Entropy, Combinatorial Dimensions and Random Averages

Shahar Mendelson[1] and Roman Vershynin[2]

[1] RSISE, The Australian National University
Canberra, ACT 0200, Australia
shahar.mendelson@anu.edu.au
[2] Department of Mathematical Sciences, University of Alberta
Edmonton, Alberta T6G 2G1, Canada
Vershynin@yahoo.com

Abstract. In this article we introduce a new combinatorial parameter which generalizes the VC dimension and the fat-shattering dimension, and extends beyond the function-class setup. Using this parameter we establish entropy bounds for subsets of the n-dimensional unit cube, and in particular, we present new bounds on the empirical covering numbers and gaussian averages associated with classes of functions in terms of the fat-shattering dimension.

1 Introduction

Empirical entropy estimates play an important role in Machine Learning since they are one of the only ways in which one can control the "size" of the function class one is interested in, and thus obtain sample complexity estimates.

Thanks to the development of the theory of empirical processes in recent years, entropy bounds are no longer the best way to obtain sample complexity estimates. Rather, the notion of random averages (such as the Rademacher or gaussain complexities) seems to be a better way to obtain generalization results. However, entropy bounds are still extremely important, since they are almost the only way in which one can establish bounds on the random complexities. Thus, finding ways to control the empirical entropy of function classes remains an interesting problem.

In an attempt to tackle this issue, several combinatorial parameters were introduced, in the hope that using them, one would be able to bound the entropy.

This approach was pioneered by the work of Vapnik and Chervonenkis [17] who introduced the VC dimension to obtain bounds on the L_∞ empirical entropy of Boolean classes of functions. Other results regarding the empirical entropy of Boolean classes with respect to empirical L_p norms for $1 \leq p < \infty$ were established by Dudley (see [8]) and then improved by Haussler [6].

Extending these results to classes of real-valued, uniformly bounded functions is very difficult. To that end, the notion of the *fat-shattering dimension*

J. Kivinen and R. H. Sloan (Eds.): COLT 2002, LNAI 2375, pp. 14–28, 2002.

was introduced and in [1] the authors presented empirical L_∞ bounds and generalization results using this parameter. Only recently, in [9], improved empirical entropy bounds were established for L_p spaces, when $1 \leq p < \infty$.

The motivation for this article is the fact that the combinatorial parameters used in Machine Learning literature were also used to tackle problems in Convex Geometry [13]. Our original aim was to study a new combinatorial parameter which would be suitable for the analysis of convex bodies in $\mathbb{R}^n$, but our methods yield improved entropy bounds in terms of the fat-shattering dimension as well.

All the results presented here are based on new entropy estimates for subsets of the unit cube in $\mathbb{R}^n$, which is denoted by B_∞^n. For example, we show that if $K \subset B_\infty^n$ is convex, then its controlled by its *generalized Vapnik-Chervonenkis dimension*. This parameter, denoted by $\mathrm{VC}(A,t)$, is defined for every $0 < t < 1$ as the maximal size of a subset σ of $\{1,\ldots,n\}$, such that the coordinate projection of K onto $\mathbb{R}^\sigma$ contains a coordinate cube of the form $x + [0,t]^\sigma$. This notion carries over to convexity the "classical" concept of the VC dimension, denoted by $\mathrm{VC}(A)$, and defined for subsets A of the discrete cube $\{0,1\}^n$ as the maximal size of the subset σ of $\{1,\ldots,n\}$ such that $P_\sigma A = \{0,1\}^\sigma$, where P_σ is the coordinate projection onto the coordinates in σ (see [8] §14.3).

To see how this relates to function classes, assume that F consists of functions which are all bounded by 1. For every sample $s_n = \{x_1,...,x_n\}$ let $\mu_n = n^{-1}\sum_{i=1}^n \delta_{x_i}$, where δ_x is the point evaluation functional at x. Set $F/s_n = \left\{\left(f(x_i)\right)_{i=1}^n | f \in F\right\}$ and note that $F/s_n \subset B_\infty^n$. One can see that the $L_p(\mu_n)$ covering numbers of F at scale t are simply the number of translates of $tn^{1/p}B_p^n$ needed to cover F/s_n, where $B_p^n = \{(a_i)_{i=1}^n | \sum_{i=1}^n |a_i|^p \leq 1\}$. Thus, by investigating the structure of the sets F/s_n, the problem of estimating the empirical entropy numbers is reduced to the analysis of the entropy of subsets of B_∞^n.

If K is a convex body then a volumetric bound on the entropy shows that for every $0 < t \leq 1$,

$$\log N(K, n^{1/p}B_p^n, t) \leq \log(5/t) \cdot n,$$

where $N(K,B,t)$ is the covering number of K at scale t, that is, the number of translates of tB needed to cover K.

One question is whether it is possible to replace the dimension n on the right-hand side of this estimate by the VC dimension $\mathrm{VC}(K,ct)$, which is generally smaller? This is perfectly true for the *Boolean* cube: the known theorem of R. Dudley that lead to a characterization of the uniform central limit property in the Boolean case states that if $A \subset \{0,1\}^n$ then

$$\log N(A, n^{1/2}B_2^n, t) \leq C \log(2/t) \cdot \mathrm{VC}(A).$$

This estimate follows by a random choice of coordinates and an application of the Sauer-Shelah Lemma (see [8] Theorem 14.12). The same problem for convex bodies is considerably more difficult, since in that case, one needs to find a cube in $P_\sigma K$ with well separated faces, not merely disjoint. The same difficulty occurs in the case of classes of real-valued functions. We prove the following theorem.

Theorem 1. *There are absolute constants $C, c > 0$ such that for every convex body $K \subset B_\infty^n$, every $1 \le p < \infty$ and any $0 < t < 1$,*

$$\log N(K, n^{1/p} B_p^n, t) \le C p^2 v \log^2(2/t), \tag{1}$$

and

$$\log N(K, B_\infty^n, t) \le C v \log^2\big(\frac{n}{tv}\big), \tag{2}$$

where $v = \mathrm{VC}(K, ct)$.

Observe that (2) is best complemented by the lower bound $\log N(K, B_\infty^n, t) \ge \mathrm{VC}(K, ct)$, for some absolute constant $c > 0$, which follows from the definition of the generalized VC dimension and a comparison of volumes.

A very similar theorem can be formulated for general subsets of the unit cube, which leads to this next result regarding function classes.

