_\mu(f-g)^2 - \mathbb{E}_{\mu_n}(f-g)^2\big| < \varepsilon$ for every $f, g \in \mathcal{F}$, and if $S_1, ..., S_m$ are ε-sufficient statistics in $L_2(\mu_n)$, then they are also 2ε-sufficient statistics in $L_2(\mu)$.

We shall utilize this fact and establish the desired estimates for general $L_2(\mu)$ spaces. To that end, we need the following sample complexity estimates for $(\mathcal{F}-\mathcal{F})^2$. Recall that for every $\varepsilon > 0$ and $0 < \delta < 1$, $n_{\mathcal{F}}(\varepsilon, \delta)$ denotes the sample complexity estimate of the class $\mathcal{F}$ associated with the accuracy ε and the confidence δ.

Lemma 4.3 *Let $\mathcal{F}$ be a GC class of functions whose range is a subset of $[0,1]$ and set $G = (\mathcal{F}-\mathcal{F})^2$.*

1. *If $\mathcal{F}$ is a $\{0,1\}$ class and $VC(\mathcal{F}) = d$ then there is some absolute constant C such that*

$$n_G(\varepsilon, \delta) = O\Big(\frac{d}{\varepsilon\delta}\Big)$$

for every $\varepsilon > 0$ and $0 < \delta < 1$.

2. *If $P_\varepsilon(\mathcal{F}) \leq \gamma\varepsilon^{-p}$, then there are constants C_p which depend only on p such that for every $\varepsilon > 0$ and every $0 < \delta < 1$, $n_G(\varepsilon, \delta) \leq C_p\frac{\gamma}{\varepsilon^2}\Big(\frac{1}{\varepsilon^p}\log^3\frac{1}{\varepsilon} + \log\frac{1}{\delta}\Big)$.*

The proof of the Lemma is standard, hence it is omitted. An argument similar to the one used in the proof may be found in [10].

Corollary 4.4 *Let $\mathcal{F}$ be a GC class of functions into $[0,1]$ and let μ be a probability measure on Ω.*

1. *If $VC(\mathcal{F}) = d$ then $S_{\mathcal{F},\mu}(\varepsilon) \leq C\frac{d}{\varepsilon}$ for some absolute constant C. Moreover, the statistics are supported on a sample of $C'\big(\frac{d}{\varepsilon}\log\frac{1}{\varepsilon}\big)$ points at the most, where C' is some absolute constant.*
2. *If $P_\varepsilon(\mathcal{F}) \leq \frac{\gamma}{\varepsilon^p}$ then*

$$S_{\mathcal{F},\mu}(\varepsilon) \leq \begin{cases} C_p\frac{\gamma}{\varepsilon}\log^2\frac{\gamma}{\varepsilon} & \text{if } 0<p<2, \\ C_2\frac{\gamma}{\varepsilon}\log^4\frac{\gamma}{\varepsilon} & \text{if } p=2, \\ C_p\frac{\gamma}{\varepsilon}\frac{1}{\varepsilon^{p-\frac{4}{p}}}\log^5(\frac{\gamma}{\varepsilon}) & \text{if } p>2. \end{cases}$$

where C_p is a constant which depends only on p.

Each functional S_i is supported on a sample of at most $D_p(\frac{1}{\varepsilon^{p+2}}\log^3\frac{1}{\varepsilon})$ elements, where D_p depends only on p.

Again, we shall prove only the first part of the Corollary. The other claims follow in a similar fashion.