Theorem 2. *There are absolute constants c and C such that for every $0 < t < 1$, every $1 \le p < \infty$ and every empirical measure μ_n,*

$$\log N(F, L_p(\mu_n), t) \le C p^2 v \log^2(2/t) \tag{3}$$

and

$$\log N(F, L_\infty(s_n), t) \le C v \log^2\big(\frac{n}{tv}\big), \tag{4}$$

where $v = \mathrm{fat}_{ct}(F)$ and s_n is the sample on which μ_n is supported.

The main difference between (3) and (4) is that the first is dimension free, in the sense that is does not depend on the cardinality of the set on which the sample is supported. Note that (3) is linear in the fat-shattering dimension. One can also show that both these bounds are optimal up to the power of the logarithmic factor. For example, a volumetric bound show that for (3) one needs at least a factor of $\log(2/t)$.

These two theorems show that the $\|\cdot\|_\infty$-entropy of a subset of B_∞^n is governed by the appropriate combinatorial parameter, up to a logarithmic factor in $1/t$.

The other main result we present, deals with estimating the gaussian averages associated with a class of functions. Given a class F and a sample $s_n = \{x_1, ..., x_n\}$, let $E(s_n) = \mathbb{E} \sup_{f \in F} |\sum_{i=1}^n g_i f(x_i)|$, where $(g_i)_{i=1}^n$ are independent standard gaussian random variables.

There are many results which show that the gaussian averages (and likewise, the Rademacher averages) play an important part in the quest for generalization bounds [10,11], hence their importance in the Machine Learning context.

It is easy to see that $E(s_n) = \mathbb{E} \sup_{f \in F} |\langle f/s_n, \sum_{i=1}^n g_i e_i \rangle|$, where $f/s_n = \big(f(x_1), ...f(x_n)\big)$ and $(e_i)_{i=1}^n$ is the standard basis in the n-dimensional inner product space ℓ_2^n.

This naturally leads to the use of the polar of a subset of ℓ_2^n; if F is a bounded subset of ℓ_2^n, let the *polar* of F be $F^\circ = \{x \in \ell_2^n | \sup_{f \in F} |\langle f, x \rangle| \le 1\}$. Let K be the symmetric convex hull of F. As a convex and symmetric set, it is the unit ball of a norm denoted by $\| \ \|_K$. It is easy to see that F° is the unit ball

of the norm dual to $\| \ \|_K$, and for every $x \in \mathbb{R}^n$, $\|x\|_{F^\circ} = \sup_{f \in F} \langle f, x \rangle$. Thus, $E(s_n) = \mathbb{E}\| \sum_{i=1}^n g_i e_i \|_{(F/s_n)^\circ}$.

We shall present bounds on $E = \mathbb{E}\| \sum_{i=1}^n g_i e_i \|_{F^\circ}$ for arbitrary subsets $F \subset B_\infty^n$ in terms of their generalized VC dimension or their fat-shattering dimension.

Both are relatively easy once we know (1) or (3). Indeed, replacing the entropy by the VC dimension in Dudley's entropy inequality it follows that there are absolute constants C and c such that

$$E \le C \int_{cE/\sqrt{n}}^{\infty} \sqrt{\log N(K, B_2^n, t)}\, dt \le C\sqrt{n} \int_{cE/n}^{1} \sqrt{\mathrm{VC}(K, ct)} \log(2/t)\, dt. \quad (5)$$

and

$$E(s_n) \le C\sqrt{n} \int_{cE(s_n)/n}^{\sigma_F(s_n)} \sqrt{\mathrm{fat}_{ct}(F)} \log(2/t)\, dt, \quad (6)$$

where $\sigma_F^2(s_n) = \sup_{f \in F} n^{-1} \sum_{i=1}^n f^2(x_i)$.

Inequality (5) improves the main theorem of M. Talagrand in [15]. An additional application which follows from (3) is that if F is a class of uniformly bounded functions which has a relatively small fat-shattering dimension, then it satisfies the uniform central limit theorem. This extends Dudley's characterization for VC classes to the real-valued case.

The paper is organized as follows; in Section 2 we prove the bound for the B_p^n-entropy in abstract finite product spaces, and then derive (1) by approximation. In Section 3 we apply (1) to prove (3) and obtain the uniform CLT result. In Section 4 we obtain the bounds on the gaussian complexities of a class. Finally, in Section 5 we prove (2) for the B_∞^n-entropy by reducing it to (1) through an independent lemma that compares the B_p^n-entropy to the B_∞^n-entropy.

Throughout this article, positive absolute constants are denoted by C and c. Their values may change from line to line, or even within the same line.

2 B_p^n-Entropy in Abstract Product Spaces

We will introduce and work with the notion of the VC dimension in an abstract setting that encompasses the classes considered in the introduction.

We call a map $d : T \times T \to \mathbb{R}_+$ a *quasi-metric* if d is symmetric and reflexive (that is, $\forall x, y$, $d(x, y) = d(y, x)$ and $d(x, x) = 0$). We say that points x and y in T are separated if $d(x, y) > 0$. Thus, d does not necessarily separate points or satisfy the triangle inequality.

Definition 1. *Let (T, d) be a quasi-metric space and let n be a positive integer. For a set $A \subset T^n$ and $t > 0$, the VC-dimension $\mathrm{VC}(A, t)$ is the maximal cardinality of a subset $\sigma \subset \{1, \ldots, n\}$ such that the inclusion*

$$P_\sigma A \supseteq \prod_{i \in \sigma} \{a_i, b_i\} \quad (7)$$

holds for some points $a_i, b_i \in T$, $i \in \sigma$ with $d(a_i, b_i) \ge t$. If no such σ exists, we set $\mathrm{VC}(A, t) = 0$. When there is a need to specify the underlying metric, we denote the VC dimension by $\mathrm{VC}_d(A, t)$.

Since $\mathrm{VC}(A,t)$ is decreasing in t and is bounded by n, which is the "usual" dimension of the product space, the limit $\mathrm{VC}(A) := \lim_{t\to 0_+} \mathrm{VC}(A,t)$ always exists. Equivalently, $\mathrm{VC}(A)$ is the maximal cardinality of a subset $\sigma \subset \{1,\dots,n\}$ such that (7) holds for some pairs (a_i,b_i) of separated points in T.

This definition is an extension of the "classical" VC dimension for subsets of the discrete cube $\{0,1\}^n$, where we think of $\{0,1\}$ as a metric space with the $0-1$ metric. Clearly, for any set $A \subset \{0,1\}^n$ the quantity $\mathrm{VC}(A,t)$ does not depend on $0<t<1$, and hence $\mathrm{VC}(A) = \max\Big\{|\sigma| : \ \sigma \subset \{1,\dots,n\},\ P_\sigma A = \{0,1\}^\sigma\Big\}$, which is precisely the "classical" definition of the VC dimension.

The other example discussed in the introduction was the VC dimension of convex bodies. Here $T=\mathbb{R}$ or, more frequently, $T=[-1,1]$, both with respect to the usual metric. If $K \subset T^n$ is a convex body, then $\mathrm{VC}(K,t)$ is the maximal cardinality of a subset $\sigma \subset \{1,\dots,n\}$ for which the inclusion $P_\sigma K \supseteq x+(t/2)B_\infty^\sigma$ holds for some vector $x \in \mathbb{R}^\sigma$ (which automatically lies in $P_\sigma K$). It is easy to see that if K is symmetric, we can set $x=0$. Also note that for every convex body (that is a convex, symmetric subset of $\mathbb{R}^n$ with a nonempty interior) $\mathrm{VC}(K)=n$.

The generalized VC dimension may be controlled by the fat-shattering dimension, in the following sense. Assume that F is a subset of the unit ball in $L_\infty(\Omega)$, which is denoted by $B\big(L_\infty(\Omega)\big)$. Let $s_n=\{x_1,...,x_n\}$ be a subset of Ω and set $F/s_n = \big\{\big(f(x_1),...,f(x_n)\big)\big|f\in F\big\} \subset \mathbb{R}^n$. If $\mathrm{VC}(F/s_n,t)=m$, there is a subset $\sigma \subset \{1,...,n\}$ of cardinality m such that $P_\sigma F/s_n \supset \prod_{i\in\sigma}\{a_i,b_i\}$ where $|b_i-a_i| \ge t$. By selecting $s(x_i)=(b_i+a_i)/2$ as the witness to the shattering (see [2]), it is clear that $(x_i)_{i\in\sigma}$ is $t/2$-shattered by F, and thus $VC(F/s_n,t) \le \mathrm{fat}_{t/2}(F)$.

The main results of this article rely on (and are easily reduced to) a discrete problem: to estimate the generalized VC-dimension of a set in a product space T^n, where (T,d) is a *finite* quasi-metric space. T^n is usually endowed with the normalized Hamming quasi-metric $d_n(x,y) = n^{-1}\sum_{i=1}^n d(x(i),y(i))$ for $x,y\in T^n$.

In this section we bound the entropy of a set $A\subset T^n$ with respect to d_n in terms of $\mathrm{VC}(A)$.

Theorem 3. *Let (T,d) be a finite quasi-metric space with* $\mathrm{diam}(T)\le 1$, *and set n to be a positive integer. Then, for every set $A\subset T^n$ and every $0<\varepsilon<1$,*

$$\log N(A,d_n,\varepsilon) \le C\log^2(|T|/\varepsilon)\cdot \mathrm{VC}(A),$$

where C is an absolute constant.

Before presenting the proof, let us make two standard observations. We say that points $x,y\in T^n$ are separated on the coordinate i_0 if $x(i_0)$ and $y(i_0)$ are separated. Points x and y are called ε-separated if $d_n(x,y)\ge\varepsilon$.

Clearly, if A' is a maximal ε-separated subset of A then $|A'|\ge N(A,d_n,\varepsilon)$. Moreover, the definition of d_n and the fact that $\mathrm{diam}(T)\le 1$ imply that every two distinct points in A' are separated on at least εn coordinates. This shows that Theorem 3 is a consequence of the following statement.

Theorem 4. *Let (T, d) be a quasi-metric space for which* $\mathrm{diam}(T) \leq 1$. *Let $0 < \varepsilon < 1$ and consider a set $A \subset T^n$ such that every two distinct points in A are separated on at least εn coordinates. Then*

$$\log |A| \leq C \log^2(|T|/\varepsilon) \cdot \mathrm{VC}(A). \tag{8}$$

The first step in the proof of Theorem 4 is a probabilistic extraction principle, which allows one to reduce the number of coordinates without changing the separation assumption by much. Its proof is based on a simple discrepancy bound for a set system.

Lemma 1. *There exists an absolute constant $c > 0$ for which the following holds. Let $\varepsilon > 0$ and assume that $\mathcal{S}$ is a system of subsets of $\{1, \ldots, n\}$ which satisfies that each $S \in \mathcal{S}$ contains at least εn elements. Let $k \leq n$ be an integer such that $\log |\mathcal{S}| \leq c\varepsilon k$. Then there exists a subset $I \subset \{1, \ldots, n\}$ of cardinality $|I| = k$, such that*

$$|I \cap S| \geq \varepsilon k/4 \quad \textit{for all } S \in \mathcal{S}.$$

Proof. If $|\mathcal{S}| = 1$ the lemma is trivially true, hence we may assume that $|\mathcal{S}| \geq 2$. Let $0 < \delta < 1/2$ and set $\delta_1, \ldots, \delta_n$ to be $\{0, 1\}$-valued independent random variables with $\mathbb{E}\delta_i = \delta$ for all i. By the classical bounds on the tails of the binomial law (see [7], or [8] 6.3 for more general inequalities), there is an absolute constant $c_0 > 0$ for which

$$\mathbb{P}\Big\{ \Big| \sum_{i=1}^{n} (\delta_i - \delta) \Big| > \frac{1}{2}\delta n \Big\} \leq 2\exp(-c_0 \delta n). \tag{9}$$

Let $\delta = k/2n$ and consider the random set $I = \{i : \ \delta_i = 1\}$. For any set $B \subset \{1, \ldots, n\}$, $|I \cap B| = \sum_{i \in B} \delta_i$. Then (9) implies that

$$\mathbb{P}\{|I \cap B| \geq \delta |B|/2\} \geq 1 - 2\exp(-c_0 \delta |B|).$$

Since for every $S \in \mathcal{S}$, $|S| > \varepsilon n$, then $\mathbb{P}\{|I \cap S| \geq \varepsilon k/4\} \geq 1 - 2\exp(-c_0 \varepsilon k/2)$. Therefore, $\mathbb{P}\big\{\forall S \in \mathcal{S}, \ |I \cap S| \geq \varepsilon k/4\big\} \geq 1 - 2|\mathcal{S}| \exp(-c_0 \varepsilon k/2)$. By the assumption on k, this quantity is larger than $1/2$ (with an appropriately chosen absolute constant c). Moreover, by a similar argument, $|I| \leq k$ with probability larger than $1/2$. This proves the existence of a set I satisfying the assumptions of the lemma. □