Proof: Let $\mathcal{F}$ be a $\{0,1\}$ class such that $VC(\mathcal{F}) = d$. Let $\varepsilon > 0$, fix some $\delta \in (0,1)$ and put $n = O\big(\frac{d}{\varepsilon\delta}\log\frac{1}{\varepsilon}\big)$, which is the sample complexity estimate for $(\mathcal{F}-\mathcal{F})^2$. Since $\delta < 1$ there is some empirical measure μ_n such that for every $f,g \in \mathcal{F}$, $\big|\mathbb{E}_\mu(f-g)^2 - \mathbb{E}_{\mu_n}(f-g)^2\big| < \varepsilon$. By Theorem 4.2 there exist a set of $m = O(\frac{d}{\varepsilon})$ linear empirical functionals $S_1,...,S_m$ such that if $S_i(f) = S_i(g)$ for every $1 \leq i \leq m$, then $\|f-g\|^2_{L_2(\mu_n)} < \varepsilon$. Therefore, $S_1,...,S_m$ are 2ε sufficient statistics in $L_2(\mu)$. Our claim follows by taking $\delta \to 1$.

■

4.2 Example

As an example, let $\mathcal{F}$ be the class of all the functions $f : [0,1] \to [0,1]$ such that for every $x,y \in [0,1]$, $|f(x)-f(y)| \leq |x-y|$. To estimate the fat shattering dimension of $\mathcal{F}$, note that if $\{\omega_1 < \omega_2 < ... < \omega_n\}$ is ε shattered, then for every $1 \leq i \leq n$, there is some $f \in \mathcal{F}$ such that

$$\varepsilon \leq f(\omega_{i+1}) - f(\omega_i) = |f(\omega_{i+1}) - f(\omega_i)| \leq \omega_{i+1} - \omega_i.$$

Hence,

$$1 \geq \omega_n - \omega_1 = \sum_{i=1}^n \omega_{i+1} - \omega_i \geq n\varepsilon,$$

and $n \leq \frac{1}{\varepsilon}$. On the other hand, it is easy to see that

$$VC_\varepsilon(\mathcal{F}) \geq \Big[\frac{1}{\varepsilon}\Big].$$

By the connection between the parametric Pollard dimension and the fat shattering dimension, it follows that

$$\Big[\frac{1}{\varepsilon}\Big] \leq P_\varepsilon(\mathcal{F}) \leq \frac{1}{\varepsilon^2}.$$

Let μ be a probability measure on $[0,1]$ and set some $\varepsilon \in (0,1)$. By the sample complexity estimate, there is some absolute constant C such that

$$n \equiv n_{(\mathcal{F}-\mathcal{F})^2}(\varepsilon,\delta) \leq \frac{c}{\varepsilon^2}\Big(\frac{1}{\varepsilon^2}\log^3\frac{1}{\varepsilon} + \log\frac{1}{\delta}\Big).$$

Thus, there is a sample $\{\omega_1,...,\omega_n\}$ such that if $f,g \in \mathcal{F}$ and if for every $1 \leq i \leq n$ $f(\omega_i) = g(\omega_i)$ then $\mathbb{E}_\mu(f-g)^2 < \varepsilon$. Therefore, the set $\{\delta_{\omega_1},...,\delta_{\omega_n}\}$ are ε sufficient statistics. Since such a sample exists for every $\delta \in (0,1)$, then

$$S_{\mathcal{F},\mu}(\varepsilon) \leq \liminf_{\delta\to 0} n_{(\mathcal{F}-\mathcal{F})^2}(\varepsilon,\delta) \leq \frac{c}{\varepsilon^4}\log^3\frac{1}{\varepsilon}.$$

Moreover, for every $\delta \in (0,1)$ the set of statistics is supported on the selected sample, hence, on a set of $n_{(\mathcal{F}-\mathcal{F})^2}(\varepsilon,\delta)$ elements at the most.