Proof of Theorem 4. We may assume that $|T| \geq 2$, $\varepsilon \leq 1/2$, $n \geq 2$ and $\max(4, \exp(4c)) \leq |A| \leq |T|^n$, where $0 < c < 1$ is the constant in Lemma 1. The first step in the proof is to use previous lemma, which enables one to make the additional assumption that $\log |A| \geq c\varepsilon n/4$. Indeed, assume that the converse inequality holds, and for every pair of distinct points $x, y \in A$, let $S(x, y) \subset \{1, \ldots, n\}$ be the set of coordinates on which x and y are separated. Put $\mathcal{S}$ to be the collection of the sets $S(x, y)$ and let k be the minimal positive integer for which $\log |\mathcal{S}| \leq c\varepsilon k$. Since $|A| \leq |\mathcal{S}| \leq |A|^2$, then

$$c\varepsilon(k-1) \leq \log |\mathcal{S}| \leq 2 \log |A| \leq \frac{1}{2} c\varepsilon n,$$

which implies that $1 \le k \le n$. Thus, by Lemma 1 there is a set $I \subset \{1, \dots, n\}$, $|I| = k$, with the property that every pair of distinct points $x, y \in A$ is separated on at least $\varepsilon|I|/4$ coordinates in I. Also, since $4c \le \log|A| \le \log|\mathcal{S}| \le c\varepsilon k$, then $\varepsilon|I|/4 \ge 1$ and thus $|P_I A| = |A|$. Clearly, to prove the assertion of the theorem for the set $A \subset T^n$, it is sufficient to prove it for the set $P_I A \subset T^I$ (with $|I|$ instead of n), whose cardinality already satisfies $\log|P_I A| = \log|A| \ge c\varepsilon(k-1)/2 \ge c\varepsilon|I|/4$. Therefore, we can assume that $|A| = \exp(\alpha n)$ with $\alpha > c\varepsilon$ for some absolute constant c.

The next step in the proof is a counting argument, which is based on the proof of Lemma 3.3 in [1] (see also [3]).

A set is called a *cube* if it is of the form $D_\sigma = \prod_{i\in\sigma}\{a_i, b_i\}$, where σ is a subset of $\{1, \dots, n\}$ and $a_i, b_i \in T$. We will be interested only in *large cubes*, which are the cubes in which a_i and b_i are separated for all $i \in \sigma$. Given a set $B \subset T^n$, we say that a cube D_σ *embeds* into B if $D_\sigma \subset P_\sigma B$. Note that if a large cube D_σ with $|\sigma| \ge v$ embeds into B then $\mathrm{VC}(B) \ge v$.

For all $m \ge 2$, $n \ge 1$ and $0 < \varepsilon \le 1/2$, let $t_\varepsilon(m, n)$ denote the maximal number t such that for every set $B \subset T^n$, $|B| = m$, which satisfies the separation condition we imposed (that is, every distinct points $x, y \in B$ are separated on at least εn coordinates), there exist t large cubes that embed into B. If no such B exists, we set $t_\varepsilon(m, n)$ to be infinite. The number of possible large cubes D_σ for $|\sigma| \le v$ is smaller than $\sum_{k=1}^{v}\binom{n}{k}|T|^{2k}$, as for every σ of cardinality k there are less than $|T|^{2k}$ possibilities to choose D_σ. Therefore, if $t_\varepsilon(|A|, n) \ge \sum_{k=1}^{v}\binom{n}{k}|T|^{2k}$, there exists a large cube D_σ for some $|\sigma| \ge v$ that embeds into A, implying that $\mathrm{VC}(A) \ge v$. Thus, to prove the theorem, it suffices to estimate $t_\varepsilon(m, n)$ from below. To that end, we will show that for every $n \ge 2$, $m \ge 1$ and $0 < \varepsilon \le 1/2$,

$$t_\varepsilon(2m \cdot |T|^2/\varepsilon, n) \ge 2t_\varepsilon(2m, n-1). \tag{10}$$

Indeed, fix any set $B \subset T^n$ of cardinality $|B| = 2m \cdot |T|^2/\varepsilon$, which satisfies the separation condition above. If no such B exists then $t_\varepsilon(2m \cdot |T|^2/\varepsilon, n) = \infty$, and (10) holds trivially. Split B arbitrarily into $m \cdot |T|^2/\varepsilon$ pairs, and denote the set of the pairs by $\mathcal{P}$. For each pair $(x, y) \in \mathcal{P}$ let $I(x, y) \subset \{1, \dots, n\}$ be the set of the coordinates on which x and y are separated, and note that by the separation condition, $|I(x, y)| \ge \varepsilon n$.

Let i_0 be the random coordinate, that is, a random variable uniformly distributed in $\{1, \dots, n\}$. The expected number of the pairs $(x, y) \in \mathcal{P}$ for which $i_0 \in I(x, y)$ is

$$\mathbb{E}\sum_{(x,y)\in\mathcal{P}} \mathbf{1}_{\{i_0 \in I(x,y)\}} = \sum_{(x,y)\in\mathcal{P}} \mathbb{P}\{i_0 \in I(x, y)\} \ge |\mathcal{P}| \cdot \varepsilon = m|T|^2.$$

Hence, there is a coordinate i_0 on which at least $m|T|^2$ pairs $(x, y) \in \mathcal{P}$ are separated. By the pigeonhole principle, there are at least $m|T|^2/\binom{|T|}{2} \ge 2m$ pairs $(x, y) \in \mathcal{P}$ for which the (unordered) set $\{x(i_0), y(i_0)\}$ is the same.

Let $I = \{1, \dots, n\} \setminus \{i_0\}$. It follows that there are two subsets of B, denoted by B_1 and B_2, such that $|B_1| = |B_2| = 2m$ and

$$B_1 \subset \{b_1\} \times T^I, \quad B_2 \subset \{b_2\} \times T^I$$

for some separated points $b_1, b_2 \in T$. Clearly, the set B_1 satisfies the separation condition and so does B_2. It is also clear that if a large cube D_σ embeds into B_1, then it also embeds into B, and the same holds for B_2. Moreover, if the same cube D_σ embeds into both B_1 and B_2, then the large cube $\{b_1, b_2\} \times D_\sigma$ embeds into B (since $\{b_1, b_2\} \times D_\sigma \subset P_{\{i_0\}\cup\sigma}B$). Therefore, $t_\varepsilon(|B|, n) \geq 2t_{\frac{\varepsilon n}{n-1}}(|B_1|, n-1) \geq 2t_\varepsilon(|B_1|, n-1)$, establishing (10).

Since $t_\varepsilon(2, n) \geq 1$, an induction argument yields that $t_\varepsilon(2(|T|^2/\varepsilon)^r, n) \geq 2^r$ for every $r \geq 1$. Thus, for every $m \geq 4$

$$t_\varepsilon(m, n) \geq m^{\frac{1}{2\log(|T|^2/\varepsilon)}}.$$

(It is remarkable that the right hand side does not depend on n). Therefore, $\mathrm{VC}(A) \geq v$ provided that v satisfies

$$t_\varepsilon(|A|, n) \geq \exp\left(\frac{\alpha n}{2\log(|T|^2/\varepsilon)}\right) \geq \sum_{k=1}^{v} \binom{n}{k} |T|^{2k}. \tag{11}$$

To estimate v, one can bound the right-hand side of (11) using Stirling's approximation $\sum_{k=1}^{v} \binom{n}{k} \leq [\gamma^\gamma(1-\gamma)^{1-\gamma}]^{-n}$, where $\gamma = v/n \leq 1/2$. It follows that for $v \leq n/2$, $\sum_{k=1}^{v} \binom{n}{k}|T|^{2k} \leq (\frac{|T|n}{v})^{2v}$. Taking logarithms in (11), we seek integers $v \leq n/2$ satisfying that

$$\frac{\alpha n}{2\log(|T|^2/\varepsilon)} \geq 2v \log\left(\frac{|T|n}{v}\right).$$

This holds if

$$v \leq \cdot \left(\frac{\alpha n}{\log(|T|^2/\varepsilon)}\right) \Big/ 8\log\left(\frac{4|T|\log(|T|^2/\varepsilon)}{\alpha}\right),$$

proving our assertion since $\alpha > c\varepsilon$. □

Corollary 1. *Let $n \geq 2$ and $p \geq 2$ be integers, set $0 < \varepsilon < 1$ and $q > 0$. Consider a set $A \subset \{1, \ldots, p\}^n$ such that for every two distinct points $x, y \in A$, $|x(i) - y(i)| \geq q$ for at least εn coordinates i. Then*

$$\log|A| \leq C\log^2(p/\varepsilon) \cdot \mathrm{VC}(A, q).$$

Proof. We can assume that $q \geq 1$. Define the following quasi-metric on $T = \{1, \ldots, p\}$:

$$d(a, b) = \begin{cases} 0 & \text{if } |a-b| < q, \\ 1 & \text{otherwise.} \end{cases}$$

Then $N(A, d_n, \varepsilon) = |A|$. By Theorem 3, $\log|A| \leq C\log^2(p/\varepsilon) \cdot \mathrm{VC}_d(A)$, which completes the proof by the definition of the metric d. □

Now we pass from the discrete setting to the "continuous" one - namely, we study subsets of B_∞^n. Recall that the Minkowski sum of two convex bodies $A, B \subset \mathbb{R}^n$ is defined as $A + B = \{a + b |\ a \in A,\ b \in B\}$.

Corollary 2. *For every $A \subset B_\infty^n$, $0 < t < 1$ and $0 < \varepsilon < 1$,*

$$\log N(A, \sqrt{n}B_2^n, t) \le C \log^2(2/t\varepsilon) \cdot \mathrm{VC}(A + \varepsilon B_\infty^n, t/2).$$

Proof. Clearly, we may assume that $\varepsilon \le t/4$. Put $p = \frac{1}{2\varepsilon}$ and let

$$T = \{-2\varepsilon p, -2\varepsilon(p-1), \dots, -2\varepsilon, 0, 2\varepsilon, \dots, 2\varepsilon(p-1), 2\varepsilon p\}.$$

Since $t - \varepsilon > 3t/4$, then by approximation one can find a subset $A_1 \subset T^n$ for which $A_1 \subset A + \varepsilon B_\infty^n$ and $N(A_1, \sqrt{n}B_2^n, t - \varepsilon) \ge N(A, \sqrt{n}B_2^n, t)$. Therefore, there exists a subset $A_2 \subset A_1$ of cardinality $|A_2| \ge N(A, \sqrt{n}B_2^n, t)$, which is $\frac{3t}{4}\sqrt{n}$-separated with respect to the $\|\cdot\|_2$-norm. Note that every two distinct points $x, y \in A_2$ satisfy that $\sum_{i=1}^n |x(i) - y(i)|^2 \ge (9t^2/16)n \ge t^2 n/2$ and that $|x(i) - y(i)|^2 \le 4$ for all i. Hence $|x(i) - y(i)| \ge t/2$ on at least $t^2 n/16$ coordinates i. By corollary 1 applied to A_2, $\log |A_2| \le C \log^2(2/t\varepsilon) \cdot \mathrm{VC}(A_2, t/2)$, and since $A_2 \subset A_1 \subset A + \varepsilon B_\infty^n$, our claim follows. □

From this we derive the entropy estimate (1).

Corollary 3. *There exists an absolute constant C such that for any convex body $K \subset B_\infty^n$ and every $0 < t < 1$,*

$$\log N(K, \sqrt{n}B_2^n, t) \le C \log^2(2/t) \cdot \mathrm{VC}(K, t/4).$$

Proof. This estimate follows from Corollary 2 by selecting $\varepsilon = t/4$ and recalling the fact that for every convex body $K \subset \mathbb{R}^n$ and every $0 < b < a$,

$$\mathrm{VC}(K + bB_\infty^n, a) \le \mathrm{VC}(K, a - b).$$

The latter inequality is a consequence of the definition of the VC-dimension and the observation that if $0 < b < a$ are such that $aB_\infty^n \subset K + bB_\infty^n$, then $(a - b)B_\infty^n \subset K$. □

Note that Corollary 2 and Corollary 3 can be extended to the case where the covering numbers are computed with respect to $n^{1/p}B_p^n$ for $1 < p < \infty$, thus establishing the complete claim in (1).