Let us compare this direct method with our approach. The beginning of the selection process is the same: select a sample $\mathcal{S}_n = \{\omega_1,...,\omega_n\}$ such that if μ_n is an empirical measure supported on $\mathcal{S}_n$ then for every $f,g \in \mathcal{F}$,

$$\big|\mathbb{E}_\mu(f-g)^2 - \mathbb{E}_{\mu_n}(f-g)^2\big| < \frac{\varepsilon}{2}. \tag{4.2}$$

Next, we construct ε sufficient statistics for $\mathcal{F}/\mu_n$. By Theorem 4.2 there is a random selection process which produces $m \leq \frac{C_2}{\varepsilon}\log^4 n$ linear empirical equations $(S_i)_1^m$ which are supported on $\mathcal{S}_n$ such that if $f,g \in \mathcal{F}$ and $S_i(f) = S_i(g)$ for every $1 \leq i \leq m$ then $\mathbb{E}_{\mu_n}(f-g)^2 < \varepsilon$. Since for every $\delta \in (0,1)$ n may be selected as $n_{(\mathcal{F}-\mathcal{F})^2}(\varepsilon,\delta)$, then up to a logarithmic factor in $\frac{1}{\delta}$,

$$m \leq \frac{C}{\varepsilon}\log^4\frac{1}{\varepsilon}.$$

Now, by (4.2) it follows that $(S_i)_1^m$ are ε sufficient statistics for $\mathcal{F}$ in $L_2(\mu)$. Thus,

$$S_{\mathcal{F},\mu} \leq \frac{C}{\varepsilon}\log^4\frac{1}{\varepsilon},$$

which is much better than the estimate obtained by the direct method.

Let us sum-up the selection scheme for a class $\mathcal{F}$: Fix the desired confidence and accuracy parameters.

1. Randomly select an i.i.d. sample $\{\omega_1,..,\omega_n\}$ according to μ, where $n = n_{(\mathcal{F}-\mathcal{F})^2}(\varepsilon/2,\delta/2)$.
2. Let m be as in Theorem 4.2 and assume that $m \geq C_1\log\frac{2}{\delta}$, where C_1 is the absolute constant as in Theorem 4.1. Set G be an $m \times n$ matrix whose entries are realizations of standard independent Gaussian random variables.
3. For every $1 \leq i \leq m$, let

$$S_i = \sum_{j=1}^n g_{ij}\delta_{\omega_j}.$$

Then, with probability larger than $1-\delta$, $(S_i)_1^m$ are ε sufficient statistics in $L_2(\mu)$.

Note that our result is even stronger than what we have claimed. Not only did we prove the existence of sufficient statistics, we where able to formulate a simple random construction scheme which produces ε sufficient statistics with arbitrarily large probability.

4.3 Improving the computational complexity

In this final application we indicate how constructing sufficient statistics in empirical L_2 spaces may aid in reducing the computational complexity of a learning problem.

Assume that h is the target concept and that μ_n is an empirical measure such that

$$\sup_{f\in\mathcal{F}} \left|\mathbb{E}_\mu (f-h)^2 - \mathbb{E}_{\mu_n}(f-h)^2\right| < \varepsilon \ .$$

Normally, when trying to approximate a function h with respect to the $L_2(\mu_n)$ norm, one tries to solve the system of n empirical linear equations $\delta_{\omega_i}(h) = \delta_{\omega_i}(f)$ (i.e. the equations $f(\omega_i) = h(\omega_i)$) subjected to the constraint that the solution belongs to $\mathcal{F}$. By using linear functionals on $L_2(\mu_n)$ which are linear combinations of the point evaluation functionals $\{\delta_{\omega_1}, ..., \delta_{\omega_n}\}$, it is enough to solve $S_{\mathcal{F},\mu_n}(\varepsilon) << n$ linear empirical equations with the same constraint to ensure that the solution approximates h in $L_2(\mu_n)$.

Below is a summary of the learning procedure together with complexity estimates in terms of the ℓ-norm. The proof of the claims in the example below are based on the same idea as in the proof of Theorem 4.2.

Example 4.5 *Let $\mathcal{F}$ be a class of functions on a set Ω, all of which have a range contained in $[0,1]$ and set $h \in \mathcal{F}$ to be the target concept. Let ε, δ be the accuracy and confidence parameters, set n to be the sample complexity estimate of $\mathcal{F}$ associated with an accuracy of ε and confidence of $\delta/2$, and put $\ell_n = \sup_{\mu_n} \ell(\mathcal{F}/\mu_n)$.*

1. *Select a sample $(\omega_1, ..., \omega_n)$ according to μ and let $\big(h(\omega_1), ..., h(\omega_n)\big)$ be the values of h on the sample.*
2. *Let $m = C\max\{\ell_n^2/\varepsilon, \log 2/\delta\}$, where C is some absolute constant, and put G to be an $m \times n$ matrix such that each element g_{ij} is a realization of a standard Gaussian random variable.*
3. *Find a solution $f \in \mathcal{F}$ to the system $\sum_1^n g_{ij}h(\omega_i) = \sum_1^n g_{ij}f(\omega_i)$ which consists of m empirical linear equations.*

Then, by the selection of m, $\|f-h\|^2_{L_2(\mu_n)} < \varepsilon$ with probability larger than $1-\delta/2$. Combining this with the selection of n it follows that with probability larger than $1-\delta$, $\|f-h\|^2_{L_2(\mu)} < \varepsilon$.

It is important to note that this learning procedure does not improve the sample complexity estimates. One has to start with an empirical measure for which

$$\sup_{f\in\mathcal{F}} \left|\mathbb{E}_\mu (f-h)^2 - \mathbb{E}_{\mu_n}(f-h)^2\right|$$

are "close", where h is the target concept. This is done by randomly selecting a sample according to μ, and the size of the sample is determined by the given accuracy and confidence parameters. On the other hand, the *computational* complexity improves. As an example, let $\mathcal{F}$ be a class of functions into $[0,1]$ such that for every $\varepsilon > 0$, $P_\varepsilon(\mathcal{F}) = O(\varepsilon^{-2})$. Given the accuracy and confidence parameters ε and δ, then $m = O(\varepsilon^{-1})$, while $n = O(\varepsilon^{-4})$ up to a logarithmic factor in $\frac{1}{\varepsilon}$ and $\frac{1}{\delta}$.

This learning rule may be adjusted to have a pre processing feature. Indeed, given $\varepsilon, \delta \in (0,1)$, if one selects $n = n_{(\mathcal{F}-\mathcal{F})^2}(\varepsilon, \delta)$ then the empirical functionals found here (which are determined by the Gaussian matrix G) do not depend on the target concept h. For every pair $f, h \in \mathcal{F}$, if $\sum_1^n g_{ij}h(\omega_i) = \sum_1^n g_{ij}f(\omega_i)$ for every $1 \le j \le m$, then with probability larger than $1-\delta$ $\|f-h\|^2_{L_2(\mu)} < \varepsilon$. The price one has to pay for this pre processing feature is a worse sample complexity estimate.

5 ℓ-norm estimates

This appendix is devoted to empirical ℓ-norm estimates of GC classes based on their VC or parametric Pollard dimension. Recall that in both these cases, there are known estimates for the covering numbers of $\mathcal{F}$: if $\mathcal{F}$ has a finite VC dimension then by Haussler's inequality (see [8] or [15]) its covering numbers in $L_2(\mu)$ are polynomial in $1/\varepsilon$ for every probability measure μ. Even when $\mathcal{F}$ does not have a finite VC dimension but its parametric Pollard dimension $P_\varepsilon(\mathcal{F})$ is polynomial in $1/\varepsilon$, then its log-covering numbers in $L_2(\mu_n)$ are polynomial in $1/\varepsilon$. We shall use those estimates to establish ℓ-norm estimates for the sets $\mathcal{F}/\mu_n$.