3 Fat-Shattering Dimension and Entropy

Turning to the case of function classes, we will show that one can obtain empirical entropy bounds in terms of the fat-shattering dimension.

Let $F \subset B\big(L_\infty(\Omega)\big)$ and fix a set $s_n \in \Omega$. For every $f \in F$ let $f/s_n = \sum_{i=1}^n f(x_i)e_i \in F/s_n$. Recall that, $\|f - g\|_{L_2(\mu_n)} = \|f/s_n - g/s_n\|_{\sqrt{n}B_2^n}$, implying that for every $t > 0$, $N\big(F, L_2(\mu_n), t\big) = N\big(F/s_n, \sqrt{n}B_2^n, t\big)$. Also, observe that for any $t > 0$,

$$\mathrm{VC}\big(F/s_n + \frac{t}{8}B_\infty^n, \frac{t}{2}\big) \le \mathrm{fat}_{\frac{t}{4}}\big(F/s_n + \frac{t}{8}B_\infty^n\big) \le \mathrm{fat}_{\frac{t}{8}}(F/s_n) \le \mathrm{fat}_{\frac{t}{8}}(F). \quad (12)$$

Theorem 5. *There is an absolute constant C such that for any class $F \subset B\big(L_\infty(\Omega)\big)$, any integer n, every empirical measure μ_n and every $t > 0$,*

$$\log N\big(F, L_2(\mu_n), t\big) \leq C\mathrm{fat}_{t/8}(F)\log^2(2/t).$$

Proof. Let $s_n = \{x_1, ..., x_n\}$ be the points on which μ_n is supported, and apply corollary 2 for the set F/s_n. We obtain that

$$\log N(F/s_n, \sqrt{n}B_2^n, t) \leq C\log^2(2/t) \cdot \mathrm{VC}(F/s_n + \frac{t}{8}B_\infty^n, t/2)$$

and our claim follows from (12). □

Remark. It is possible to show that this bound is essentially tight. Indeed, fix a class $F \subset B\big(L_\infty(\Omega)\big)$ and put $H(t) = \sup_n \sup_{\mu_n} \log N\big(F, L_2(\mu_n), t\big)$ (that is, the supremum is taken with respect to all the empirical measures supported on a finite set). By theorem 5, $H(t) \leq C\mathrm{fat}_{\frac{t}{8}}(F)\log^2\big(2/t\big)$. On the other hand it was shown in [9] that $H(t) \geq c\mathrm{fat}_{16t}(F, \Omega)$ for some absolute constant c.

This uniform entropy estimate gives us the opportunity to prove a result which is important in the context of empirical processes.

Empirical covering numbers play a central role in the theory of empirical processes. They can be used to characterize classes which satisfy the *uniform law of large numbers* (see [4] or [18] for a detailed discussion). Indeed, if $F \subset B\big(L_\infty(\Omega)\big)$ then F satisfies the uniform law of large numbers if and only if $\sup_{\mu_n} \log N\big(F, L_2(\mu_n), \varepsilon\big) = o(n)$ for every $\varepsilon > 0$, where the supremum is taken with respect to all empirical measures supported on at most n elements of Ω. In [1] it was shown that $F \subset B\big(L_\infty(\Omega)\big)$ satisfies the uniform law of large numbers if and only if $\mathrm{fat}_\varepsilon(F) < \infty$ for every $\varepsilon > 0$.

Another important application of covering numbers estimates is the analysis of the *uniform central limit property*.

Definition 2. *Let $F \subset B\big(L_\infty(\Omega)\big)$, set P to be a probability measure on Ω and assume G_P to be a gaussian process indexed by F, which has mean 0 and covariance*

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

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

A property stronger than the universal Donsker property is the uniform Donsker property. For such classes, ν_n^P converges to G_P uniformly in P in some sense. A detailed discussion on uniform Donsker classes is beyond the scope of this article. We refer the reader to [5] or [4] for additional information.

It is possible to show that the uniform Donsker property is connected to estimates on covering numbers [4].

Theorem 6. *Let $F \subset B\big(L_\infty(\Omega)\big)$. If*

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

then F is a uniform Donsker class.

Having this entropy condition in mind, it is natural to try to find covering numbers estimates which are "dimension free", that is, do not depend on the size of the sample. In the Boolean case, such bounds where first obtained by Dudley (see [8] Theorem 14.13), and then improved by Haussler [6,18] who showed that for any empirical measure μ_n and any Boolean class F,

$$N(F, L_2(\mu), \varepsilon) \le Cd(4e)^d \varepsilon^{-2d},$$

where C is an absolute constant and $d = \mathrm{VC}(F)$. In particular this shows that every VC class is a uniform Donsker class.

We can extend Dudley's result from VC classes to the real valued case.

Corollary 4. *If $F \subset B(L_\infty(\Omega))$ and $\int_0^1 \sqrt{\mathrm{fat}_{t/8}(F)} \log \frac{2}{t}\, dt$ converges then F is a uniform Donsker class.*

In particular this shows that if $\mathrm{fat}_\varepsilon(F)$ is "slightly better" than $1/\varepsilon^2$, then F is a uniform Donsker class.

4 Gaussian Complexities

The result we present below improves the main result of M. Talagrand from [15].

Theorem 7. *There are absolute constants $C, c > 0$ such that for every convex body $K \subset B_\infty^n$*

$$E \le C\sqrt{n} \int_{cE/n}^1 \sqrt{\mathrm{VC}(K, ct)} \log(2/t) dt,$$

where $E = \mathbb{E}\| \sum_{i=1}^n g_i e_i \|_{K^\circ}$, and $(e_i)_{i=1}^n$ is the canonical vector basis in $\mathbb{R}^n$.

For the proof, we need a few standard definitions and facts from the local theory of Banach spaces, which may be found in [12].

Given an integer n, let S^{n-1} be the unit Euclidean sphere with the normalized Lebesgue measure σ_n, and for every measurable set $A \subset \mathbb{R}^n$ denote by $\mathrm{vol}(A)$ its Lebesgue measure in $\mathbb{R}^n$. For a convex body K in $\mathbb{R}^n$, put $M_K = \int_{S^{n-1}} \|x\|_K \, d\sigma_n(x)$ and let M_K^* denote M_{K°, where K° is the polar of K. Recall that for any two convex bodies K and L, $M_{K+L}^* \le M_K^* + M_L^*$. Urysohn's inequality states that $\left(\frac{\mathrm{vol}(K)}{\mathrm{vol}(B_2^n)}\right)^{1/n} \le M_K^*$.