Let us recall Haussler's result:

Theorem 5.1 *Let $\mathcal{F}$ be a class of $\{0,1\}$ valued functions, such that $VC(\mathcal{F}) = d$. Then, there is an absolute constant C such that for every probability measure μ on Ω, $N(\varepsilon, \mathcal{F}, L_2(\mu)) \le Cd(4e)^d\varepsilon^{-2d}$.*

Using this estimate it is easy to derive the following:

Theorem 5.2 *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 $\ell(\mathcal{F}) \le Cd^{1/2}$.*

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

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

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 3.2,

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

and our claim follows. ∎

Next, Assume that $P_\varepsilon(\mathcal{F}) = O(\varepsilon^{-p})$ for some $p > 0$. The following estimate is due to Alon, Ben–David, Cesa–Bianchi and Haussler (see [1]).

Theorem 5.3 *Let $\mathcal{F}$ be a class of functions on Ω, all of which have a range contained in $[0,1]$ and set $d = P_{\varepsilon/4}(\mathcal{F})$. Then, for every empirical measure μ_n,*

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

We may apply the same idea used in the proof of Theorem 5.2 to classes which have a "small" parametric Pollard dimension.

Theorem 5.4 *Let $\mathcal{F}$ be a class of functions into $[0,1]$ such that $P_\varepsilon(\mathcal{F}) \le \gamma\varepsilon^{-p}$ for some $0 < p < 2$ and $\gamma \ge 1$. Then, there are constants C_p such that for every empirical measure μ_n,*

$$\ell(\mathcal{F}/\mu_n) \le C_p \gamma^{\frac{1}{2}}(1+\log n),$$

where $C_p = 2^p C \int_0^1 \frac{1}{\varepsilon^{p/2}} \log\frac{1}{\varepsilon} d\varepsilon$ for some absolute constant C.

Proof: By Theorem 5.3 it follows that there is some absolute constant C such that

$$\log N\big(\varepsilon, \mathcal{F}, L_2(\mu_n)\big) \le C\frac{4^p\gamma}{\varepsilon^p}\Big(1+\log^2\frac{n}{\varepsilon^2}\Big).$$

Since $\mathcal{F}$ is a subset of the unit ball of $L_2(\mu_n)$, then for every $\varepsilon \ge 1$ it takes only a single ball of cover $\mathcal{F}$. Thus, by Theorem 3.2,

$$\ell(\mathcal{F}/\mu_n) \le 2^p C\gamma^{\frac{1}{2}}(1+\log n)\int_0^1 \frac{1}{\varepsilon^{\frac{p}{2}}}\log\frac{1}{\varepsilon}d\varepsilon.$$

■

The case of $p \ge 2$ is much more difficult, because one can not use the upper bound in Theorem 3.2. However, it is possible to estimate the ℓ-norm, as described in the following Theorem:

Theorem 5.5 *Let $\mathcal{F}$ be a class of functions whose range is contained in $[0,1]$. Assume further that $P_\varepsilon(\mathcal{F}) \le \gamma\varepsilon^{-p}$ for some $p \ge 2$. Then, there is some absolute constant C, such that for every empirical measure μ_n,*

1. *if $p > 2$ then*

$$\ell(\mathcal{F}/\mu_n) \le C\gamma^{\frac{1}{2}}\alpha_p(1+\log n)(n^{\frac{1}{2}-\frac{1}{p}}-1)+n^{\frac{1}{2}-\frac{1}{p}},$$

where $\alpha_p = 2^{p/2}(2^{p/2-1}-1)^{-1}$, and,

2. *if $p = 2$ then*

$$\ell(\mathcal{F}/\mu_n) \le C(1+\gamma^{\frac{1}{2}})\log^2 n.$$

Although we can not apply the upper bound of theorem 3.2 directly, we shall use the same idea used in the proof of that Theorem.

Recall that $\mathcal{F}$ may be viewed as an subset of $L_2(\mu_n)$, where μ_n is an empirical measure supported on the sample $\{\omega_1,...,\omega_n\}$. Each $f \in \mathcal{F}$ is identified as an element of $L_2(\mu_n)$ (which is denoted by f/μ_n) by the map $T(f) = \sum_{i=1}^n f(\omega_i)\chi_{\omega_i}$, where χ_{ω_i} is the characteristic function of $\{\omega_i\}$. In terms of the orthonormal basis of $L_2(\mu_n)$, $f/\mu_n = n^{-1/2}\sum_{i=1}^n f(\omega_i)e_i$. Let $(g_i)_{i=1}^n$ be independent standard Gaussian random variables. For every $f \in \mathcal{F}$, let $Z_f = n^{-1/2}\sum_{i=1}^n f(\omega_i)g_i$. Thus, each Z_f is a random variable on some probability space (Y,P), and denote by $\|\ \|_2$ the norm in $L_2(Y,P)$. From the definition of the ℓ-norm it is easy to see that $\ell(\mathcal{F}/\mu_n) = \big\|\sup_{f\in\mathcal{F}} Z_f\big\|_2$. It is possible to show (see, for example, [12]) that there is some absolute constant $C > 0$ such that

$$\ell(\mathcal{F}/\mu_n) \le C\mathbb{E}\big|\sup_{f\in\mathcal{F}} Z_f\big| = \mathbb{E}(\sup_{f\in\mathcal{F}\cup-\mathcal{F}} Z_f).$$

Also, note that the map $V : L_2(\mu_n) \to L_2(Y,P)$ given by $V(\sum_{i=1}^n a_i e_i) = \sum_{i=1}^n a_i g_i$ is an isometry into $L_2(Y,P)$. Thus, for every $f \in \mathcal{F}$, $Z_f = V(f/\mu_n)$.

The following Lemma plays a crucial part in the proof of the upper bound in Theorem 3.2. It is based on the classical inequality of Slepian (see [12] or [5]).

Lemma 5.6 *Let $\{Z_1,...,Z_N\}$ be Gaussian random variables. Then, there is some absolute constant C such that*

$$\mathbb{E}\sup_i Z_i \le C\sup_{i,j}\|Z_i - Z_j\|_2 \log^{\frac{1}{2}} N.$$

Proof of Theorem 5.5: We will assume that $\mathcal{F}$ is symmetric. The proof in the non-symmetric case is essentially the same. Set $\mathcal{Z}_\mathcal{F} = \{Z_f | f \in \mathcal{F}\}$ and note that since $V : L_2(\mu_n) \to L_2(Y,P)$ is an isometry for which $V(\mathcal{F}/\mu_n) = \mathcal{Z}_\mathcal{F}$ then

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

Therefore, by Theorem 5.3 and since $P_\varepsilon(\mathcal{F}) \le \gamma\varepsilon^{-p}$, there is some absolute constant C such that

$$\log N(\varepsilon, \mathcal{Z}_\mathcal{F}) \le C\Big(1+4^p\gamma\varepsilon^{-p}\log^2\frac{n}{\varepsilon^2}\Big).$$

Let $\varepsilon_k = 2^{-k}$, put $N = [p^{-1}\log_2 n]$ and set $H_k \subset \mathcal{Z}_\mathcal{F}$ to be a $2\varepsilon_k$ cover of $\mathcal{Z}_\mathcal{F}$, such that

$$\log|H_k| \le C\Big(1+4^p\gamma\varepsilon_k^{-p}\log^2\frac{n}{\varepsilon_k^2}\Big).$$