Next, put $\ell(K) = \mathbb{E}\| \sum_{i=1}^n g_i e_i \|_K$, where $(g_i)_{i=1}^n$ are independent standard gaussian random variables and $(e_i)_{i=1}^n$ is the canonical basis of $\mathbb{R}^n$. It is well

known that $\ell(K) = c_n\sqrt{n}M_K$, where $c_n < 1$ and $c_n \to 1$ as $n \to \infty$. Recall that by Dudley's inequality (see [14]) there is an absolute constant C_0 such that for every convex body K,

$$\ell(K^\circ) \le C_1 \int_0^\infty \sqrt{\log N(K, B_2^n, \varepsilon)}\, d\varepsilon.$$

It is possible to slightly improve Dudley's inequality using an additional volumetric argument. This observation is due to A. Pajor, and we omit its proof.

Lemma 2. *There exist absolute constants C and c such that for any set $K \subset \mathbb{R}^n$*

$$\ell(K^\circ) \le C \int_{cM_K^*}^\infty \sqrt{\log N(K, B_2^n, \varepsilon)}\, d\varepsilon.$$

Proof of Theorem 7. By Lemma 2, there exist absolute constants C and c such that

$$E = \ell(K^\circ) \le C \int_{cE/\sqrt{n}}^\infty \sqrt{\log N(K, B_2^n, t)}\, dt.$$

Since $K \subset B_\infty^n \subset \sqrt{n}B_2^n$, the integrand vanishes for all $t \ge \sqrt{n}$. Therefore, changing the integration variable and using Corollary 3,

$$\begin{aligned} E &\le C \int_{cE/\sqrt{n}}^{\sqrt{n}} \sqrt{\log N(K, B_2^n, t)}\, dt = C\sqrt{n} \int_{cE/n}^1 \sqrt{\log N(K, n^{1/2}B_2^n, t)}\, dt \\ &\le C\sqrt{n} \int_{cE/n}^1 \sqrt{VC(K, ct)} \log(2/t)\, dt, \end{aligned}$$

as claimed. □

In a similar way to theorem 7 one can obtain the following:

Theorem 8. *There are absolute constants C and c for which the following holds. Let F be a class of functions which are all bounded by 1, set $s_n = \{x_1, ..., x_n\}$ to be a sample and let $G_n = n^{-1/2} \sup_{f\in F} |\sum_{i=1}^n g_i f(x_i)|$. Then,*

$$G_n \le C \int_{cG/\sqrt{n}}^{\sigma_F(s_n)} \sqrt{\mathrm{fat}_{ct}(F)} \log(2/t)\, dt, \tag{13}$$

where $\sigma_F^2(s_n) = n^{-1}\sum_{i=1}^n f^2(x_i)$ and μ_n is the empirical measure supported on s_n.

To complement this result, let us mention another one of Talagrand's results, which enables one to estimate the expectation of σ_F^2 when s_n is selected randomly according to an underlying probability measure.

Lemma 3. *[16] There is an absolute constant C such that for any class F of functions bounded by 1,*

$$\mathbb{E}_\mu \sup_{f\in F} \sum_{i=1}^n f^2(X_i) \le n\tau^2 + C\mathbb{E} \sup_{f\in F} \Big|\sum_{i=1}^n g_i f(X_i)\Big|,$$

where (X_i) are independent, distributed according to μ and $\tau^2 = \sup_{f\in F} \mathbb{E}_\mu f^2$.

Results of the nature of theorem 8 combined with lemma 3 were used to obtain estimates or the "localized" gaussian averages in [10,11] to obtain improved complexity estimates for various learning problems.

5 Application: B_∞^n-Entropy

In this section we prove estimate (2), which improves the main combinatorial result in [1]. Our result can be equivalently stated as follows.

Theorem 9. *Let $K \subset B_\infty^n$ be a convex body, set $t > 0$ and put $v = \mathrm{VC}(K, t/8)$. Then,*

$$\log N(K, B_\infty^n, t) \le Cv \cdot \log^2(n/tv), \tag{14}$$

where C is an absolute constant.

This estimate should be compared with the Sauer-Shelah lemma for subsets of the Boolean cube $\{0,1\}^n$. It says that if $A \subset \{0,1\}^n$ then for $v = \mathrm{VC}(K)$ we have $|A| \le \binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{v}$, so that

$$\log |A| \le 2v \cdot \log(n/v)$$

(and note that, of course, $|A| = N(K, B_\infty^n, t)$ for all $0 < t < 1/2$).

We reduce the proof of (14) to an application of the B_p^n-entropy estimate (1). As a start, note that for $p = \log n$, $B_\infty^n \subset n^{1/p} B_p^n \subset eB_\infty^n$. Therefore, an application of (1) for this value of p yields

$$\log N(K, B_\infty^n, t) \le Cv \cdot \log^2(n/t),$$

which is slightly worse than (14).

To deduce (14) we need a result that compares the B_∞^n-entropy to the B_p^n-entropy, and which may be useful in other applications as well.

Lemma 4. *There is an absolute constant $c > 0$ such that the following holds. Let A be a subset of B_∞^n such that every two distinct points $x, y \in A$ satisfy $\|x - y\|_\infty \ge t$. Then, for every integer $1 \le k \le n/2$, there exists a subset $A' \subset A$ of cardinality*

$$|A'| \ge \binom{n}{k}^{-1} (ct)^k |A|,$$

with the property that every two distinct points in A' satisfy that $|x(i) - y(i)| \ge t/2$ for at least k coordinates i.