Hence, for every k and every Z_f there is some $Z_f^k \in H_k$ such that $\big\|Z_f - Z_f^k\big\|_2 \le 2\varepsilon_k$. By writing

$$Z_f = \sum_{k=1}^N (Z_f^k - Z_f^{k-1}) + Z_f - Z_f^N$$

it follows that

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

By the definition of Z_f^k and by Lemma 5.6, there is an absolute constant C such that

$$\mathbb{E}\sup_{f\in\mathcal{F}}(Z_f^k - Z_f^{k-1}) \le$$

$$\mathbb{E}\sup\big\{Z_i - Z_j | Z_i \in H_k, Z_j \in H_{k-1}, \|Z_i - Z_j\|_2 \le 4\varepsilon_k\big\} \le$$

$$C\sup_{i,j}\|Z_i - Z_j\|_2 \log^{\frac{1}{2}}|H_k|\,|H_{k-1}| \le$$

$$C\varepsilon_k\Big(1+2^p\gamma^{\frac{1}{2}}\varepsilon_k^{-\frac{p}{2}}\log\frac{n}{\varepsilon_k^2}\Big).$$

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

$$\Big(\sum_{i=1}^{n} \frac{f(\omega_i) - f'(\omega_i)}{\sqrt{n}}^2\Big)^{\frac{1}{2}} =$$
$$\|f/\mu_n - f'/\mu_n\|_{L_2(\mu_n)} = \|Z_f - Z_{f'}\|_2 \leq \varepsilon_N,$$

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

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

Therefore,

$$\mathbb{E} \sup_{f \in \mathcal{F}} Z_f - Z_f^N \leq \varepsilon_N \mathbb{E}\Big(\sum_{i=1}^{n} g_i^2\Big)^{\frac{1}{2}} = \varepsilon_N \sqrt{n}.$$

Combining the two estimates and since $\varepsilon_k = 2^{-k}$ and $N = [p^{-1} \log_2 n]$,

$$\sup_{f \in \mathcal{F} \cup -\mathcal{F}} Z_f \leq$$
$$C \sum_{k=1}^{N} 2^{-k} \Big(1 + 2^p \gamma^{\frac{1}{2}} 2^{\frac{kp}{2}} \log 4^k n\Big) + 2^{-N} \sqrt{n} \leq$$
$$C\big(1 + 2^p \gamma^{\frac{1}{2}} \log 2^{2N} n\big) \sum_{k=1}^{N} 2^{(-1+\frac{p}{2})k} + 2^{-N} \sqrt{n} \leq$$
$$C\big(1 + 2^p \gamma^{\frac{1}{2}} \log n\big) c_p (n^{\frac{1}{2} - \frac{1}{p}} - 1) + n^{\frac{1}{2} - \frac{1}{p}},$$

where $c_p = 2^{p/2-1}(2^{p/2-1} - 1)^{-1}$ and the claim follows.

■

Remark 1 *Using a similar argument, it is possible to show that if $P_\varepsilon(\mathcal{F}) \leq \gamma \varepsilon^{-2}$ then there is some absolute constant C such that for every empirical measure μ_n,*

$$\ell(\mathcal{F}/\mu_n) \leq C(\gamma^{\frac{1}{2}} + 1) \log^2 n.$$

References

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

[2] P. Assouad, R.M. Dudley: Minmax nonparametric estimation over classes of sets, Preprint.

[3] G. Darmois: Sur les limits de la dispersion de certains estimations, Rev. Inst. intern. Statist., 13,9-15, 1945.

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

[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. of Theoretical Probability, 4, 485-510, 1991

[7] 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*, Lecture notes in Mathematics 1317, 84-106.

[8] D. Haussler: Sphere packing numbers for subsets of Boolean n-cube with bounded Vapnik-Chervonenkis dimension, Journal of Combinatorial Theory A 69, 217-232

[9] B.O. Koopman: On Distributions admitting a sufficient statistic, Trans. Am. Math. Soc., 39, 399-409, 1936.

[10] S. Mendelson: ℓ-norm and its application to Learning Theory, To appear in Positivity.

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

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

[13] E.J.G. Pitman: Sufficient statistics and intrinsic accuracy, Proc. Camb. Phil. Soc., 32, 567-579, 1936.

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

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

[16] V. Vapnik, A. Chervonenkis: Necessary and sufficient conditions for uniform convergence of means to mathematical expectations, Theory Prob. Applic. 26(3), 532-553, 1971