Proof. We can assume that $0 < t < 1/8$. Set $s = t/2$. The separation assumption imply that $N(A, B_\infty^n, s) \ge |A|$. Denote by D_k the set of all points x in $\mathbb{R}^n$ for which $|x(i)| \ge 1$ on at most k coordinates i. One can see that $N(A, D_k, s) = N(A, sD_k, 1) = N(A, sD_k \cap 3B_\infty^n, 1)$. Then, by the submultiplicative property of the covering numbers,

$$\begin{aligned} N(A, B_\infty^n, s) &\le N(A, sD_k \cap 3B_\infty^n, 1) \cdot N\big(sD_k \cap 3B_\infty^n, B_\infty^n, s\big) \\ &\le N(A, sD_k, 1) \cdot N\big(sD_k \cap 3B_\infty^n, B_\infty^n, s\big). \end{aligned} \tag{15}$$

To bound the second term, write D_k as

$$D_k = \bigcup_{|\sigma|=k} \Big(\mathbb{R}^\sigma + (-1,1)^{\sigma^c}\Big),$$

where the union is taken with respect to all subsets $\sigma \subset \{1,\dots,n\}$ and σ^c is the complement of σ. Thus,

$$sD_k \cap 3B_\infty^n = \bigcup_{|\sigma|=k} \Big(3B_\infty^\sigma + (-s,s)^{\sigma^c}\Big).$$

Denote by $N'(A,B,t)$ the number of translates of tB by vectors in A needed to cover A. Therefore,

$$\begin{aligned} N\big(sD_k \cap 3B_\infty^n, B_\infty^n, s\big) &\le \sum_{|\sigma|=k} N\big(3B_\infty^\sigma + (-s,s)^{\sigma^c}, B_\infty^n, s\big) \\ &\le \sum_{|\sigma|=k} N'(3B_\infty^\sigma, B_\infty^n, s). \end{aligned}$$

The latter inequality holds because any cover of $3B_\infty^\sigma$ by translates of sB_∞^n automatically covers $3B_\infty^\sigma + (-s,s)^{\sigma^c}$. Hence, for some absolute constant C,

$$\begin{aligned} N\big(sD_k \cap 3B_\infty^n, B_\infty^n, s\big) &\le \binom{n}{k} N'(3B_\infty^k, B_\infty^k, s) \\ &\le \binom{n}{k} (C/s)^k \end{aligned}$$

by a comparison of the volumes, and by (15) we obtain

$$N(A, D_k, s) \ge \binom{n}{k}^{-1} (cs)^k N(A, B_\infty^n, s) \ge \binom{n}{k}^{-1} (ct)^k |A|,$$

from which the statement of the lemma follows by the definition of D_k. □

Proof of Theorem 9. Fix $0 < t < 1$, and define α by $\log N(K, B_\infty^n, t) = \exp(\alpha n)$. Hence, there exists a set $A \subset K$ of cardinality $|A| = \exp(\alpha n)$, where every two distinct points $x, y \in A$ satisfy that $\|x-y\|_\infty \ge t$. Applying Lemma 4 we obtain a subset $A' \subset A \subset K$ of cardinality

$$|A'| \ge \binom{n}{k}^{-1} (ct)^k e^{\alpha n},$$

such that for every two distinct points in A', $|x(i) - y(i)| \ge t/2$ on at least k coordinates i. Selecting $k = \frac{c\alpha n}{\log(2/t\alpha)}$ we see that $|A'| \ge e^{\alpha n/2}$.

The proof is completed by discretizing A' and applying Corollary 1 with $p = 4/t$ and $\varepsilon = k/n$ in the same manner as we did in the previous section. Therefore

$$\begin{aligned} \alpha n/2 = \log|A'| &\le C \log^2\Big(\frac{4n}{tk}\Big) \cdot \mathrm{VC}(A' + (t/4)B_\infty^n, t/2) \\ &\le C \log^2(1/t\alpha) \cdot \mathrm{VC}(K, t/4), \end{aligned}$$

and thus $\alpha n \le c \log^2(n/tv) \cdot v$, as claimed. □

Remark. As the proof shows, only a slight modification is needed to obtain the analogous result for function classes. Due to the lack of space, we omit the formulation and the proof of this assertion.

References

1. N. Alon, S. Ben-David, N. Cesa-Bianchi, D. Hausser, *Scale sensitive dimensions, uniform convergence and learnability*, Journal of the ACM 44 (1997), 615–631. 15, 20, 23, 26
2. M. Anthony, P. L. Bartlett, *Neural Network Learning, Theoretical Foundations*, Cambridge University Press, 1999. 18
3. P. Bartlett, P. Long, *Prediction, learning, uniform convergence, and scale-sensitive dimensions*, J. Comput. System Sci. 56 (1998), 174–190. 20
4. R. M. Dudley, *Uniform central limit theorems*, Cambridge University Press, 1999. 23
5. E. Giné, J. Zinn, *Gaussian charachterization of uniform Donsker classes of functions*, Annals of Probability, 19 (1991), 758–782. 23
6. D. Haussler, *Sphere packing numbers for subsets of Boolean n-cube with bounded Vapnik-Chervonenkis dimension*, Journal of Combinatorial Theory A 69 (1995), 217–232. 14, 24
7. W. Hoeffding, *Probability inequalities for sums of bounded random variables*, J. Amer. Statist. Assoc. 58 (1963), 13–30. 19
8. M. Ledoux and M. Talagrand, *Probability in Banach spaces*, Springer, 1991. 14, 15, 19, 24
9. S. Mendelson, *Rademacher averages and phase transitions in Glivenko-Cantelli classes*, IEEE transactions on Information Thery, Jan 2002. 15, 23
10. S. Mendelson, *Improving the sample complexity using global data* To appear, IEEE transactions on Information Theory. 16, 26
11. S. Mendelson, *Geometric parameters of Kernel Machines*, These proceedings. 16, 26
12. V. Milman, G. Schechtman, *Asymptotic theory of finite dimensional normed spaces*, Lecture Notes in Math., vol. 1200, Springer Verlag, 1986. 24
13. A. Pajor, *Sous espaces ℓ_1^n des espaces de Banach*, Hermann, Paris, 1985. 15
14. G. Pisier, *The volume of convex bodies and Banach space geometry*, Cambridge University Press, 1989. 25
15. M. Talagrand, *Type, infratype, and Elton-Pajor Theorem*, Inventiones Math. 107 (1992), 41–59. 17, 24
16. M. Talagrand, *Sharper bounds for Gaussian and empirical processes*, Annals of Probability, 22(1), 28–76, 1994. 25
17. V. Vapnik, A. Chervonenkis, *Necessary and sufficient conditions for uniform convergence of means to mathematical expectations*, Theory Prob. Applic. 26(3), 532-553, 1971. 14
18. A. W. Van–der–Vaart, J. A. Wellner, *Weak convergence and Empirical Processes*, Springer-Verlag, 1996. 23, 24